



Canadian Journal of Higher Education
Revue canadienne d'enseignement supérieur
Volume 37, No. 1, 2007, pages 27 - 43
www.ingentaconnect.com/content/csshe/cjhe

What's the "Use" of Student Ratings of Instruction for Administrators? One University's Experience

Tanya Beran
Claudio Violato
Don Kline

University of Calgary

ABSTRACT

At most Canadian and American community colleges and universities, student ratings have been implemented as a means of evaluating course instruction. Although concerns regarding the validity of student ratings from instructors' perspectives have been studied quite extensively, issues associated with the use of student ratings information by administrators have been largely ignored. In this study, we surveyed 52 administrators at a major Canadian university about the types of ratings they use, how useful they are, and their purpose. Our findings indicate that administrators are interested in knowing about instructor characteristics and teaching procedures. In addition, ratings are being used for instructor and department evaluation as well as scheduling courses. In general, administrators regard student ratings positively and think that they are useful. However, they have some reservations.

RÉSUMÉ

Dans la majorité des collèges et des universités, les étudiants évaluent leurs professeurs. Les inquiétudes de la validité de ces évaluations selon les professeurs, ont été bien étudiées, mais l'utilisation et les perspectives des administrateurs n'ont pas été étudiées. Dans cet étude 52 administrateurs dans une grande université Canadienne ont répondu aux

questions selon les évaluations. Les résultats indiquent que les administrateurs s'intéressent aux caractéristiques des professeurs et de leurs apprentissages. De plus, les administrateurs utilisent les évaluations pour organiser les cours. En général, les administrateurs ont donné des réponses très positives avec peu de réservations.

Since their inception in the 1920s, the use of student ratings for the evaluation of instruction has proliferated to the point that they are now used regularly at almost all universities and community colleges in Canada and the United States (Algozzine et al., 2004). Student ratings have been administered with the intention of providing instructional feedback to instructors (their "formative" function), to inform students in regard to course selection, and to make personnel decisions about instructors (their "summative" function). Indeed, administrators may make extensive use of student ratings information more than do instructors (Beran, Violato, Kline, & Frideres, 2005). Considering the potential impact of the use or misuse of student ratings information on instructor promotion and tenure decisions, it is important to investigate the type of information that administrators consider useful, the reasons they use the ratings, the type(s) of information used, and their attitudes about the adequacy of this method. The major purpose of the present paper is to further investigate the use of student ratings of instruction by administrators within the context of a major research university.

In an attempt to meet demands for accountability in higher education, student ratings have become widespread and standardized (Hourcade, Parette, & Anderson, 2003; Nasser & Fresko, 2002). Even decisions about financial support to a university may be affected by ratings results (Hourcade et al., 2003). The use of student ratings has also been described as "a politically expedient performance measure for quality monitoring" (Penny & Coe, 2004, p. 215). Hence, ratings have obtained considerable prominence in demonstrating value in higher education.

An extensive body of research on the reliability and validity of student ratings has been accumulated (Ali & Sell, 1998; Greenwald, 2002; Ory & Ryan, 2001; Schmelkin, Spencer, & Gellman, 1997). Most researchers consider student ratings to be useful measures of the instructional behaviours that contribute to teaching effectiveness (Marsh, 1987; Schmelkin, et al., 1997). And although such ratings do not measure student learning directly, higher ratings tend to indicate greater learning than do lower ratings (Abrami & Apollonia, 1986; Cohen, 1981). Reliability and validity are also indicators of an instrument's utility or usefulness (Messick, 1988). For administrators, the utility of student ratings can be determined according to how they use ratings, the type of ratings information they use, and their perception about the usefulness of the ratings.

Purpose of Student Ratings

Several groups, including students, instructors, and administrators, have a stakeholder interest in student ratings. Student ratings can provide information to instructors about the quality of their teaching, to students to assist them in course selection, and to course developers on the effectiveness of instructional strategies (Kulik, 2001; Nasser & Fresko, 2002; Newport, 1996; Schmelkin, et al., 1997).¹ However, the use of student ratings information by administrators has been largely ignored in the research. It is generally assumed that administrators use ratings to inform decisions about promotion and tenure, but empirical evidence regarding such a function is rare. Anecdotal evidence suggests that student ratings are frequently used for administrative purposes such as making personnel decisions (Haskell, 1997; Nasser & Fresko, 2002; Schmelkin, et al., 1997). Specifically, ratings may influence an administrator's decisions about instructor salary adjustment, tenure, and promotion. In some research universities, student ratings may be the only source of teaching information available, and may have the second highest weighting value after publications in the evaluation of university instructors (Abrami, 2001; Haskell, 1997).

In addition to assisting instructors to improve their teaching through ratings feedback, administrators may be able to monitor specific course improvements. This information may provide administrators with the ability to track changes in teaching skills more generally. Such information can also be aggregated to determine the teaching quality in a department or program in relation to other programs. Procedural decisions may also be informed by ratings. For example, ratings may determine instructors' course assignments in subsequent terms. Kulik (2001) mentions additional purposes such as hiring instructors, obtaining accreditation, and rewarding instructors for exceptional teaching. Thus, administrators may use ratings to make other administrative decisions in addition to promotion and merit.

Content of Student Ratings

Student rating instruments tend to include a variety of items (e.g., Arreola, 2000). These may include ratings of instructors' enthusiasm, organization, and interactions with students (e.g., kindness, attention, and respect shown to students). Instruments also often include items regarding instructors' teaching approaches such as the types of materials provided, clarity of explanations, expectations for assignments, and fairness in marking.

Researchers have indicated that multiple dimensions of ratings (Centra, 1993; Feldman, 1976a; Marsh, 1982; 1987; Marsh & Roche, 1993) such as the teacher, the course, assessment issues, classroom rapport and workload/difficulty are separable issues. Students may provide positive ratings for some teaching aspects such as enthusiasm, for example, but lower ratings for organization. Depending on the instrument, however, students may also develop a general perception of the course and rate all instrument items similarly (Greenwald,

1997). Student consistency in ratings across items has been explored (Beran et al., 2005). Consistency in administrators' perceptions of the usefulness of these items is unknown, however. It is also unknown whether the type of information they find useful is related to their purpose(s) in using the ratings. It is possible, for example, that administrators rely more on information about instructors' teaching approaches than their personal characteristics to evaluate teaching quality.

Administrators' Reactions to Student Ratings

Considering the lack of previous research on administrators' views of student ratings, and that most administrators are or have been faculty members themselves, a review of research on instructors' perceptions of student ratings may be useful in understanding how administrators in their management role perceive the use of ratings.

In using student ratings, administrators presumably need to consider faculty reactions to the types and quality of information that they provide. Instructors' opinions about student ratings can range widely, from supportive to hostile (Newport, 1996; Schmelkin, et al., 1997; Wachtel, 1998). Many researchers have concluded that instructors' views of student ratings are generally negative (Abrami, 2001; Centra, 1993; Fries & McNinch, in press; Nasser & Fresko, 2002; Theall & Franklin, 2001; Wachtel, 1998). Reasons for opposition include the concern that student ratings may be biased by characteristics of the instructors and courses such as grades on assignments (Eiszler, 2002; Feldman, 1976b). The competence of students in evaluating some aspects of teaching ability (e.g., determining if materials are updated) has also been questioned (Lowman, 1984). The introduction of student ratings has also been observed to have decreased job satisfaction and grading standards of some instructors (Birnbaum, 2000; Haskell, 1997; Ryan, Anderson, & Birchler, 1980). The use of students' ratings for summative rather than formative purposes may also raise concerns for faculty. Given such reactions to student ratings, it follows that many instructors may be concerned about how administrators use student ratings in making personnel decisions (Nasser & Fresko, 2002; Sproule, 2000).

Administrators, however, are often directed by institutional policy as well as student demand to implement student ratings, seemingly with little opportunity of expressing their unique perspectives about the usefulness of such instruments. To be comprehensive, an evaluation of the usefulness of student ratings must include the views of administrators who are major users of information derived from student ratings of instruction. Thus, we investigated the reasons why administrators used the ratings, the types of ratings information that they considered useful, their need for information not provided by the ratings instrument, and their views about the rating process.

METHOD

Participants

Our study was conducted at a major Canadian research intensive doctoral/medical university with approximately 20,000 undergraduate students, 5,000 graduate students, and 1,800 full time faculty and sessional instructors. A survey of the usefulness of student ratings was sent to all university Department Heads and Deans, 52 of whom completed and returned it. This corresponded to a response rate of 53%. Of these, 27 (52%) were Department Heads, 6 (12%) were Deans, and 19 (36%) were Associate Deans. The majority of the university's faculties were represented in the surveys returned, although nearly two thirds of the respondents ($n = 33$, 63%) did not indicate their faculty or department.

Instrument

A panel of faculty members from the Faculty of Education, and Departments of Psychology and Sociology who were experienced in questionnaire development and psychometric research developed the administrator survey. In addition to their professional expertise in such evaluation, the panel members had prior experience with student evaluation of their teaching using the university's 12-item student ratings instrument (see Appendix A), and hence were familiar with its content validity.

The administrator survey consisted of three general parts (see Appendix B). The first of these (13 items) asked administrators to indicate the usefulness for functions they served for both individual (9 items: e.g., merit evaluations, tenure and promotion recommendations, teaching awards, identification of teaching problems.) and unit-level issues (4 items: e.g., course timetabling, documenting unit teaching, promoting the unit). The second section (13 items) asked administrators to rate the usefulness of each of the 13 specific items that comprise the university-wide student ratings scale.

Administrator responses in both of the first two sections were rated on a 4-point scale from "Not Useful at All" to "Very Useful," with higher values indicating greater usefulness. For each item, respondents could also indicate whether an item was "Not Applicable" to their role or unit. The third section asked administrators to report their opinions on a range of issues, including the cost and people involved in administering the ratings instrument the frequency of administration of the instrument, the need for information beyond that provided by the instrument and the use of measures other than the ratings instrument for evaluating teaching.

Procedure

The survey instrument described above was sent to all university Deans, Head and Head-equivalent administrators in the university. Except for some pre-defined waiver exceptions (e.g., new experimental courses, low enrollment

courses, instructor illness) university policy mandated that the instrument be administered in “every course, every term,” beginning in the fall of the 1998-99 academic year. The system also allowed for the evaluation of up to four intra-course segments where each segment was associated with a different instructor (i.e., in a sequential “team-taught” course).

Administrators receive these ratings as a mean, frequency distribution, and standard deviation on each rating item for each course/instructor combination. The number of student respondents and course enrollees are also reported. Mean and standard deviation comparison information for the department and instructors at the same level (i.e., junior level, senior level) is also provided for each rating item. In addition, the mean student rating of the course workload and the total number of times the instructor had taught the course were indicated. Finally, an optional 60-word summary written by the instructor(s) about the course could also be included. (To facilitate their selection of courses, much of the same information, including the instructor’s summary is also made available to students on a restricted access Web site.)

RESULTS

To explore the basic structure and the dimensionality of the survey instrument, a principal component analysis with varimax rotation was conducted on the response data from the first section of the survey (administrator reasons for using the student ratings information). The analysis was selected to allow us to determine the factor structure of the scale. Four components with eigenvalues greater than 1.0 were extracted. Table 1 presents the *means, standard deviations, component loadings, eigenvalues and percent variance* from these 13 items. When an item had a loading of greater than .40 on two components the higher loading was used to assign it to a component (Kerlinger & Lee, 2000). The first component was comprised of items about the evaluation of *teaching quality*, and the second measured improvements and *progress in teaching*. The third component reflected administrators’ use of student ratings to *evaluate and promote* the teaching of instructors. Finally, two items related to administrators’ use of student rating information to develop *teaching schedules* loaded highly on a fourth component. Inter-item consistencies (Cronbach’s alpha) comprising each of the four components were .89, .84, .79, and .71, respectively. The means in Table 1 showed high agreement among administrators that they use student ratings to identify the quality of teaching, reward teaching, and determine merit, whereas they tended to disagree that ratings were used to timetable courses.

A second principal component analysis was carried out on the data from the second section of the survey regarding the administrative utility of information from each of the specific items from the student rating scale (see Table 2). Two components with eigenvalues greater than 1.0 were extracted: *teaching procedures* (e.g., course follows outline, support materials helpful) and the *instructor characteristics* (e.g., enthusiasm, students treated respectfully). The internal consistencies (Cronbach’s alpha) for each of these components were .91, and .93, re-

spectively. Administrators' mean ratings indicated that the information that they found to be most useful included knowing the instructors' overall instructional ability and the respect shown to students. The information that they found least useful was the instructor's ability to follow the course outline.

Table 1. Means, Standard Deviations, Component Loadings*, Eigenvalues and Percent of Explained Variance of Purpose of Student Ratings

Items	Mean	SD	Components			
			Evaluate teaching	Monitor progress	Evaluate unit	Create schedule
Faculty merit	3.26	.69	.87	.31	.07	.10
Tenure	3.10	.78	.79	.45	.17	.11
Promotion	3.12	.77	.85	.36	.14	.17
Identifying good/poor teaching	3.42	.59	.65	.00	.27	-.40
Teaching awards	3.31	.82	.43	.73	-.12	.14
Remediation of teaching problems	3.00	.96	.13	.83	.29	-.04
Reappointment of sessional instructors	3.17	.82	.19	.77	.22	.07
Tracking teaching	2.95	.74	.34	.72	.09	.16
Assigning courses to faculty	2.11	.85	-.05	.34	.14	.80
Deciding on timetable	1.83	.78	.21	-.10	.34	.81
Documenting overall quality of unit's teaching	2.71	.84	.24	.05	.86	.24
Analyzing trends in unit's teaching	2.55	.83	.11	.25	.78	.09
Promoting the unit	2.39	.87	.05	.12	.86	.11
Eigenvalue			5.56	2.08	1.40	1.08
% of variance			23.16	22.75	19.30	12.57

* The cells containing the highest component loading item are bolded.

Table 2. Means, Standard Deviations, Component Loadings*, Eigenvalues and Percent of Variance for Administrative Utility

Items	Mean	SD	Components	
			Teaching procedures	Instructor characteristics
Overall quality of instruction	3.37	.76	-.03	.73
Course outline provided enough detail	2.70	.88	.77	.34
Course followed course outline	2.63	.92	.86	.27
Organization of content	3.23	.80	.61	.66
Student questions responded to appropriately	3.17	.79	.59	.73
Course communicated with enthusiasm	3.13	.88	.55	.74
Opportunities for assistance were available	3.10	.79	.44	.83
Students were treated respectfully	3.27	.90	.54	.73
Evaluation methods were fair	2.90	.85	.67	.57
Student work was graded in a reasonable time	2.93	.85	.62	.52
Students learned a lot in the course	3.00	1.01	.75	.21
Support materials were helpful	2.70	.92	.84	.33
Number of students in the class completing ratings	3.00	.88	.72	.15
Eigenvalue			8.64	1.09
% of variance			42.06	32.74

* The cells containing the highest component loading items are bolded.

To examine the relationships among the four purposes and the two aspects of administrative utility, the sum of all items that loaded highest on each component (bolded in table) was calculated to obtain a single subscale score. Pearson's product moment correlations indicated that the teaching procedures subscale was related to teaching quality ($r = .56$), monitoring teaching progress ($r = .49$), assessing unit quality ($r = .72$) and course scheduling ($r = .47$). The instructor characteristic subscale was related to the evaluation of teaching quality ($r = .44$) and to monitoring teaching progress ($r = .44$). These moderate to high correlations suggest that these factors are related, but also measure distinct constructs.

Table 3. Pearson's Correlations of Purpose and Administrative Utility

	Purpose			
	Evaluate teaching	Monitor progress	Evaluate unit	Create schedule
Administrative utility				
Teaching procedures	.56**	.49*	.72**	.47*
Instructor characteristics	.44*	.44*	.37	.33

Note. ** $p < .01$, * $p < .05$.

Content Analysis of Open-ended Items

The content analysis method, as presented in Berg (2006), when applied to administrator responses to the open-ended question about the suitability of the student ratings instrument in Section 3 of the survey, yielded several different themes. This included concerns regarding the limitations of the information provided by the student rating instrument. For example, some administrators indicated that the ratings are a valid indicator of teaching quality ($n = 16$, 31%), and some noted that ratings should be contextualized and supported by other information sources ($n = 12$, 23%). Some administrators ($n = 11$, 21%) stated that program size limited the utility of ratings information because of difficulties in maintaining student confidentiality (e.g., student demographics or particular responses might compromise their anonymity), limited student choices in program selection, and inadequate resources for administering the student ratings system (e.g., no one available to attend class and collect forms). A few administrators also indicated that the information provided by the student ratings was not useful for all teaching situations, such as when two or more instructors in a team taught course are present during the evaluation period ($n = 5$, 10%). Finally, a few philosophical disagreements were registered about student ratings: evaluation of the course and the instructor should be separated ($n = 2$, 4%), student ratings drive grade inflation ($n = 1$, 2%), or the ratings imply that instructors rather than students are solely responsible for student learning ($n = 1$, 2%). A total of 19 (36%) did not provide a written response.

In summary, the majority of administrators stated that they used the ratings results for various administrative functions, with the primary purposes being to

identify the quality of teaching of individual faculty members and the overall effectiveness of their unit. They also reported that information regarding overall course instruction was the most useful.

DISCUSSION

Our study showed that administrators use the student ratings to a moderate or high extent for a variety of administrative decisions. General information about both instructors' teaching procedures and characteristics was considered more useful for evaluating teaching quality than was information derived from specific survey items. Also, a minority of administrators expressed concerns with the use of ratings information in their open-ended written comments.

This study revealed that administrators use student ratings information for four different functions: evaluating individual teaching, monitoring progress, evaluating teaching at the unit level, and developing course schedules. In regard to the evaluation of individual instructors, administrators reported that student ratings were very helpful in making decisions about merit, promotion and tenure, and the identification of good and poor teaching. Moreover, they indicated that the teaching ability of instructors can be monitored in terms of changes in ratings over time, for recommendations on teaching awards, and for the remediation of teaching problems. Administrators also considered ratings to be useful in gauging and communicating the teaching effectiveness of the department or faculty. Finally, administrators found student ratings information useful for curriculum planning functions including the assignment of courses to faculty and course timetabling. They did not, however, find the ratings useful when scheduling courses. It is likely that other resource factors such as availability of instructors and/or suitable instructional space (e.g., labs), and times that courses need to be offered guide this type of decision making.

Given that the student ratings instrument was developed to measure teaching quality, it is not surprising that administrators find the ratings most useful for evaluating teaching quality. Thus, more general anecdotal reports of frequent use were empirically supported in this study. Despite instructors' reservations about being evaluated for tenure and promotion on the basis of students' ratings (Nasser & Fresko, 2002), administrators indicate that ratings are useful for this purpose.

Although student ratings were generally considered useful, a minority of administrators also reported concerns about their validity. Ratings were sometimes considered poor indicators of teaching quality (e.g., "simplistic") because they were not appropriate for all programs and teaching situations, they evaluated both the instructor and the course, or both. They also expressed concerns about the consequences of misuse including their potential effects on grade standards and over-emphasis on the responsibility of instructors in influencing student learning.

Why do administrators report ratings to be generally useful yet express concern about them? Although important for university-wide assessment, student ratings may not be sufficient alone in determining the quality of teaching. Indeed, consistent with university policy, many administrators identified the need

for additional information about teaching. Although it is generally recognized that multiple sources of evaluation are required by the university student ratings policy, limited resources (e.g., no other suitable evaluation instrument readily available, shortage of personnel time) in some units may limit evaluation to the use of rating scales as the most efficient and expedient way of obtaining information about teaching quality. Although some departments at this university use teaching portfolios and peer ratings from colleagues, these practices are not used in every department. As a result, most administrators use the student ratings. It is also important to consider that many administrators are, or at least were, instructors who have had experience with being evaluated by student ratings. Although asked to complete the survey as administrators, it is possible that they considered their own teaching experiences as well. These latter personal experiences may also explain any tension about student ratings. On one hand, it is possible that some administrators feel pressured to conform to policy and procedures, inflating the reported usefulness of rating information. On the other hand, recognition by administrators of the limitations of ratings instruments may also contribute to the more appropriate use of the information derived from them.

Information about instructors' teaching procedures and more personal instructional characteristics were considered useful for many purposes. Teaching procedures were more likely to be used for evaluating teaching within the unit as well as planning schedules. Both types of information were considered useful for evaluating and monitoring teaching. Thus, students' reactions to the respect, enthusiasm, and assistance demonstrated by instructors are taken into consideration when administrators evaluated teaching effectiveness. If teaching competence is enhanced by a positive relationship with students (Fereshteh, 1996; Hargreaves, 1998), then it seems appropriate that such characteristics be assessed in evaluating teaching quality. However, instructors with a less personable approach may find themselves judged more harshly by both students and administrators.

The mean scores indicated that administrators considered students' ratings of the overall quality of instruction and respect shown by instructors for students to be the most useful types of information. Given the lack of consensus about the specific qualities that comprise effective teaching, this finding is consistent with the recommendation that administrators use general, rather than detailed, judgments about teaching (Algozzine et al., 2004; d'Apollonia & Abrami, 1997). Indeed, based on weighted composites of multidimensional student ratings items, Cashin and Downey (1992) found that global items are the most useful indicator for teaching effectiveness. Of least importance to administrators was information regarding the consistency between instruction and the course outline. Apparently, administrators believe that instructors can teach effectively regardless of whether they follow an *a priori* outline.

Practical Implications

Just as faculty are more likely to make teaching changes when given specific support from an advisor to help interpret the ratings and make specific

behavioral adjustments (Penny & Coe, 2004), administrators are also likely to need training and support in the practice of appropriate evaluation. For example, many administrators come from academic backgrounds that include little or no training in statistics or psychometric measurement. There also may be little incentive or opportunity for learning the institutional policies that guide the administration and use of student ratings information. Abrami (2001) provides several suggestions for enhancing teaching evaluation including open communication about the results and interpretation guidelines to ensure both the accurate understanding of rating results and their limitations. Also, several respondents indicated that standardized administration of a student ratings instrument did not provide sufficient introductory information about the questions to assist students in completing the items. However, longer and more complex instruments are cumbersome to develop and their administration could remain inadequate. Hence, evaluations may only provide a general indication of teaching quality that should be supplemented by complementary sources of information regarding instructional effectiveness. All user groups, including administrators, faculty, and students should be aware of this limitation when using student ratings.

Although the present study informs our understanding of how administrators use student ratings, several limitations should be considered. Our sample size may not be adequate to represent the majority of administrators' perceptions of these instruments. Administrators at this university may have attitudes towards ratings that are more positive or negative than administrators more generally. A similar study at other universities that have used student ratings evaluation over a longer duration may show different results. This university may also be unique in that it has a campus-wide instrument. This consistency may have created an accepted norm that may prevent administrators from providing critical reflections on the survey. Also, administrators may differ in their beliefs as to whether student ratings results should be shared with students, and these differences may have affected their perceptions of the usefulness of student ratings. It is also possible that the ways in which results are shared with faculty (e.g., through annual discussion, written report, rankings) affected administrators' judgments of their use.

The results of the principal components analyses show that several items can load highly on more than one component. Thus, dimensions of the various uses and content of student ratings may not be distinct. Alternate measures of the usefulness of ratings for administrators should be developed to test the generalizability of our findings. Also, being asked to identify the department of each administrator may have affected their honesty in responding to questions.

Negative perceptions regarding the use of student ratings held by some administrators in this study raise questions that deserve further research examination. An issue relevant to understanding administrators' ratings use is their ability to accurately interpret their meaning, which can lead to the possibility of misuse (Franklin & Theall, 1989). As noted by Abrami (2001), it is uncertain whether administrators use the ratings appropriately to inform personnel or oth-

er decisions. Moreover, it is not yet clear what constitutes appropriate interpretation: should general or specific ratings or relative or absolute ratings be used? In addition, knowing more about the differences in the tasks and characteristics of administrators that affect the perceived administrative utility of student ratings would enhance our understanding of the issues associated with their use. Similarly, the impact of unit-specific procedures, traditions, approaches to teaching and organizational culture has yet to be explored. For example, although the policy on student ratings for the university in our study indicates that they should not be relied upon as the sole indicator of teaching effectiveness, administrator awareness and interpretation of this policy was not studied.

Whether students should be evaluating instructors is a topic of longstanding debate. Although students may be well-suited to evaluate instructors as the experienced "consumers" of the instruction, their competence in evaluating instructors has also been questioned (Newport, 1996). The lack of resolution on this issue coupled with the recognition that student ratings are necessarily subjective may influence how administrators use and their concerns about student ratings. Research to determine how these interact with different administrative tasks and the pedagogical traditions of different units and disciplines is also needed. ♣

NOTES

- ¹ Teacher evaluation includes student ratings as one method of feedback which can be interpreted in the context of additional teaching information such as a teaching portfolio.

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.

Abrami, P. C. & d'Apollonia, S. (1999). Current concerns are past concerns. *American Psychologist*, 54(7), 519-520.

Algozzine, B., Beattie, J., Bray, M., Flowers, C., Gretes, J., Howley, L., Mohanty, G., & Spooner, F. (2004). Student evaluation of college teaching: A practice in search of principles. *College Teaching*, 52(4), 134-141.

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. (2000). *Developing a comprehensive faculty evaluation system: A handbook for college faculty and administrators on designing and operating a comprehensive faculty evaluation system* (2nd ed.). Bolton, MA: Anker Publishing.

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. *Canadian Journal of Higher Education*, 35(2), 49-70.

Berg, B. L. (2006). *Qualitative research methods for the social sciences* (6th ed). NY: Pearson.

Birnbaum, M. H. (2000). *A survey of faculty opinions concerning student evaluations of teaching*. Retrieved on January 21, 2005, from <http://psych.fullerton.edu/mbirnbaum/faculty3.htm>

Cashin, W. E., & Downey, R. G. (1992). Using global student ratings items for summative evaluation. *Journal of Educational Psychology*, 84(4), 563-572.

Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Francisco: 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.

d'Apollonia, S., & Abrami, P. (1997). *Variables moderating the validity of student ratings of instruction: A meta-analysis*. Paper presented at the 77th annual meeting of the American Educational Research Association, New York.

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

Feldman, K. A. (1976a). The superior college teacher from the student's view. *Research in Higher Education*, 5(3), 243-288.

Feldman, K. A. (1976b). Grades and college students' evaluations of their courses and teachers. *Research in Higher Education*, 4(1), 69-111.

Fereshteh, H. (1996). The nature of teaching, effective instruction, and roles to play: A social foundations's perspective. *Contemporary Education*, 68(1), 73-75.

Franklin, J., & Theall, M. (1989). *Who reads ratings: Knowledge, attitude and practice of users of student ratings of instruction*. Paper presented at the Annual Meeting of the American Education Research Association, San Francisco.

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

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist*, 52(11), 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.

Hargreaves, A. (1998). The emotional practice of teaching. *Teaching and Teacher Education*, 14(8), 835-854.

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.

Hourcade, J., Parette, P., & Anderson, H. (2003). Accountability in collaboration: A framework for evaluation. *Education and Training in Developmental Disabilities*, 38(4), 398-404.

Kerlinger, F.N. & Lee, H.B. (2000) *Foundations of behavioral research*, Harcourt College Publishers.

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.

Lowman, J. (1984). *Mastering the techniques of teaching*. San Francisco, Jossey-Bass.

Marsh, H. W. (1982). SEEQ. A reliable, valid, and useful instrument for collecting student's evaluation of university teaching. *British Journal of Educational Psychology*, 52(1), 77-95.

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), 253-388.

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. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33-48). Hillsdale, NJ: Erlbaum.

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

Newport, F. J. (1996). Rating teaching in the USA: Probing the qualifications of student raters and novice teachers. *Assessment & Evaluation in Higher Education*, 21(1), 17-21.

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.

Penny, A. R., & Coe, R. (2004). Effectiveness of consultation on student ratings feedback: A meta-analysis. *Review of Educational Research*, 74(2), 215-253.

Ryan, J. J., Anderson, J. A., & Birchler, A. B. (1980). Student evaluation: The faculty responds. *Research in Higher Education*, 12(4), 317-333.

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.

Sproule, R. (2000). Student evaluation of teaching: A methodological critique of evaluation practices. *Education Policy Analysis Archives, 8*(50). Retrieved on January 21, 2005, from <http://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*. (No. 109, pp. 45-56). San Francisco: Jossey-Bass.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education, 29*(2), 191-121.

CONTACT INFORMATION

Dr. Tanya Beran
Division of Applied Psychology
University of Calgary, AB
T2N 1N4
E-mail: tnaberan@ucalgary.ca

Dr. Beran is an Assistant Professor at the University of Calgary teaching in the area of evaluation and measurement. As part of a university-wide evaluation of student ratings, she was asked by the university administration to report on the validity, reliability, and utility of students ratings of evaluation. She has published papers on the application of validity and structural equation modeling to school psychology.

Donald Kline is currently a Professor of Psychology and Surgery (Ophthalmology) at the University of Calgary. His research and teaching interests include vision, perception and aging. After receiving the Ph.D. at the University of Southern California, he taught at the University of Notre Dame. The recipient of numerous teaching awards, including the national 3M Fellowship, he has longstanding interests in curriculum design and teaching effectiveness

Dr. Claudio Violato is Professor and Director of the Medical Education and Research Unit in the Faculty of Medicine at the University of Calgary. He specializes in medical education and educational psychology. In addition to 10 books, Dr. Violato has published more than 200 scientific and technical articles and reports in journals such as *Canadian Journal of Education, Academic Medicine, Journal of Psychology, British Medical Journal and Paediatrics*.

APPENDIX A

Student Ratings Instrument

1. The overall quality of instruction was (unacceptable to excellent).
2. The course outline or other course descriptive information provided enough detail about the course (e.g., goals, reading list, topics covered, assignments, exams, due dates, grade weightings).
3. The course as delivered followed the outline and other course descriptive information.
4. The course content was presented in a well-organized manner.
5. Student questions and comments were responded to appropriately.
6. The course content was communicated with enthusiasm.
7. Opportunities for course assistance were available (e.g., instructor office hours, out-of-class appointments, e-mail, telephone, websites).
8. Students were treated respectfully.
9. The evaluation methods used for determining the course grade were fair.
10. Students' work was graded in a reasonable amount of time.
11. I learned a lot in this course.
12. The support materials (e.g., readings, audio-visual materials, speakers, field trips, equipment, software, etc.) used in this course helped me to learn.

APPENDIX B

Administrator Survey

Please rate the usefulness of the following functions:

1. For making recommendations/decision regarding faculty merit.
2. For making recommendations/decisions regarding tenure.
3. For making recommendations/decisions regarding promotion.
4. For identifying unusually good or poor teaching.
5. For making recommendations/decisions regarding teaching awards.
6. For making recommendations/decisions regarding remediation of teaching problems.
7. For making recommendations/decisions regarding reappointment of sessional instructors.
8. For tracking improvement or decline in a faculty member's teaching over time.
9. For deciding the course(s) to timetable for a particular faculty member.

Please rate the usefulness of the following student ratings items:

1. Overall quality of instruction.
2. Course outline or descriptive material provided enough detail.
3. Course as delivered followed the course outline.
4. Course content presented in a well-organized manner.
5. Student questions and comments responded to appropriately.
6. Course content communicated with enthusiasm.
7. Opportunities for course assistance were available.
8. Students were treated respectfully.
9. Evaluation methods for determining grades were fair.
10. Student work graded in a reasonable amount of time.
11. I learned a lot in this course.
12. The support materials used in the course helped students to learn.
13. The proportion of students in the class completing the rating.

Other issues:

1. The resources used under the current policy are worth the benefits.
2. The class time taken under the current policy is worth the benefits.
3. Faculty members seldom complain to me about the current frequency of administration of the student ratings instrument.
4. The unit's student ratings coordinator appears to agree with the current policy.