GUIDELINES for the INTERPRETATION of STUDENT EVALUATIONS

Queen's University

Student ratings are the most commonly used method of evaluating teaching. Reviews of the extensive body of research on student evaluation of teaching generally conclude that student ratings tend to be reliable, valid, relatively unbiased, and useful. Murray (1994) selects from the research the following key findings:

a) Student evaluations are generally consistent across raters, rating forms, courses and time periods for a given instructor.

b) Student evaluations correlate moderately to highly with evaluations made of the same instructor by independent observers.

c) Student evaluations correlate significantly with various objective indicators of student performance, such as student performance on standardized examinations.

d) Student evaluations have generally low correlations with extraneous factors such as class size, severity of grading, etc.

Those conclusions are the basis for the implementation of a university-wide student rating procedure at Queen's, using the QUEST (now USAT) form. However, QUEST/USAT results require careful interpretation if they are to be fairly and effectively used for the assessment and improvement of teaching.

The following guidelines are intended for anyone attempting to interpret the results of the QUEST/USAT form or to use these results for administrative purposes. It is important to understand how student, course, or teacher characteristics can influence student evaluations; individually the effects of these characteristics on ratings are small, but the combination of several factors can have a significant effect. The purpose of the guidelines is to ensure that factors affecting student evaluations are seriously considered in interpreting the results. Some of these guidelines indicate ways to reduce the likelihood of bias, while others reflect the fact that one must always be sensitive to the potential for bias, and allow that sensitivity to inform one's interpretation and use of the results.

1. *Complement results from student evaluations with data from other sources.*

Since no one source of evidence about teaching is definitive, multiple sources of data (peer observation, self-assessment, measures of student achievement, for example) should be used wherever possible, as is done in the compilation of a teaching dossier.

2. *Use several sets of evaluation results.*

- For decisions about the overall teaching effectiveness of an instructor, review evaluation results from different courses and different types of courses (lectures and seminars, graduate and undergraduate, for example) taught over an extended period of time. Results from at least 5 classes is best, including ratings from two or more courses from three or more terms.

- For course improvement, multiple evaluations of a single course over time and/or with different instructors can be useful.

3. *Ensure that a sufficient number of students have responded -- the absolute number and the*

proportion of students responding are both important.

- The response rate from each class should be at least 65% to ensure that results are representative.

- Treat data from fewer than 10 students with caution; data from 10 to 15 students are acceptable; with 25 students, reliability estimates are excellent. When results are only available from very small classes, combine data from several classes to improve reliability.

4. *Consider course, student, and teacher characteristics.*

- Results of student evaluations need to be interpreted in context. While the effect of course, student, and teacher characteristics on results of student evaluations does not always represent bias, complete disregard for the impact of such factors may lead to a misleading assessment of teaching effectiveness. Marsh (1987) found that several factors combined could account for as much as 15% of the variance in student ratings he analyzed.

- Characteristics known to have an impact, and which may be a source of bias, include:

  Gender. There is some evidence that female instructors may receive lower ratings from male students and higher ratings from female students. The ratings of male instructors do not seem to be affected by the gender of the student.

  Class size. There is a slight tendency for smaller classes to receive higher ratings.

  Course level. Higher level courses, especially graduate courses, tend to be rated more positively.

  Discipline. Mathematics, natural science and engineering courses tend to be rated more negatively. In male-dominated disciplines, female instructors tend to be rated more negatively than male instructors, especially by male students.

  Required or elective. Elective courses (and the teachers of these courses), especially courses in the major field, tend to receive higher ratings than do required courses.

  Difficulty level of course. Within disciplines, teachers of courses perceived to be more difficult tend to be rated more highly. Note that courses taught by women tend to be perceived as "easier" than courses taught by men.

  Race. No studies have been reported that investigate systemic racial bias in student ratings. Centra (1993) proposes that results of gender studies suggest ways that race may affect ratings, for example a class of same-race teacher and students may result in a somewhat higher rating than one where race differs. Ng's (1994) experience as a minority female faculty member suggests that minorities in positions of authority often have their knowledge and expertise questioned.

- Other contextual factors with important consequences for student ratings are:

  Course topics. Courses (and instructors) that explicitly challenge students' assumptions about certain topics, especially sensitive or controversial topics such as gender, race, or sexual orientation, can generate resistance among students, which may be reflected in negative ratings.

  Teaching style. The questions used on the rating form, and the students completing the evaluation, generally make assumptions about the nature of good teaching -- although the QUEST/USAT form attempts to avoid such limitations by having very general questions). The ratings of instructors with a teaching style that matches the assumptions of the students and the form are easier to interpret than ratings of teachers with a less conventional style. Instructors with unconventional assumptions/styles may wish to gather additional data, for example by use of optional questions. Persons interpreting results should be sure to review all data, and not restrict their view to overall effectiveness ratings.

  An additional concern is that students' assumptions about the nature of good teaching can be influenced by gender role expectations. For example, female teachers are generally rated more highly than males in terms of their interpersonal skills, yet women's high interpersonal skills ratings do not translate into higher "overall" ratings and women with lower rated interpersonal skills may receive quite poor "overall" ratings.

  Another caveat concerning teaching style is that course characteristics, especially class size, affect the appropriateness and effectiveness of an instructor's preferred teaching style. Low ratings may indicate a mis-match rather than an absolute measure of ineffectiveness; while it is generally desirable that instructors adapt style to situation, there are limits to this principle in practice.

5. *Use good, fair judgement and remain open to situational explanations for ratings results*.

- Individual instructors may receive high student ratings in a situation that the research literature suggests should result in poor evaluations. For example, a female instructor teaching a large required course in a male dominated discipline may be rated very highly. It is tempting to conclude that such cases refute the literature and deny the usefulness of interpretive guidelines. Alternatively, one might conclude that the instructor has responded admirably to a challenging teaching situation. High student ratings may also reflect a high level of departmental support for the instructor and/or the course.

- Do not give undue weight to any single one of the above factors, or to other factors -- such as time of day when class was taught or leniency in grading -- which have no apparent effect on student ratings, *unless* an instructor provides evidence in his or her self-report for the influence of these variables; if you or others have such evidence, that evidence should be taken into consideration.

- Above all, remember that student ratings are only one source of data.

LITERATURE CITED

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

Marsh, H.W. (1987) Student evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research,* 11, 253-288.

Murray, H. (1994). *The impact/consequences of student ratings of teaching*. Presentation at the CSSHE Annual Meeting, Calgary, June 16-18, 1994.

Ng, R. (1994). Sexism and racism in the university: Analyzing a personal experience. *Canadian Women Studies* 14 (2), 41-46.

LITERATURE REVIEWED

Basow, S. (1995). *Student evaluations of college professors: When gender matters*. Unpublished manuscript. Lafayette College, Psychology Department.

Basow, S. and Silberg, N.T. (1987). Student evaluations of college professors: Are male and female professors rated differently? *Journal of Educational Psychology*, 79 (3), 308-314.

Cashin, W.E. (1990). *Student ratings of teaching: Recommendations for use*. IDEA Paper No. 22. Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

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

Feldman, K.A. (1993). College students' views of male and female college teachers: Part II--Evidence from students' evaluations of their classroom teachers. *Research in Higher Educatio*n, 34 (2), 151-191.

Franklin, J. and Theall, M. (1994). *Student ratings of instruction and sex differences revisited*. Tucson, Arizona: University of Arizona, Test Scoring and Evaluation Services.

Marsh, H.W. (1987) Student evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research,* 11, 253-288.

- **From the Report of the SCAD Sub-Committee on Teaching Performance, November 1995.**

- **Distributed by the Instructional Development Centre**.