Evaluation of Teaching at the U of A

Report of the Sub-committee of the Committee on the Learning Environment (CLE)

Last revised: January 9, 2009
(Final changes: preamble, executive summary, recommendations)

Members of the Sub-Committee

Heather Kanuka and Paula Marentette (co-chairs)
John Braga
Katy Campbell
Steve Harvey
Robert Holte
John Nychka
Dan Precht
Daphne Read
Chris Skappak
Connie Varnhagen

EXECUTIVE SUMMARY

A review of the literature conducted for this report revealed that the research on student ratings is immense, and shows inconsistency exists in all areas of teaching evaluation. The reasons are varied, for example: (1) each study seeks to answer different questions with different instruments; (2) there is extreme diversity of research methodologies used; and (3) there is considerable variance on the quality of the studies, especially with respect to sample size.

This being said and acknowledged, within the educational research community, conclusions have been made. The most consistent conclusion is that "properly constructed, appropriately administered, and correctly interpreted student rating can be valid and reliable measures indicating the quality of teaching" (Arreola, 2007, p. 98). Among educational researchers, there is some consensus that when USRI instruments have undergone rigorous psychometric and statistical procedures it results in valid and reliable ratings.

Acknowledging the limitations of generalizing the literature to the UofA, this committee has concluded that a professionally developed instrument with appropriately established metrics can result in valid and reliable teaching evaluation instrument. Although originally based on an evaluation system imported from the University of Michigan, the validity and reliability of the USRI currently in use at the University of Alberta is unknown and needs to be revisited.

The recommendations are as follows.

1. The purpose of the USRI needs to be determined:
   - Is it to improve teaching at the University of Alberta?
   - Is it to provide data for evaluating teaching for FEC?

2. USRI instrument
   a) The use and administration of the USRI (or equivalent instrument) needs be considered in a broader context. Specifically, a teaching evaluation instrument (with proper metrics) should be used in a broader context within course and program evaluation (for examples, see Appendix D from Australia and the UK).
   b) If a decision is made to continue with the administration of teaching evaluation instruments (i.e., the USRI), based on our review of the literature we recommend that a professionally developed instrument be created by an expert in this area to ensure validity and reliability.

3. Multi-faceted Evaluation
   The USRI is designed to be a part of a broader teaching evaluation. Chairs, Deans, Supervisors and Faculty continue to struggle with this in FEC (see Appendix A). As per GFC policy, we need an accompanying set of possibilities and/or examples to be used as a guide for facilitating effective multi-faceted evaluation.

4. GFC Policy
   Quite simply, existing policy is in need of updating.

PREAMBLE

The creation of this task force arose out of a need to better understand teaching evaluations. This need was identified by the members of the Committee of the Learning Environment. The Committee on the Learning Environment (CLE) is responsible for making recommendations concerning policy matters and action matters. The overarching purposes of this task force are to (1) examine alternative teaching evaluation instruments, (2) review existing policies for teaching evaluations at the UofA, as well as at other universities, and (3) provide recommendations for improving teaching evaluations to the CLE committee members.

The Teaching Evaluation subcommittee focused on the following:

- Review existing research on university teaching evaluations
- Review existing instruments on university teaching evaluations at other (similar) Universities (e.g., North America, Australia, UK)
- Review existing teaching evaluation policies at other (similar) Universities (e.g., North America, Australia, UK)
- Building on the information gathered on policies, instruments and the literature, develop recommendations for improving teaching evaluations at the UofA
- Disseminate findings and recommendations to the CLE committee

This subcommittee is advisory to the CLE committee members.

The report that follows is comprised of:

- Part 1: a review of the research literature on teaching evaluations instruments and policies
- Part 2: a description of the policy context for research evaluation at the University of Alberta
- Part 3: a series of recommendations
- Appendix A: Evaluating teaching at FEC
- Appendix B: History of student rating systems at the University of Alberta,
- Appendix C: A review of existing instruments on university teaching evaluations and policies in North America
- Appendix D: A review of existing instruments on university teaching evaluations and policies in Australia and United Kingdom

REVIEW OF THE RESEARCH LITERATURE, TEACHING EVALUATION INSTRUMENTS, AND POLICIES

Faculty Evaluation: Fast, Fair, Cheap—pick any two (Arreola, 2007)

A subtle but significant shift over the last decade has occurred in the social and economic paradigm within which institutions of higher education must operate (Arreola, 2007). There is a greater demand for accountability for what is being taught by the public and what is being learned by the accreditation bodies, increasing diversity of the student population, changing student attitudes and expectations, and a growing expectation by the governments that universities will assume greater responsibility for funding their own operations. Perhaps more importantly, the value of a university degree has been questioned (see, for example, "Declining by Degrees" online at http://www.decliningbydegrees.org/).

At the root of these changes is a shift from evaluating university processes to evaluating outcomes. The question universities must be able to respond to has changed from "What resources do you have and what are you doing with them?" to "How much have your students learned and what can they actually do as a results of the experience?" (National Centre for Higher Education Management Systems, online at http://www.nchems.org/). This shift has also impacted how universities treat their students. Specifically, the traditional practice of "filtering" students is changing from a preferred function of the university (where only the most persistent, talented and brightest survive) to an instructional/teaching function whereby universities must provide accountability for their completion rates and their graduates must be at a specified level of employable competence. The latter part of this expectation is of particular importance, as it requires universities to have graduates that are not only knowledgeable within their field of study, but also capable practitioners. This requires stepping up the learning transactions from simple "transmitters of information" to include learning activities that require the students actually do something with this information, and to do it with competence.

These shifts have led to a need for the University of Alberta to create a taskforce to function as an advisory body that will provide recommendations about whether the current teaching evaluation process (the USRI instrument and associated teaching policy for FEC) is sufficiently rigorous and whether the USRI instrument is measuring what it should be measuring. Also important is to provide recommendations on whether the USRI instrument is consistent with what the Academic Plan purports as effective teaching and learning.

Part 1: Review of the Research Literature

Scope of the Literature Review

Students' evaluation of teaching is not a new phenomenon. Students' feedback on instruction has been collected since the 1920s (Arreola, 2007; Doyle, 1983 as cited by d'Apollonia & Abrami, 1997). More recently, student evaluations of teaching have become much more common. McDaniel (2006) notes there has been an increase from 29% to 86% of postsecondary institutions using student evaluations over the last decade in the United States. With the increasing frequency of evaluations, the body of research on this topic grows. It is "probably...the most thoroughly studied of all forms of personnel evaluation, and one of the best in terms of being supported by empirical research" (Marsh, 1984, p. 749 as cited by Gump, 2007, p. 55). Similarly, Cashin (1998) states that by 1988 over 1,300 articles and books had been published on this topic (as cited by Ory, 2001), and Centra (2003) points out that the ERIC clearinghouse lists over 2,000 studies on this topic.

In preparing this literature review, over 125 articles from academic journals in one database (Academic Search Complete), which had been published since 2000 using the search subject "student evaluation of teachers" were found. A second database (Education Fulltext) search produced another 105 items. For this review 35 articles, published since 2000 were used. This literature review is by no means an exhaustive study of the research; however, it is sufficiently deep enough to identify key issues in the literature.

Identifying Themes

A number of recurring themes can be found in the literature about student evaluations of teaching, which is also known as SET, student rating of instruction (SRI), or student instructional ratings. Wachtel (1998) identified five areas of research: "(1) characteristics related to the administration of evaluations (e.g., anonymity of ratings, timing of evaluations, presence of the instructor during evaluation); (2) course characteristics (e.g., class size, selectivity); (3) instructor characteristics (e.g., gender, reputation); (4) student characteristics (e.g., age, expectations, prior subject interest); and (5) reaction to the use of evaluations (e.g. by faculty or students)" (as cited by Gump, 2007). Ali and Sell (1998) identified 14 questions in their literature review on student ratings of instruction (see Table 1). In this literature review, the following themes were examined:

- Validity of results
- Bias in evaluations
- Can students effectively measure quality of teaching?
- Need for effective tools
- Correlation between higher grades and higher ratings
- Impact on quality of teaching
- Evaluating faculty for tenure and promotion

TABLE 1: COMMONLY EXPRESSED QUESTIONS REGARDING STUDENT RATINGS OF INSTRUCTION (ALI & SELL, 1998)

1. Are student rating forms reliable and valid?
2. Are student rating schemes only a popularity contest?
3. Are [those faculty] with excellent publication records and expertise the only persons qualified to teach and to evaluate their peers' instruction?
4. Do grades or marks students receive in the course affect ratings of the course and the instructor?
5. Does the immaturity and level of experience of students preclude their being able to make consistent judgments about the instructor and instruction?
6. Are students able to make accurate judgments prior to having been away from the course, and possibly the university, for several years?
7. Does class size affect student ratings?
8. Does gender of the student and/or instructor affect student ratings?
9. Does the level of the course (1st year, 2nd year, etc.) affect student ratings?
10. Does the rank of the instructor (instructor, assistant professor, associate professor, professor) affect student ratings?
11. Does student workload affect student ratings?
12. Can student ratings be used meaningfully to improve instruction?
13. Does the ideology or value system of the instructor affect student ratings?
14. What impact does the publication of student ratings have on course selection, quality of instruction or instructors?

Validity of Results

The question of whether student evaluations of teaching produce valid results has been considered by numerous researchers. Greenwald's (1997) four different validity concerns highlight some of the issues they have raised:

1. Conceptual structures: Are ratings conceptually unidimensional or multidimensional?
2. Convergent validity: How well are ratings measures correlated with other indicators of effective teaching?
3. Discriminant validity: Are ratings influenced by variables unrelated to effective teaching?
4. Consequential validity: Are ratings results used in a fashion that is beneficial to the educational system? (p. 1185)

In response to these types of concerns, Ory (2001) identified five types of research that have been conducted to evaluate the validity of student evaluations of teaching: "multisection, multitrait-multimethod, bias, laboratory designs, and dimensionality" (p. 8). He concludes, based on these various types of research, that evaluations can be valid. Greenwald (1997) agrees that based on examining historical literature from the past 25-year period "more publications favored validity than invalidity" (p. 1182). d'Apollonia and Abrami (1997) further affirm that "across different students, courses, and settings, student ratings are consistently valid" (p. 1203). Marsh and Roche (1997) conclude that "SETs are significantly and consistently related to ratings by former students, students' achievement in multisection validity studies, teachers' self-evaluations, and extensive observations of trained observers on specific processes such as teachers' clarity. This pattern of results supports their construct validity" (p. 1188).

Of course, other researchers have presented different perspectives. Pounder (2007), for example, identifies three types of factors that influence SET ratings: (1) student related factors (gender, academic level and maturity, punishing teachers for low grades), (2) course related factors (grading, class size, course content), and (3) teacher related factors (gender, age, experience and rank, teachers' influencing tactics, teachers' behavioural traits). Merritt (2008) also disagrees about the validity of assessments: "evaluations collected from students after no more than five minutes exposure to a professor accurately predict assessments gathered at semester's end" (p. 239).

And, finally, some researchers remain undecided. Gump (2007), for example, examined the literature related to the leniency hypothesis. He concludes that: "there is little consensus on the validity of SETs, the perceptions of this validity (or lack thereof), and the ways in which SETs should (or should not) be used by the various constituents involved" (p. 65).

Bias in Evaluations

"Bias exists when a student, teacher, or course characteristic affects the evaluations made, either positively or negatively, but is unrelated to any criteria of good teaching, such as increased student learning" (Centra & Gaubatz, 2000 as cited by Centra, 2003, p. 498). The question of bias continues to be a

persistent issue as multiple studies support each side. Tables 2 and 3 provide examples of some of the factors researched in relation to student evaluations of teaching.

Some research has provided evidence that bias can arise from characteristics of the student, the course and the teacher (e.g., Pounder, 2007). Typically, these studies focus on characteristics of the instructor including the English-language training see (Üstünlüoglu, 2007), the amount of physical touch demonstrated by the instructor (Lannutti, Laliker, & Hale, 2001), the amount of self-disclosure shared by the instructor (Lannutti & Strauman, 2006), the age and experience of instructors (Blackhart, Peruche, DeWall, & Joiner, 2006; Davidovitch & Soen, 2006), nonverbal mannerism which reflects gender, race, and cultural background (Merritt, 2008), and gender (Laube, Massoni, Sprague, & Ferber, 2007). Likewise, using data from an unendorsed online evaluation site (ratemyprofessors.com), researchers found that there are "strong positive correlations between Quality and Easiness and between Quality and Hotness" (Felton, Koper, Mitchell, & Stinson, 2008, p. 54) suggesting that the same factors may influence scores on officially sanctioned teaching evaluations.

Alternatively, some research explored other factors that can introduce bias. For example, McPherson (2003) uses a statistical model based on data collected over 18 semesters from almost a thousand courses in a department of economics. He concludes the characteristics of the course influence evaluation scores: "level of the class, time at which the class meets, the level of experience of the instructors, and the class size are all found to be significant determinants of SET scores" (p. 1). Pounder, Youmans and Jee (2007) found that evaluation results can be influenced by outside factors related to the administration of the evaluation. In this case, they found that offering students chocolate before they completed their evaluation resulted in higher ratings. d'Apollonia and Abrami (1997) echo the need to have clear procedures which are consistently applied in administering evaluations. From this perspective, Pounder's (2007) literature review concludes,

> most studies have called into question the value of the SET system. It seems that there are so many variables unrelated to the actual execution of teaching influencing SET scores that they tend to obscure accurate assessment of teaching performance. Equally, SET research has generally failed to demonstrate that there is a concrete relationship between teaching performance and student achievement. (p. 186)

Alternatively, some studies find that there is little or no bias in evaluations. For example, Ali and Sell (1998) conclude their literature review by stating, "the literature clearly demonstrates that student rating forms that are psychometrically sound, are reliable, valid, relatively free from bias, and useful in improving teaching" (1998, Concluding comments section, ¶ 2). Similarly, Smith, Yoo, Farr, Salmon, and Miller (2007) studied whether the sex of instructors and students influenced student evaluation scores. While they found that female teachers scored significantly higher on all areas of the teaching evaluation, a factor analysis found that less than one percent of the variance was explained by gender.

Table 2: Summary of Published Research Findings:
Factors Determining Instructor Evaluation Scores
(Chonko et al., 2002, p. 272)

Authors and year	*Factors considered in instructor evaluations scores*
Painter & Granzin, 1972	Communication skills
Tauber, 1973	Perception of fair grading
Kerin, Peterson, & Martin, 1975	Enthusiasm and subject knowledge
Ross, 1977	Personality
Aleamoni, 1981	Class size, grades, grade expectations
Glass, McGaw, & Smith, 1981	Class size
Homan & Kremer, 1983	Student attitudes
Marsh, 1984	Class size, grades, grade expectations
Cardy & Dobbins, 1986	Instructor traits: Warmth, supportiveness, and personality
Miller, 1987	Test frequency
Scherr & Scherr, 1990	Expected grade
Goldberg & Callahan, 1991	Class standing
Langbein, 1994	Faculty traits, overall GPA, hours spent on class, times met with instructor
Tatro, 1995	Grade expectations
Greenwald, 1997	Class size, grades, grade expectations
McKeachie, 1997	Class size, grades, grade expectations
Williams & Ceci, 1997	Instructor traits: warmth, supportiveness, and personality
Bergman & Dobie, 1999	Accessibility
Clayson, 1999	Instructor traits: warmth, supportiveness, and personality

However, there are also some factors that influence results that may be directly linked to the quality of teaching, which would mean they are not biasing factors. For example, Cohen (2005) conducted a principal component analysis and a smallest space analysis, and found that the following factors influence ratings: the course, the instructor, and the interaction between the course or instructor and students. Similarly Remedios and Lieberman (2008) explain,

> students' ratings of courses are largely determined by the degree to which they feel involved, as measured by the extent to which they find their courses stimulating, interesting and useful. In turn, this sense of involvement largely depends on how well students thought a course was organized and taught. Factors such as grades and course difficulty seemed to play at most a very small role. (p. 110)

Similarly, Dziuban, Wang and Cook (n.d.) found that facilitation and communication skills, and the ability to make students feel supported are the most important predictors of ratings. Schrodt et al. (2008) found that power relationships in the classroom impact student evaluations of teaching. These researchers found that:

> When college instructors discuss current theory and research in the classroom, deliver clearly organized lectures, and demonstrate an advanced knowledge in the content area of the course, such behaviors positively predict higher teaching evaluations. At the same time, instructors who use expert power simultaneously sharing personal stories with students, relating to them in ways that are open and approachable, and generally identifying with students' perspectives are even more likely to reap the benefits of such behaviors by receiving higher evaluations. (p. 195)

While these behaviours are examples of using referent and expert power, they also demonstrate characteristics of effective teaching.

Table 3: Overview of Relationships Found Between Students' Ratings and Background Characteristics

Background characteristic	*Summary of findings*
Prior subject interest	Classes with higher interest rate classes more favourably, although it is not always clear if interest existed before the start of the course or was generated by the course or the instructor.
Expected grade – actual grade	Class-average grades are correlated with class-average students' evaluations of teaching, but the interpretation depends on whether higher grades represent grading leniency, superior learning, or pre-existing differences.
Reason for taking a course	Elective courses and those with a higher percentage of students taking the course for general interest tend to be rated higher.
Workload-difficulty	Harder, more difficult courses requiring more effort and time are rated somewhat more favourably.
Class size	Mixed finders but most studies show smaller classes are rated somewhat more favourably although some find curvilinear relationships where large classes also are rated favourably.
Level of course or year in school	Graduate-level courses are rated somewhat more favourably; weak, inconsistent findings suggest upper division courses are rated higher than lower division courses.
Instructor's rank	Mixed findings but little or no effect.
Sex of instructor or student	Mixed findings but little or no effect.
Academic discipline	Weak tendency for higher ratings in humanities and lower ratings in sciences, but too few studies to be clear.
Purpose of ratings	Somewhat higher ratings if ratings are known to be used for tenure-promotion decisions.
Administrative conditions	Somewhat higher if ratings are not anonymous and the instructor is present when ratings are being completed.
Students' personality	Mixed findings but apparently little effect, particularly because different "personality types" may appear in somewhat similar numbers in different classes.

Note: Particularly for the more widely studied characteristics, some studies have found little or no relation or even results opposite to those reported here. The size, or even the direction, of relations may vary considerably, depending on the particular component of students' ratings that is being considered. Few studies have found any of these characteristics to be correlated more than .30 with class-average students' ratings, and most relations are much smaller.

Can Students Effectively Measure Quality of Teaching?

Another important issue is whether or not students have the skills and knowledge to measure the quality of teaching they receive. There seems to be general agreement that students are capable evaluators since their ratings correlate with those of peer evaluators and trained evaluators (Marsh & Roche, 1997). Similarly, Dolmans, Janssen-Noordman and Wolfhagen (2006) studied students' evaluation of tutors in problem-based learning tutorials. They found that students "are not only able to distinguish between poorly and excellently performing tutors, but are also able to distinguish between tutors with different deficiencies" (p. 159).

However, the research in this area has conflicting results. Chonko, Tanner and Davis (2002), for example, surveyed freshmen in their eighth week of introduction to business course in their first semester of college. These researchers found that the students' expectations of a good instructor did not necessarily reflect all of the qualities of good instruction. They conclude that students' evaluations may be more focused on qualities that make a course appealing than on qualities that will ensure students are learning skills that will benefit them in the long term. These authors fear that a customer service model of education is detrimental for students and the quality of instruction they receive.

A related question asks whether students should be involved in evaluating instructors. The large percentage of universities that incorporate some form of student evaluation appears to signal agreement on this issue. For example Greenwald and Gillmore (1997) state that, despite questions about validity, student ratings of teaching are the most readily available and least expensive form of evaluation, and therefore, should continue to be used. A recent study affirms that students also agree they should have a role in evaluating instruction. In their study of student, faculty and administrators' perceptions of student evaluations, Campbell and Bozeman (2008) identified three conclusions about students' beliefs about their role in this process: "The large majority of community college students strongly believe that students should complete formal evaluations of their instructors;" ... "administrators should inform faculty about the ratings;" and "a summary of the results should be available online" (pp. 18-19). In addition, "participants believe that students, in general, take the process of evaluating their instructors seriously, that student surveys are a valuable method of evaluating instructors, that students provide fair evaluations of their instructors, and that students know the qualities of effective teachers" (p. 19). The same study found that most students believe their ratings do not have a significant effect on "a teacher's grading system, dismissal or promotion status, or salary increases" (p. 20).

Need for Effective Tools

In order for students to effectively evaluate instruction, they need effective tools. Some current evaluation tools have been criticized for focusing on the negative aspects of teaching. Zimmerman (2008), for example, points out that evaluation questionnaires tend to "manipulate students into providing negative responses, we encourage them to cast about for some negative remark, any negative remark" (2008 ¶ 7). He also questions that students submit anonymous feedback so they do not have to be accountable for their comments. Moreover,

popular evaluation tools tend to focus on specific aspects of teaching. Kolitch and Dean (1999), for example, found that "the majority of items emphasized an information transmission model for teaching, leaving them to conclude most instruments narrowly define an 'effective' course" (as cited by Dziuban et al., n.d., p. 2). Marsh and Roche (1997) also argue that many "homemade" (p. 1188) evaluation tools do not address the multidimensionality of teaching, so they do not effectively measure the quality of instruction or provide helpful feedback for instructors. Their recommendation is that all evaluation tools should be subject to "rigorous psychometric evaluation" (p. 1188).

In an ongoing debate, Marsh and Roche (1997) and d'Apollonia and Abrami (1997) present opposing sides of the multi-dimensional vs. global evaluation debate. On one hand, Marsh and Roche contend that more attention needs to be paid to "the importance of recognizing the multidimensionality of teaching and SETs in understanding research evidence in relation to the validity, perceived bias, and usefulness of SETs" (p. 1187). They argue that an effective evaluation tool will consider nine different factors: "Learning/Value, Instructor Enthusiasm, Organization/Clarity, Group Interaction, Individual Rapport, Breadth of Coverage, Examinations/Grading, Assignments/Readings, and Workload/Difficulty" (p. 1187). D'Apollonia and Abrami, on the other hand, identify three aspects of teaching that are measured by evaluations: "delivering instruction, facilitating interactions, and evaluating student learning" (p. 1198). They argue that global evaluation ratings are sufficient for making personnel decisions.

Students also need to be motivated to complete evaluations. Chen and Hoshower (2003) found that undergraduate students are most motivated to complete evaluations when the results were used to improve the quality of teaching. Their second motivation was to improve the quality of the course. They were less motivated to participate if their results were used for personnel decisions or to help students choose courses. The authors recommend that evaluations should clearly describe how the results will be used and that students should be made aware of how their results have had an impact.

More recently, many institutions have begun implementing online student evaluation systems. Donovan, Mader and Shinsky (2007) surveyed education students at a college offering both online and traditional, paper-based evaluations. They found that the large majority of students in all categories (undergraduate and graduate, male and female) preferred completing the evaluations online, with 88.4% of respondents indicating they preferred the online evaluation. The majority of students also preferred completing the surveys on their own time rather than going to the computer lab as a class. Students indicated the following reasons for preferring the online evaluations, from most (87.2%) to least common (48.2%) responses: convenience, anonymity, privacy, more time to reflect, able to write more comments, and not taking class time (p. 167). However, students also expressed concerned with security and anonymity on the online system and remembering to complete the evaluations on their own time during the busy weeks at the end of term. These authors also found that students' preference for completing evaluations online increased as they completed more using the online evaluations, but that graduate students, in particular, had fewer opportunities to complete the online evaluations. If online evaluations are used, students need to be given time to complete the evaluations; they need to receive reminders to complete the evaluations; and they need extra support to complete the evaluations the first few times until they become familiar with the format. More information about online evaluations can be found in volume 96 of New Directions for Teaching & Learning (Ballantyne, 2003; Bothell & Henderson, 2003; Bullock, 2003; Hardy,

2003; Hoffman, 2003; Jonhson, 2003; Lleywellyn, 2003; McGhee & Lowell, 2003; Sorenson & Reiner, 2003; Tucker, Jones, Straker, & Cole, 2003).

Correlation between Higher Grades and Higher Ratings

Numerous studies seem to agree there is a correlation between grades and evaluation scores (see, for example, Guinn & Vincent, 2006; Blackhart et al., 2006); however, whether that means that giving students higher marks causes them to rate instructors higher, or students who get higher marks have learned more so they rate their instructor's teaching better (Centra, 2003), or students who do better than they expected are more satisfied with the course, is unclear (Remedios & Lieberman, 2008).

Greenwall and Gillmore (1997) identify five possible theories to explain this correlation: "1. Teaching effectiveness influences both grades and ratings" ... "2. Students' general academic motivation influences both grades and ratings" ... "3. Students' course-specific motivation influences both grades and ratings" ... "4. Students infer course quality and own ability from received grades" and "5. Students give high ratings in appreciation for lenient grading" (pp. 1210-1211). Though it is also noted by Maurer (2006) that another explanation might be cognitive dissonance (i.e. students expect to get a low mark, and in order to justify it to themselves, they blame it on poor instruction).

One theory that has been frequently studied is the leniency hypothesis, which claims that when instructors are more lenient in their marking, students' evaluation scores are higher. SET scores are higher in social sciences where there is more flexibility in grading, proving the connection between grade inflation and higher evaluation scores according to Johnson (2003) (as cited in Felton et al., 2008). However, this theory has been rejected by Marsh and Roche (1997), who identified numerous methodological problems with these studies before concluding, "Whereas a grading-leniency effect may produce some bias in SETs, support for this suggestion is weak, and the size of such an effect is likely to be unsubstantial" (p. 1193). Likewise, McPherson (2003) determined that "instructors cannot 'buy' higher SET scores by awarding higher grades" (p. 1). Similarly Gump (2007) finds that there are many contradictory studies on this hypothesis, reflecting problems with researchers' bias, limited information about sample size, and nonexperimental research methodologies. In fact, there is ambiguity in the research about whether instructors perceive the leniency hypothesis as valid, and how that perception might influence their teaching.

Centra (2003) also rejects the leniency hypothesis. He analyzed results from over 50,000 college courses offered over a 5-year period, examining the relationships between expected grades and level of difficulty/workload, and course evaluation ratings. He found a smaller correlation between expected grade and rating than has been reported in other studies. He concluded that students give the highest ratings to courses that they label as "just right" in terms of level of difficulty/workload, which he claims affirms the fact that instructors who are aware of their students' skill levels will be rated as most effective.

Impact on Quality of Teaching

In a recent editorial, Zimmerman (2008) questions whether evaluations provide instructors with any useful information. He points out that evaluation

questionnaires tend to "manipulate students into providing negative responses, we encourage them to cast about for some negative remark, any negative remark" (¶ 7). Other studies confirm that faculty members question the effectiveness of these evaluations. For example, Campbell and Bozeman (2008) found that faculty members did not feel that evaluations had a significant impact on their teaching. Similarly, administrators believed that student valuations only "were marginally effective" in improving teaching (pp. 21-22). The faculty members interviewed felt that institutional systems were not in place to help them use these results effectively, while the administrators felt that faculty members, especially fulltime ones, need to use self-reflection to improve their teaching. Other studies confirm that very few instructors (2.5 to 10.3%) made changes in their teaching as a result of the evaluations (Nasser and Fresko, 2002) and that those who attended academic development workshops do not receive significantly higher scores (Davidovitch & Soen, 2006).

Swain (2006) points out that all teaching evaluation programs need to have a plan in place for using teaching evaluation feedback. It should include details on the analysis, what to do with results, who takes action, keeping students informed of what has happened, letting students know their comments make a difference and showing them the impact of their comments. Likewise, instructors who receive evaluation reports plus other interventions (i.e. resources, consultation) receive higher ratings on subsequent reports than instructors who receive only their results (Marsh & Roche, 1997). Marsh and Roche also recommend providing targeted intervention to address specific dimensions identified in the evaluation.

One example of a resource developed to assist instructors is the strategy and resources guide developed at Brigham Young University. It is a thorough list of resources and strategies for faculty members to improve their teaching. The information is organized to address each item in the teaching evaluation. For each item, the following information is provided: the evaluation item is explained in more detail, a list of strategies and resources are provided, and each strategy is examined in detail with specific advice and referrals to related resources (Clark, Johnson, Sorenson, Birch, & Bradley, 2007).

Evaluating Faculty for Tenure and Promotion

Student evaluations of teaching are often used in making decisions about tenure and promotion: "Seldin (1993) noted an 86 per cent use of the student evaluation of teaching (SET) as a central feature of personnel decisions in US higher education" (Pounder, 2007, p. 178). Interestingly, some research finds that institutions that value teaching the most tend to make less use of student evaluations:

> Read, Rama, and Raghunandan (2001) surveyed a large number of accounting departments to clarify the importance placed on teaching effectiveness and the reliance on student ratings when making promotion and tenure decisions. They discovered an inverse relationship. Institutions that de-emphasized teaching importance gave more credence to student ratings than institutions that emphasized the importance of teaching. (Dziuban et al., n.d., p. 5)

The use of student evaluations "has been a source of contention since the practice was introduced" (Felton, Mitchell, & Stinson, 2004, p. 46). Many faculty members express concerns about validity and bias in the evaluations, about who

has access to results and how they are used, and whether students are qualified to evaluate teaching (Cutler, 2007; Gray & Bergmann, 2003; Gump, 2007; McDaniel, 2006; Ory, 2001). In an editorial, McDaniel (2006) poses a number of questions shared by other instructors:

- are students qualified to judge the quality of a professor's pedagogy and academic expertise?
- are students evaluating teaching effectiveness – or something else?
- are faculty rights to academic freedom compromised by the pressures to secure favorable student evaluations?
- are administrators using student evaluations to intrude on the privacy of the classroom and to manipulate faculty behavior?

The opposing perspectives of students and faculty on this issue cause further complications. While students favour publishing teaching evaluation results to improve accountability, faculty felt publishing them would lower standards (Howell & Symabluk, 2001 as cited by Dziuban et al., n.d.).

While administrators believe that teaching evaluations are only "marginally effective" in improving the quality of teaching, these evaluations are still important tools in making decisions about employment (Campbell & Bozeman, 2008, pp. 21-22). Similarly, D'Apollonia and Abrami (1997) state, "student ratings should be used to make only crude judgments of instructional effectiveness (exceptional, adequate, and unacceptable)"(p. 1205), a statement with which and McKeachie (1997) agrees.

Researchers also disagree about which results should be used and how they should be analyzed. D'Apollonia and Abrami (1997) argue that a global rating "or a single score representing a weighted average of the specific ratings" should be used (p. 1203) while Marsh and Roche (1997) argue that a multidimensional approach should be used, though they also note a weighted average could achieve the same results (d'Apollonia & Abrami, 1997). In response to critiques, McPherson (2003) suggests that adjusting scores to account for variables outside of the instructors control like time of day, class size and class level might be an alternate way of looking at results. Considering the issue from another perspective, McKeachie (1997) argues that the problems associated with student evaluations of teachings are not reflective of problems with the evaluations but of problems with the interpretation and application of their results. He identifies two different problems: (1) personnel committees do not believe student evaluations are credible, so they are not given enough attention; (2) personnel committees try to compare instructors using evaluation data without considering factors which legitimately influence evaluations, for example, "differences in goals, teaching methods, content, and a myriad of other variables" (p. 1222). However, there has been little research done on how committees use results in their decision-making processes.

Summary of Results

The literature on student ratings is immense, and this committee's review of the research shows inconsistency exists in all areas of teaching evaluation. One reason is that each study is trying to answer a different question. Where one is measuring how a single factor, like expected grades, influences evaluation scores for a single instructor over a few terms (Maurer, 2006), another study is examining the relationship between multiple factors in the students, instructors and course with a sample of over 50,000 graduate and undergraduate courses offered over a five-year period (Centra, 2003). Likewise, one study uses multiple regression analysis (Blackhart et al., 2006) while another uses decision trees (Dziuban et al., n.d.) to arrive at their conclusions. Different studies also use different instruments, each asking slightly different questions (Centra, 2003), which also makes comparing results difficult. This being said and acknowledged, within the educational research community, conclusions have been made. The most consistent conclusion is that "properly constructed, appropriately administered, and correctly interpreted student rating can be valid and reliable measures indicating the quality of teaching" (Arreola, 2007, p. 98).

Among educational researchers, there is some consensus that when USRI instruments have undergone rigorous psychometric and statistical procedures it results in valid and reliable ratings. Specifically, when 'homemade' faculty evaluation tools are not included in literature reviews, there is a fairly strong consensus in the following questions (Arreola, 2007, pp. 100-104):

- Are student ratings a popularity contest?

  No—not if the institution is using a student rating form that has been constructed using professional psychometric procedures and has demonstrated reliability and validity. Well-designed student rating forms carefully measure many different aspects of faculty performance. Alternatively, student rating forms that have not been constructed to professional psychometric standards may be unreliable and, in turn, run the risk of having such factors as popularity, temperature of the classroom, instructor gender, etc., influencing student ratings.

- Aren't student rating forms just plain unreliable and invalid?

  Yes and No. Yes—if the student rating form in use has not undergone rigorous psychometric and statistical procedures. Alternatively, well-developed instruments have been shown to be reliable and valid.

- Aren't students too immature, inexperienced and capricious to make any consistent judgments about the instructor and instruction?

  No—an extensive body of research (going back to the 1920s) shows this commonly held belief is not true.

- Isn't it true that I can buy good student ratings by just giving easy grades?

  No—and there has been more research conducted on this one question than almost any other in the field of student ratings and faculty evaluation. The reason for the numerous studies is not that the question is so difficult to answer, it's that faculty generally don't like the answer the research provides. The answer is that there is no consistent correlation between the grades a faculty member gives and the rating he or she receives when a well-designed student rating form is administered.

- Isn't it generally easier to get good rating in higher level courses?

  Yes—the research has shown that first and second year students tend to rate a course more harshly than third or fourth year students.

- Isn't there a gender bias in student ratings? (Don't female faculty tend to get lower ratings than male faculty?)

  No—there is a fairly consistent body of research showing gender bias does not exist in student rating forms.

- Isn't it true that students who are required to take a course tend to rate the course more harshly than those taking it as an elective?

  Yes—the literature shows that students who are required to take a course tend to rate it lower than students who elect to take it.

- Isn't it more difficult for math and science faculty to get good ratings?

  Yes—research has shown that courses in the math and sciences tend to get lower ratings than in the humanities.

- Don't students have to be away from the course, and possibly the institution, for several years before they are able to make accurate judgments about the instructor and instruction?

  No—collecting data on this issue is difficult to obtain. What limited research on this topic is available, leads to the conclusion that this belief is not generally true.

- Isn't it true that class size affects student ratings?

  No—the literature does not show a consistent relationship between class size and student ratings.

- Does the time of the day the course is taught affect student ratings?

  No—there is not much research is available in this area, but what has been conducted does not show the time of day has any influence on student ratings.

- Do majors in a course rate it differently than non-majors?

  No—all research conducted in this area show there is no significant relationships between students ratings and whether students were majors or non-majors.

- Does the rank of the instructor affect student ratings?

  No—while there are a few conflicting studies on this issue, in general, the research does not support the idea that faculty of higher professorial rank get higher student ratings.

- Do student ratings improve instruction?

  Yes—under the right conditions (e.g. support and resources are provided, including personal consultations by qualified educators).

While not specific to the USRI instrument at the UofA, these findings discredit many myths that are held about the effectiveness of a student rating system.

Conclusions

Acknowledging the limitations of generalizing to the UofA, it seems reasonable to conclude that a professionally developed instrument with appropriately established metrics can result in valid and reliable teaching evaluation.

Although originally based on an evaluation system imported from the University of Michigan, the validity and reliability of the USRI currently in use at the University of Alberta is unknown and needs to be revisited. This committee put forward the three possibilities for dealing with the UofA USRIs:

- Keep the current instrument
- Modify the current instrument
- Discontinue the use of the current instrument

The committee recognizes that USRIs are not the only means of evaluating teaching and must be assessed as one of many methods. The next section provides a larger context for the evaluation of teaching at the University of Alberta.

Part 2: Evaluation of Teaching at the University of Alberta

According to GFC policy on teaching evaluation, (Section 111.2, Teaching Evaluation):

1. Evaluation of teaching at the University of Alberta serves two purposes:

a. Summative - Evaluation provides a review and overview of an instructor's teaching that is an essential element in promotion and tenure decisions. In its summative form, teaching evaluation forms a basis for rewarding excellence, as well as the basis for withholding reward. (GFC 24 NOV 1997)

b. Formative - Evaluation provides helpful feedback to teachers by identifying teaching strengths and weaknesses and, in so doing, giving guidance for the improvement or refinement of teaching skills. (GFC 24 NOV 1997)

2. Evaluation of teaching shall be multifaceted. Multifaceted evaluation shall include the Universal Student Ratings of Instruction set out in Section 111.3 and other methods of assessing teaching designed within the individual Faculties to respond to the particular conditions of that Faculty. Such assessments shall include one or more of the following: input from administrators, peers, self, undergraduate and graduate students, and alumni. (GFC 09 JUN 1995) (GFC 24 NOV 1997)

3. Recognizing that the evaluation of teaching at the University shall be multifaceted, Faculty Evaluation Committee (FEC) decisions concerning tenure, promotion or unsatisfactory teaching performance must be based on more than one indicator of the adequacy of teaching. (GFC 24 NOV 1997)

4. Assessment of teaching involving input from administrators, peers, self, alumni, or undergraduate and graduate students in addition to the Universal Student Ratings of Instruction should occur annually prior to tenure. For continuing faculty (ie, Categories A1.1, A1.5 and A1.6), such assessment will occur at least triennially. (GFC 24 NOV 1997)

5. The University shall continue to support University Teaching Services in its education programming which is focused on the development and improvement of teaching and learning and its efforts to enhance research in university teaching. (GFC 28 APR 1980) (GFC 26 SEP 1988) (GFC 12 OCT 1993) (GFC 24 NOV 1997)

Two points can be made about these policy statements. First, these policy guidelines require that teaching be evaluated in a multi-faceted manner, although the particular methods of assessment are left to individual faculties. How to assess teaching is not obvious. Indeed, in GFC Section 111.1 Teaching and Learning it says "nowhere, in any document, is there a clear and concise statement of what constitutes excellent teaching. It is taken for granted that we all know." Although this statement is followed by an attempt to describe the attributes of good teaching, it is not accompanied by assessment methods for each attribute. From her meetings with the Chairs of each department at the University, Heather Kanuka, Director of UTS learned that Chairs struggle with this mandate (see summary of Chairs' comments in Appendix A).

Second, these statements on teaching evaluation have not been considered by GFC for more than ten years. This means that the present policy and assessment tools, the USRI in particular, were developed before both the Dare to Discover and Dare to Deliver documents. The values articulated in these documents should be reflected in how we assess teaching at this institution.

The evaluation of teaching involves a process of interpreting data through the lens of a set of values to determine whether the data indicate a desirable or undesirable set of conditions. In spite of GFC's policy that "Small differences in evaluation should not be considered meaningful" (111.3, section I), in fact they are because at present we do not have an effective means of conducting multifaceted evaluation. If the teaching evaluation process at the University of Alberta is to be perceived as fair, it is necessary to begin by gaining agreement of the values to be used in the evaluation process. Of necessity, this requires tying the University of Alberta's Academic Plan to the teaching evaluation process, as well as GFC policy on teaching and learning.

The teaching evaluation instrument exists in "recognition of the University's commitment to teaching" (111.3); that is, to ensure quality teaching. Improving teaching therefore becomes the priority. Also from GFC policy, we recognize that since evaluation of teaching was left to individual faculties, the responsibility for improving teaching lies with the faculties.

However, support from administration is essential to the improvement of teaching, as well as support services offered by UTS. Suggestions offered by Aerreola (2007) based on data collected from administration and teaching service units to help Faculties and Departments are as follows:

Integrate faculty evaluation and professional enrichment programs. For every element of the faculty evaluation program there needs to be corresponding professional enrichment resources offered. This will ensure that teaching staff have institutionally supported recourse when the evaluation system detects performance weaknesses.

Use a variety of sources in the evaluation system. The faculty evaluation system needs to use input from a variety of sources including peers, self, administration, as well as students. It is also important to articulate the impact that each of these sources of information has on the total evaluation.

Ensure that the faculty evaluation program is functionally valid. The areas of faculty performance that are being evaluated need to be in agreement with what the faculty and the administration believe ought to be evaluated. The extent to which faculty are either unsure of or disagree with the assumed value structure of the faculty evaluation program, they will consider the program not to be valid and will resist it. Equally important is that the faculty evaluation program is tied to the Academic Plan.

Provide detailed and confidential faculty evaluation information to each instructor. The faculty evaluation must be provided as confidential resources for faculty to use in improving and documenting the quality of their performance. The unit providing the professional development (e.g., UTS) cannot come to be seen as a 'watchdog' agency for the administration.

Establish a facilitative reward structure. Policies need to be established that treat documented professional growth and enrichment efforts in a manner similar to those of publication and research efforts.

Hire an external consultant with expertise in faculty evaluation. Bringing in an outside expert can be invaluable in ensuring that an evaluation program initiative for the University of Alberta will effectively meet the needs of both faculty and administration, and in a manner that is tied to our Academic Plan. There are researchers who have expertise in faculty evaluation and we are recommending that we bring in one of these experts to provide consultation services to review our institutional needs.

Part 3: Recommendations

This committee has deliberated thoughtfully and extensively on how to provide productive recommendations to the Committee on the Learning Environment. Our recommendations fall into four distinct categories: (1) The purpose of the use of the USRI needs to be clarified; (2) the USRI instrument, (3) multi-faceted evaluation and, (4) GFC Policy.

1. The purpose of the USRI needs to be determined:

   - Is it to improve teaching at the University of Alberta?
   - Is it to provide data for evaluating teaching for FEC?

2. USRI instrument

   c) The use and administration of the USRI (or equivalent instrument) needs be considered in a broader context. Specifically, a teaching evaluation instrument (with proper metrics) should be used in a broader context within course and program evaluation (for examples, see Appendix D from Australia and the UK).

   d) If a decision is made to continue with the administration of teaching evaluation instruments (i.e., the USRI), based on our review of the literature we recommend that a professionally developed instrument be created by an expert in this area to ensure validity and reliability.

3. Multi-faceted Evaluation

   The USRI is designed to be a part of a broader teaching evaluation. Chairs, Deans, Supervisors and Faculty continue to struggle with this in FEC (see Appendix A). As per GFC policy, we need an accompanying set of possibilities and/or examples to be used as a guide for facilitating effective multi-faceted evaluation.

4. GFC Policy

   Quite simply, existing policy is in need of updating.

Underpinning these recommendations is the assumption that if the University of Alberta is going to evaluate teaching it has a responsibility to support faculty to improve their teaching development.

References

Arreola, R. A. (2007). Developing a comprehensive faculty evaluation system. A guide to designing, building, and operating large-scale faculty evaluation. San Francisco, CA: Anker Publishing.

Ali, D. L., & Sell, Y. (1998). Issues regarding the reliability, validity and utility of student ratings of instruction: A survey of research findings. Retrieved from University of Calgary: http://www.ucalgary.ca/vpa/USRI/appendix4.html

Ballantyne, C. (2003). Online evaluations of teaching: An examination of current practice and considerations for the future. New Directions for Teaching & Learning, (96), 103-112. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=11744624&loginpage=Login.asp&site=ehost-live&scope=site

Blackhart, G. C., Peruche, B. M., DeWall, C. N., & Joiner, T. E., Jr. (2006). Faculty forum: Factors influencing teaching evaluations in higher education. Teaching of Psychology, 33(1), 37-39. doi:10.1207/s15328023top3301_9

Bothell, T. W., & Henderson, T. (2003). Do online ratings of instruction make $ense? New Directions for Teaching & Learning, (96), 69-79. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=11744627&loginpage=Login.asp&site=ehost-live&scope=site

Bullock, C. D. (2003). Online collection of midterm student feedback. New Directions for Teaching & Learning, (96), 95-102. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=11744625&loginpage=Login.asp&site=ehost-live&scope=site

Campbell, J. P., & Bozeman, W. C. (2008). The value of student ratings: Perceptions of students, teachers, and administrators. Community College Journal of Research & Practice, 32(1), 13-24. doi:10.1080/10668920600864137

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? Research in Higher Education, 44(5), 495-518. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=10691989&loginpage=login.asp&site=ehost-live&scope=site

Chen, Y., & Hoshower, L. B. (2003). Student evaluation of teaching effectiveness: An assessment of student perception and motivation. Assessment & Evaluation in Higher Education, 28(1), 71-88. doi:10.1080/0260293032000033071

Chonko, L. B., Tanner, J. F., & Davis, R. (2002). What are they thinking? Students' expectations and self-assessments. Journal of Education for Business, 77(5), 271-281. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=7214031&loginpage=login.asp&site=ehost-live&scope=site

Clark, S. J., Johnson, T., Sorenson, L., Birch, J., & Bradley, B. (2007). Teaching improvement strategies and resources related to student rating items. Retrieved from Center for Teaching & Learning, Brigham Young University: http://ctl.byu.edu/?page_id=661

Cohen, E. (2005). Student evaluations of course and teacher: Factor analysis and SSA approaches. Assessment & Evaluation in Higher Education, 30(2), 123-136. doi:10.1080/0260293042000264235

Cutler, A. (2007). A night at the circus. Journal of College Science Teaching, 36, 6-7. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=25497742&loginpage=login.asp&site=ehost-live&scope=site

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. American Psychologist, 52(11), 1198-1208. Retrieved from http://ovidsp.tx.ovid.com.login.ezproxy.library.ualberta.ca/spb/ovidweb.cgi

Davidovitch, N., & Soen, D. (2006). Using students' assessments to improve instructors' quality of teaching. Journal of Further & Higher Education, 30(4), 351-376. doi:10.1080/03098770600965375

Dolmans, D. H. J. M., Janssen-Noordman, A., & Wolfhagen, H. A. P. (2006). Can students differentiate between PBL tutors with different tutoring deficiencies? Medical Teacher, 28(6), 156-161. doi:10.1080/01421590600776545

Donovan, J., Mader, C., & Shinsky, J. (2007). Online vs. Traditional course evaluation formats: Student perceptions. Journal of Interactive Online Learning, 6(3), 158-180. Retrieved from http://www.ncolr.org.login.ezproxy.library.ualberta.ca/jiol/issues/showissue.cfm?volID=6&IssueID=21

Dziuban, C. D., Wang, M. C., & Cook, I. J. (n.d.). Dr. Fox rocks: Student perceptions of excellent and poor college teaching [working draft]. Retrieved from http://www.sc.edu/cte/docs/dr.FoxRocks.pdf

Felton, J., Koper, P. T., Mitchell, J., & Stinson, M. (2008). Attractiveness, easiness and other issues: Student evaluations of professors on ratemyprofessors.Com. Assessment & Evaluation in Higher Education, 33(1), 45-61. doi:10.1080/02602930601122803

Felton, J., Mitchell, J., & Stinson, M. (2004). Web-based student evaluations of professors: The relations between perceived quality, easiness and sexiness. Assessment & Evaluation in Higher Education, 29(1), 91-108. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=12511352&loginpage=login.asp&site=ehost-live&scope=site

Ginns, P., Prosser, M., & Barrie, S. (2007). Students' perceptions of teaching quality in higher education: The perspective of currently enrolled students. Studies in Higher Education, 32(5), 603-615. doi:10.1080/03075070701573773

Gray, M., & Bergmann, B. R. (2003). Student teaching evaluations. Academe, 89(5), 44-46. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=10928004&loginpage=login.asp&site=ehost-live&scope=site

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. American Psychologist, 52(11), 1182-1186.

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. American Psychologist, 52(11), 1209-1217. Retrieved from http://ovidsp.tx.ovid.com.login.ezproxy.library.ualberta.ca/spb/ovidweb.cgi

Guinn, B., & Vincent, V. (2006). The influence of grades on teaching effectiveness ratings at a Hispanic-serving institution. Journal of Hispanic Higher Education, 5(4), 313-321. doi:10.1177/1538192706291138

Gump, S. E. (2007). Student evaluations of teaching effectiveness and the leniency hypothesis: A literature review. Educational Research Quarterly, 30(3), 55-68. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=24293912&loginpage=login.asp&site=ehost-live&scope=site

Hardy, N. (2003). Online ratings: Fact and fiction. New Directions for Teaching & Learning, (96), 31-38. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=11744631&loginpage=Login.asp&site=ehost-live&scope=site

Hoffman, K. M. (2003). Online course evaluation and reporting in higher education. New Directions for Teaching & Learning, (96), 25-29. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=11744632&loginpage=login.asp&site=ehost-live&scope=site

Jonhson, T. D. (2003). Online student ratings: Will students respond? New Directions for Teaching & Learning, (96), 49-59. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=11744629&loginpage=Login.asp&site=ehost-live&scope=site

Lannutti, P. J., Laliker, M., & Hale, J. L. (2001). Violations of expectations and social-sexual communication in student/professor interactions. Communication Education, 50(1), 69-82. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=4100573&loginpage=login.asp&site=ehost-live&scope=site

Lannutti, P. J., & Strauman, E. C. (2006). Classroom communication: The influence of instructor self-disclosure on student evaluations. Communication Quarterly, 54(1), 89-99. doi:10.1080/01463370500270496

Laube, H., Massoni, K., Sprague, J., & Ferber, A. L. (2007). The impact of gender on the evaluation of teaching: What we know and what we can do. NWSA Journal, 19(3), 87-104. Retrieved from http://vnweb.hwwilsonweb.com.login.ezproxy.library.ualberta.ca/hww/jumpstart.jhtml?recid=0bc05f7a67b1790ec5a25dd95312f21932b15374917fd048cc5242add570a3d4c4e10f5ea26d892f&fmt=P

Lleywellyn, D. C. (2003). Online reporting of results for online student ratings. New Directions for Teaching & Learning, (96), 61-68. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=11744628&loginpage=Login.asp&site=ehost-live&scope=site

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluation of teaching effectiveness effective: The critical issues of validity, bias, and utility. American Psychologist, 52(11), 1187-1197. Retrieved from http://ovidsp.tx.ovid.com.login.ezproxy.library.ualberta.ca/spb/ovidweb.cgi

Maurer, T. W. (2006). Cognitive dissonance or revenge? Student grades and course evaluations. Teaching of Psychology, 33(3), 176-179. doi:10.1207/s15328023top3303_4

McDaniel, T. R. (2006). Student evaluations of instructors: A good thing? Academic Leader, 22(8), 8. Retrieved from http://vnweb.hwwilsonweb.com.login.ezproxy.library.ualberta.ca/hww/jumpstart.jhtml?recid=0bc05f7a67b1790ec5a25dd95312f2193252051c0ad769747e92fbc0239598149oed29155bf3ecff&fmt=H

McGhee, D. E., & Lowell, N. (2003). Psychometric properties of student ratings of instruction in online and on-campus courses. New Directions for Teaching & Learning, (96), 39-48. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=11744630&loginpage=Login.asp&site=ehost-live&scope=site

McKeachie, W. J. (1997). Student ratings: The validity of use. American Psychologist, 52(11), 1218-1225. Retrieved from http://ovidsp.tx.ovid.com.login.ezproxy.library.ualberta.ca/spb/ovidweb.cgi?&S=GHDIFPGOIDDDDALJNCILJAPLDAPPAA00&Full+Text+Link=S.sh.15.16.18%7c6%7csl_10%7c80%7c2&WebLinkReturn=Full+Text%3dL%7cS.sh.15.16.18.42%7c0%7c00000487-199711000-00005

McPherson, M. A. (2003). Revisiting the determinants of student evaluation of teachers scores. Retrieved from http://ssrn.com/abstract=410934

Merritt, D. J. (2008). Bias, the brain, and student evaluations of teaching. St. John's Law Review, 82(1), 235-287. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=28324297&loginpage=login.asp&site=ehost-live&scope=site

Nasser, F., & Fresko, B. (2002). Faculty views of student evaluation of college teaching. Assessment & Evaluation in Higher Education, 27(2), 187-198. doi:10.1080/02602930220128751

Ory, J. C. (2001). Faculty thoughts and concerns about student ratings. New Directions for Teaching & Learning, (87), 3-15. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=9179128&loginpage=Login.asp&site=ehost-live&scope=site

Pounder, J. S. (2007). Is student evaluation of teaching worthwhile?: An analytical framework for answering the question. Quality Assurance in Education: An International Perspective, 15(2), 178-191. Retrieved from http://www.emeraldinsight.com.login.ezproxy.library.ualberta.ca/Insight/viewContentItem.do;jsessionid=95F40DE105182093429C22368CA811DE?contentType=Article&contentId=1602895

Remedios, R., & Lieberman, D. A. (2008). I liked your course because you taught me well: The influence of grades, workload, expectations and goals on students' evaluations of teaching. British Educational Research Journal, 34(1), 91-115. doi: 10.1080/01411920701492043

Schrodt, P., Witt, P. L., Myers, S. A., Turman, P. D., Barton, M. H., & Jernberg, K. A. (2008). Learner empowerment and teacher evaluations as functions of teacher power use in the college classroom. Communication Education, 57(2), 180-200. doi:10.1080/03634520701840303

Smith, S. W., Yoo, J. H., Farr, A. C., Salmon, C. T., & Miller, V. D. (2007). The influence of student sex and instructor sex on student ratings of instructors: Results from a college of communication. Women's Studies in Communication, 30(1), 64-77. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=25547189&loginpage=login.asp&site=ehost-live&scope=site

Sorenson, L., & Reiner, C. (2003). Charting the uncharted seas of online student ratings of instruction. New Directions for Teaching & Learning, (96), 1-24. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=11744633&loginpage=login.asp&site=ehost-live&scope=site

Tucker, B., Jones, S., Straker, L., & Cole, J. (2003). Course evaluation on the web: Facilitating student and teacher reflection to improve learning. New Directions for Teaching & Learning, (96), 81-93. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=11744626&loginpage=Login.asp&site=ehost-live&scope=site

Üstünlüoglu, E. (2007). University students' perceptions of native and non-native teachers. Teachers & Teaching, 13(1), 63-79. doi:10.1080/13540600601106096

Youmans, R. J., & Jee, B. D. (2007). Fudging the numbers: Distributing chocolate influences student evaluations of an undergraduate course. Teaching of Psychology, 34(4), 245-247.

Zimmerman, B. (2008). Course evaluations - students' revenge? University Affairs. Retrieved from http://www.universityaffairs.ca/issues/2008/february/opinion_01.html

Appendix A

Evaluating Teaching at FEC

> Chairs require information and strategies on how to conduct broad-based and fair evaluation. Many Chairs specifically noted they need help on how to evaluate teaching for FEC; they need information on what to look for in terms of assessing and evaluating good teaching and how to reward accordingly. While some Departments and Faculties use additional measures to evaluate teaching, the majority rely entirely on USRI scores for FEC. Most chairs are aware that (1) USRIs are not a sufficient measure for evaluating teaching for FEC and (2) teaching continues to be undervalued as compared to research productivity. Many chairs also expressed a desire to receive information on other ways to evaluate teaching effectiveness for FEC (e.g., peer observation).

Evaluating Teaching at FEC

- With respect to FEC and the evaluation of teaching: We need help on how to evaluate teaching.
- I have no training in this. Knowing how to look and what to look for would be helpful.
- In our department, we really need to find a balance between punishing and improving.
- Biggest problem with teaching: not enough rewards to recognize great teaching.
- With respect to annual reports: my dean says "more publications – not better teaching."
- There is no information available on what—and how—we should be observing with 'guided observations'. I need help with multi-faceted evaluation.
- The problem with FEC in our faculty is that there is no critical reflection and feedback, and no requirement for faculty to get to know their teaching philosophy.
- A critical problem with FEC is the absence of distinction between 'courses' and 'instructors'.
- Importance given to teaching by FEC at tenure and how it is weighted and evaluated varies across campus. There needs to be consistency through policy.
- In our faculty, merit increments are awarded annually. Department chairs and FEC evaluate service, research, teaching vigorously and teaching research are evaluated symmetrically. We have to have balance. Letters are sent to those faculty members who need to improve.
- Our faculty takes teaching seriously. There is no tolerance for bad teaching and faculty members will lose increments for a track record of bad teaching.
- There is a lack of balance between teaching and research at FEC. A faculty member will never get tenure because he/she is a fabulous teacher. This needs to be addressed.
- Bottom line (while teaching is important) people recognize effort needs to go toward research. I'd say 70% of their time is spent on research efforts. Competition for grants is fierce.
- We do peer evaluations in which we are required to look at course outline, tests, etc. Faculty have a range of responses to this.

- When evaluating their teaching, Faculty are required to respond to written comments and provide contextualization. Simply thinking about their teaching makes all of the difference. Reflection is important. Students need to reflect on their learning as well. We also need to accommodate different styles of teaching.
- In our department, faculty teaching evaluations are color coded, signifying to the Dean and Chair who is having difficulty.
- We don't value teaching as much as we should. We need better metrics on evaluation. We have no good indices
- There has been a shift in FEC to value teaching more, and in theory this is good, but we don't know how to evaluate teaching other than with the IDQs. So now IDQ scores are even more important. I'm not convinced there is a relationship between IDQ scores and good teaching.
- At FEC, there is limited room for growth if your research is not balanced with good teaching.
- We examine USRI comments from the students and look for patterns over time. We also have faculty submit syllabus, teaching materials, and there are group critiques of student work. There is a peer evaluation of untenured faculty in which senior faculty observe and report back. Depending on the Unit within the department, Faculty do guided observation. I prefer giving narrative feedback spanning over several classes. If there are problems, we ask the faculty member to do a self-assessment.
- The FEC process needs to reflect and value the educational process. We need a way of quantifying educational achievement and teaching. Knowing what individual departments do to assess teaching would be useful.
- Our experience with UTS hasn't been altogether positive. It is not a big plus for faculty to do UTS workshops with respect to FEC.

Questions asked related to FEC:

- For those instructors who are in trouble – how do I assess them?
- How do I do peer observation?
- How do I do a peer consultation within the department?
- How do I build on people's strengths?
- How do I tell a tenured professor that gets brutal reviews that they need an intervention?
- Where do I access services that provide effective for support faculty in trouble?
- As chair, how do I delegate this? (I have no time to do this myself).

Comments related specifically to USRIs

- According to the results on the USRI we are doing just fine in our teaching. But the USRIs present a double edged sword – people begin teaching to USRI.
- There is a presumption that teaching and education is the same thing. There is the assessment of learning verses the assessment of teaching and we are not separating and assessing them the way we should be.
- There is a push to student centeredness rather than learning centeredness. Administration has to address this fundamental issue.
- The institution needs to clarify what it wants students to achieve by the time they leave. There needs to be quality control on assessment. There needs to be a clear sign on educational achievement rather than on things that are mechanistic, e.g. teaching to USRIs

- New instructors bring their USRI scores to me and say what does this score mean? I don't know, so a handout to support USRI would be helpful.
- Faculty do try to respond to feedback given in USRI to try to improve. The IDQ is relevant. Comments and scores speak to when there is a need for improvement within the classroom. We have an additional form that is used to evaluate teaching in smaller classes (4-9 students). I try to support faculty and get involved. The process works well.
- Co-teaching not captured in the way we evaluate. The way we evaluate is not useful.
- USRIs are one of the largest issues I have at FEC time. We don't know what is working and how to measure it. The problems I see are the full-tenured faculty who are teaching well, but are getting nailed because they require students to think critically and to do work. With newer faculty there is a real shift in how much content is covered. Content is dumbed down to get higher scores on USRIs.
- On the USRI, most student evaluations are over 4, but everyone on the U of A website is over 4. You have to be really bad at teaching to get nailed.
- If we wind up getting grades that are too high, we take a look to see what's happening in the class. We've made a policy decision to increase difficulty level of examinations. The USRI is tough to criticize. We need to be tactful. It's touchy subject.
- Feedback given on USRI gives students the opportunity to give the department a clear signal early on if there is something not working. I do go through the USRIs. Where there are poor evaluations, I read all of the comments. This year, I met with two instructors who had done poorly. With one instructor, I mentored her. With the other individual, I recommended peer consulting. This individual had difficulty with English and had presentation problems (e.g. lacked eye contact, etc.). I also encouraged this faculty member to sit in on other instructors who are doing well.
- I recommend people who are having trouble with their teaching go to UTS. But, to be fair, Faculty are trapped between scoring well on the USRIs verses ensuring the rigor of their courses. There is no correlation between USRI and what people have learned. Faculty have to navigate tension between rigor and having students like them. Do you want to improve USRIs or improve teaching? How do faculty do things in an effective way to improve teaching? Our time needs to be focussed on the teaching not on satisfying students.
- Students can't tell a good instructor from a good course. What is the teaching rating versus the course rating and how do they compare? They need a teaching measure.
- We have a young department and have had a new turn over. New faculty are eager to read USRI results. Academics (like anyone else) are easily trained to increase USRI scores and there is an element of this. On the whole however, we are doing well.
- I have concerns about faculty teaching to the USRI. Faculty USRI scores are tracked and color coded by my Dean. You do not want to be red.

Appendix B

History of Student Rating of Instruction at the University of Alberta

Prior to 1978: students' ratings collected by the Students' Union, 20,000 forms completed in 1977

1978 - 1985: questionnaires individually designed and used in many departments; Faculty of Education had a common form

1985: IDQ (Instructor Designed Questionnaire) system acquired from the University of Michigan to support individualized questionnaires from a common catalog of items

1985 - 1994: IDQ system promoted by CITL (Committee for Improvement of Teaching and Learning); 107,524 forms completed in 1993/94

1987: Students' Union proposes to publish a "Course Evaluation Guide"; seeks administrative approval to access classes to administer questionnaires (didn't happen but indicative of some of the campus politics)

1994: USRI (Universal Student Ratings of Instruction) proposed by TLC (GFC's Teaching and Learning Committee); approved by GFC; funded by VP (Academic); common set of questions for comparison across campus; all instructors to be rated in all classes; printed reports to be made available to Students' Union.

1. This course was 1=A requirement 2=An elective 3=Other
2. My university year is 1=First 2=Second 3=Third 4=Fourth 5=Post-Degree
3. Overall, this was an excellent course
4. Overall, the instructor was effective
5. The course was well organized
6. The objectives of the course were achieved
7. The instructor presented the material in an interesting and helpful manner
8. The instructor seemed to enjoy teaching
9. The instructor treated students with respect
10. The instructor was helpful in answering questions
11. The instructor was reasonably accessible outside of class
12. The workload for this course was appropriate
13. The type of assigned work was appropriate to the goals of the course
14. The instructor assessed my work fairly

1995: Committee struck by Dr. Owram with representation from CITL, AASUA, Chairs, Sociology to revise questionnaire

1. My university year is 1=First 2=Second 3=Third 4=Fourth 5=Post-Degree 6=Unclassified/Other
2. This course was 1=A requirement 2=An elective 3=Other
3. The instructor spoke audibly and clearly
4. The instructor was accessible outside of class SD to SA plus "Never Tried/Not Appicable"
5. The instructor treated students with respect
6. Overall, the instructor was 1=Poor 2=Fair 3=Acceptable 4=Very Good 5=Excellent
7. Overall, this course was 1=Poor 2=Fair 3=Acceptable 4=Very Good 5=Excellent

1998: Results for "Universal" questions published on the Web (FOIPP-approved at the Provincial level at request of Anne Marie Decore)

1998: Committee struck by TLC to revise questionnaire

- literature review
- pilot proposed forms (2 versions: random, matched samples)
- psychometric analysis of results

1999: 3rd Edition of USRI implemented; policy set to not collect ratings from classes with enrollments less than 10

1. The goals and objectives of the course were clear.
2. In-class time was used effectively.
3. I am motivated to learn more about these subject areas.
4. I increased my knowledge of the subject areas in this course.
5. Overall, the quality of the course content was excellent.
6. The instructor spoke clearly.
7. The instructor was well prepared.
8. The instructor treated the students with respect.
9. The instructor provided constructive feedback throughout this course.
10. Overall, this instructor was excellent.

1999: TLC struck a subcommittee to examine evaluation methods for alternative delivery courses. After lengthy discussion, the Committee agreed that 'alternative delivery courses' should be defined quite broadly, encompassing (but not limited to) web-based courses, courses with many instructors, distance-delivery courses, 'context-based learning' courses (such as those used in the Faculties of Nursing and Medicine and Dentistry), and courses with other non-traditional teaching and learning modes.

2001: A report from the above committee was submitted to TLC by Carolin Kreber. The committee had attempted to address the above by deriving parallel questions for various methods of delivery using a framework based on Boyer's Scholarship of Teaching. Recommendations from this report were not pursued; due, at least in part, to the turn-over of members on TLC and the departure of Dr. Decore

2003: Online USRI surveys administered to online classes

2008: There are now 7 variations of the 1999 USRI questions to accommodate

- problem based learning
- on-line classes
- lab tutors
- small group facilitators
- seminar facilitators
- teaching assistants
- tutors

2008: app. 250,000 questionnaires in 75 departments have been completed in each of the past 6 years
55 departments use more than the 10 USRI questions; some as high as 32 questions
data collected from 85 classes using the Remark Web Survey system