Context

Student evaluations of teaching and courses have been widely used internationally for various purposes such as tenure and promotion and teaching improvement. The Teaching and Course Evaluation Project (the Project) at Simon Fraser University (SFU) is an initiative sponsored by the VPA on the recommendation of the Senate Committee on University Teaching and Learning (SCUTL) and the Task Force on Teaching and Learning (TFTL) to develop new instruments and guidelines for student evaluation of teaching and courses. This report serves as one of the supporting documents for the Project to provide the SFU community with an overview of research findings and current practices on student evaluation of teaching and courses with a focus on instrument development and related methodological issues. The Project strives to develop a valid and reliable instrument that best suits SFU’s evaluation needs by closely examining the relevant research literature and through community engagement during the process of instrument development and implementation.

Key issues regarding instrumentation

**Multidimensionality of teaching effectiveness**

Student evaluations of teaching and courses intend to measure a hypothetical construct of “teaching effectiveness.” There is a vast amount of research on student evaluations of teaching, however no consensus has been reached on what constitutes “good teaching.” It is generally understood that teaching effectiveness is a multidimensional and complex construct that includes multiple and possibly networked sub-constructs. What constitutes effective teaching may also vary from institution to institution, discipline to discipline, and faculty to faculty. Therefore, special attention will be given to gathering information on local preferences, priorities and standards pertinent to effective teaching at SFU. Such standards and priorities will be developed based on the literature and SFU’s unique identity and evaluation needs at institutional, departmental, or unit levels. This rationale has prompted the planning of high level community engagement as an essential part of the Project in seeking advice, consultation, and feedback from all stakeholders before and during the process of developing a sound instrument that can best meet SFU’s evaluation needs.

---

1 The text for this executive summary is excerpted from a report prepared by Hui Niu on research findings on instrumentation and validation of student evaluation of teaching and courses. For the full report or more information on the project, please visit the project website: [www.sfu.ca/teachingandcourseeval.html](http://www.sfu.ca/teachingandcourseeval.html)

2 For a broader review on general issues regarding student evaluations of teaching and courses, see the report submitted to the Senate Committee on University Teaching and Learning by Johanne Provençal in January 2012.
Purposes of evaluating teaching effectiveness

The purpose of evaluating teaching effectiveness is twofold: collecting information for summative and formative reasons. Formative evaluation in the context of teaching quality is intended for instructors to improve their teaching. Summative evaluation is commonly used for personnel decisions on tenure, promotion, or teaching appointments. A larger context for the use of teaching assessment is as a response to the growing interest of the public to hold postsecondary institutions accountable for student learning.\(^3\)

Validity and reliability

Numerous research studies, reviews and meta-analyses have shown that student evaluation of teaching can be valid, reliable and stable. However, faculty opinions of student evaluation of teaching vary vastly from viewing it as reliable, valid and useful to unreliable, invalid and useless.

Validity of student evaluation of teaching is primarily dependent upon the interpretation and use of the results, including inferences about teaching effectiveness. Validity has been the most common concern about student evaluations of teaching in higher education. Such concerns are connected to assumptions that students are not qualified to evaluate teaching and the existence of potential bias variables. These variables have been researched extensively in the literature, and overall, researchers argue that student ratings are consistent with other data sources such as instructor self-evaluation, trained observer evaluation and former student ratings. Although more research is needed to establish validity of student evaluations, such as using more criteria for teaching effectiveness rather than mainly student achievement, researchers are optimistic that student ratings of teaching effectiveness are at least one measure that reflects teaching quality.

Reliability is an important aspect of validity. Research has shown that a student evaluation instrument can be reliable, especially when the instrument is well designed and tested according to psychometric principles. Reliability of a student evaluation tool is usually determined by inter-rater agreement (agreement across different students in the same class) and internal consistency of a scale (correlations among items measuring the same element of effective teaching).\(^4\) Stability refers to reliability over time. Longitudinal studies comparing immediate ratings with ratings at least a year after have shown high correlations and thus demonstrate good stability.

Generalizability refers to reliability across situations. It assumes that evaluation of an instructor’s general teaching effectiveness across courses is more reliable than one snapshot of his/her

\(^3\) See the report submitted to the Senate Committee on University Teaching and Learning by Johanne Provençal in January 2012 for more details about this topic.

\(^4\) For example the scale of an instructor’s organization skills consists of four items in SEEQ, a well-researched instrument developed by Herbert Marsh: 1) Instructor's explanations are clear; 2) Course materials are well prepared and carefully explained; 3) Proposed objectives agree with those actually taught so you know where the course is going; 4) Instructor gives lectures that facilitate taking notes.
performance in a particular course at a particular time. Research has indicated that student ratings of teaching effectiveness across different courses or the same course across multiple terms are consistent. It has been recommended that for summative purposes, average scores of several courses taught by the same instructor should be used to enhance generalizability.

**Potential Biases: Myths or Reality?**

Perceptions of potential biases in student evaluation of teaching can be a source of anxiety for instructors. Important concerns regarding background variables that may threaten validity of student ratings have been discussed extensively in the literature over decades. Some researchers believe these variables introduce biases into the evaluation outcomes, others argue that they are merely sources of validity evidence because they cannot be disconnected from features and dimensions of teaching effectiveness. While caution is required in interpreting effects of background variables on student ratings, there seems to be a trend in the literature indicating that some of these biases are not warranted by research evidence. However, it is important for the Project team to take into account evidence-based potential biases when developing and validating our instruments.

---

5 For example, the effects of expected grades/grading leniency is one of the most common concerns among faculty. However a student’s expected grade from a course is likely to correlate with his/her efforts toward and achievement on the subject.