Student Course Evaluations: Research, Models and Trends

Prepared by: Pamela Gravestock and
Emily Gregor-Greenleaf, University of Toronto

for the Higher Education Quality Council of Ontario

Higher Education Quality Council of Ontario
An agency of the Government of Ontario

Disclaimer:

The opinions expressed in this research document are those of the authors and do not necessarily represent the views or official polices of the Higher Education Quality Council of Ontario or other agencies or organizations that may have provided support, financial or otherwise, for this project.

Cite this publication in the following format:

Gravestock, P. and Gregor-Greenleaf, E. (2008) *Student Course Evaluations: Research, Models and Trends.* Toronto: Higher Education Quality Council of Ontario.

Published by:

The Higher Education Quality Council of Ontario

1 Yonge Street, Suite 2402
Toronto, ON Canada
M5E 1E5
Phone: (416) 212-3893
Fax: (416) 212-3899
Web: www.heqco.ca
E-mail: info@heqco.ca

Table of Contents

F. Access to results
    i. Who has access? To what?
    ii. Management of written comments
    iii. Publication of results
G. Interpretation and use of results
    i. Summative and formative purposes
    ii. Information supplied with evaluation results
    iii. Use of written comments
    iv. Tenure, promotion and merit
    v. Teaching awards
    vi. Other evidence of teaching effectiveness
H. Relationship of course evaluations to accountability measures

4. Reliability, Validity and Interpretation of Course Evaluation Data
    A. Introduction to reliability and validity
    B. Students as evaluators
    C. External Validity: Creating the instrument
        i. Defining effective teaching
        ii. Developing evaluation instruments
    D. External Validity: Reporting and interpreting evaluation results
        i. Reporting of evaluation results
        ii. Challenges to interpretation and use for summative purposes
    E. Internal Validity: The influence of variables on evaluation results
        i. Overview of studied variables
        ii. Validity testing

5. Implementing Effective Evaluation Measures: Recommendations from the Research
    A. Introduction
    B. Ensuring Validity
    C. Ensuring Utility
        i. For students
        ii. For instructors
        iii. For administrators

Note: An appendix is available in English-only under a separate cover. To request a copy of the appendix, please email info@heqco.ca.

Contributors

Pamela Gravestock is the associate director of the Office of Teaching Advancement at the University of Toronto where she works with faculty on a range of teaching and learning issues and initiatives. She is also a doctoral candidate in the Higher Education Group (Theory & Policy Studies) at OISE/UT. Her dissertation examines current practices and policies relating to the evaluation of teaching for the purpose of tenure and promotion. With Emily Gregor-Greenleaf she has developed two publications for faculty: *Developing Learning Outcomes: A Guide for Faculty* and *Gathering Formative Feedback with Mid-Course Evaluations*.

Emily Gregor-Greenleaf is a doctoral candidate in the Higher Education Group (Theory & Policy Studies) at OISE/UT. Her dissertation provides a history and analysis of the undergraduate curriculum at the University of Toronto. She is also currently a research assistant with both the Office of Teaching Advancement and the Teaching Assistants' Training Program at the University of Toronto. In the latter role, she is developing and administering a large-scale survey to evaluate TA training programs at the University of Toronto.

Acknowledgements

We would like to thank Andrew Boggs at the Higher Education Quality Council of Ontario for his support and feedback throughout this project. We would also like to acknowledge the support of the Centre for the Study of Students in Postsecondary Education at the University of Toronto. Finally, and most importantly, we would like to express our gratitude to the Higher Education Quality Council of Ontario for providing us the opportunity to pursue this large-scale study.

Section 1

Introduction

This document represents the first review and summary of existing research on student course evaluations from a Canadian perspective. The scholarship in this area is vast and of varying quality and scope. Our review is an attempt to capture and synthesize the key issues and findings regarding the validity and utility of student course evaluations. We have organized our research into the following seven sections:

Section 1*: *Introduction – provides an overview of the scope, methodology and limitations of this study.

Section 2*: *Context – identifies the current state of scholarship and interest in course evaluations and the evaluation of teaching more generally. It also reviews student, faculty and administrator perceptions of course evaluation systems.

Section 3*: *Current Policy and Practice in North America – offers an overview of evaluation instruments, policies and processes from 22 post-secondary institutions in Canada and the United States as well as policies related to course evaluations from system-level and government agencies.

Section 4*: *Reliability, Validity and Interpretation of Course Evaluation Data – summarizes and reviews the findings from previous studies conducted over the past 40 years with a particular emphasis on the last two decades.

Section 5*: *Implementing Effective Evaluation Measures: Recommendations from the Research – synthesizes research findings and identifies recommendations for improved administration and interpretation of course evaluations.

Section 6*: *Emerging Trends, Existing Gaps and Suggestions for Further Research –highlights issues currently being considered in the scholarship along with those that have been identified as areas requiring more in-depth analysis.

Section 7: Concluding Remarks – provides a brief summary of our most important findings and recommendations.

Overall, our findings indicate that while course evaluation instruments generally provide reliable and valid data, significant barriers to the effective use of such evaluation systems continue to exist due to:

- Persistent myths and misconceptions about variables affecting evaluation results;
- Unclear concepts and definitions of effective teaching;

- Insufficient education about the goals, uses and validity of course evaluations for students, faculty and administrators;
- Poor presentation and contextualization of evaluation data; and
- Inconsistent and inequitable policies and practices regarding the implementation and administration of course evaluations.

Our findings suggest that no matter the reliability and validity of the evaluation instruments themselves, the policies, processes and practices at an institution determine the degree to which evaluations are an effective measure of teaching quality.

1.A Methodology

Literature search

The bulk of information contained in this survey reports the results of a substantial review of published research on course evaluations and the evaluation of teaching. Our search was conducted across a number of academic databases and traced bibliographic references identified in the articles we discovered. Though we reviewed literature dating back to the 1970s (the period that witnessed the expansion of research on course evaluations), we focused primarily on research published in the last 20 years, as many of the earlier studies were repeated or had their findings challenged. As well, more recent studies frequently included summaries of earlier scholarship.

The organization of this review is the result of an iterative process that reflected the development of our understanding of the material. We have attempted to incorporate all the major themes we identified in the research.

Survey of postsecondary institutions

A second part of our study was a survey of publicly available information about course evaluation policies and practice at a range of North American institutions and postsecondary systems. The institutions selected for this survey, and the motivation for their selection, is described in the introduction to *Section 3: Current Policy and Practice in North America*. We drew information from institutional websites and the sites of governance and organizational bodies, using search terms including “course evaluations,” “teaching evaluations,” “evaluation of teaching” and “student feedback,” among others. While these institutions were selected to address a range of institutional types and mandates, as well as a range of jurisdictions, we cannot claim to be able to make general conclusions about course evaluation policy and practice from the institutions surveyed here; instead, the discussion of our findings highlights common or particularly unique policies and procedures discovered through this survey.

1.B Limitations

No literature review on this subject can be comprehensive given the vast amount of research that currently exists (and continues to grow). Even as we conducted our review,

new publications emerged: raising new issues and rehashing old ones, presenting alternative approaches and conclusions and reporting new findings. We made every effort to locate as many sources as possible, covering the full breadth of relevant issues. However, in some cases, we reviewed but did not refer to sources that are included in later literature reviews or studies if we felt that their findings were accurately represented in the later publications. The scope of this study did not permit us to fully review or re-analyze findings from earlier research, nor did we conduct our own primary research into the issues discussed here. As evidenced by this review, many of the key issues have already been thoroughly, and adequately, addressed in the scholarship. However, there remain a few areas that demand further research. These are detailed throughout and more specifically in *Section 6.B: Existing Gaps and Suggestions for Further Research*.

A further limitation of this review is the lack of Canadian data. The bulk of the research on course evaluations has been conducted by American scholars at U.S. institutions. While there are obvious similarities between the higher education sectors in Canada and the United States, there are also significant differences in terms of structure, organization and accountability measures, not to mention cultural and demographic variations. In addition, institutional policies and practices (particularly in relation to tenure and promotion) vary within and between these two countries. As such, we are aware that there may be limits to the degree that research findings can be generalized across sectors. While we attempted to incorporate some additional Canadian data through the institutional scan, our survey, as noted above, does not provide a comprehensive review of institutional policies and practices in either jurisdiction; rather, we provide a sample to demonstrate a range of current activities.

Our review of several hundred publications relating to course evaluations revealed a surprising amount of disagreement between scholars. On one hand, certain questions pertaining to reliability, validity and utility have resulted in a general consensus supported by strong research. Even so, subsequent studies frequently reintroduce into the debate issues long considered resolved, at times needlessly muddying the waters. And so, while these issues may appear to be resolved for a time, their reentry into the discussion often raises new questions or reframes old questions in new contexts. On the other hand, there are some issues that have been continually debated, seemingly with little hope of resolution. Many of these debates are detailed in *Section 4: Reliability, Validity and the Interpretation of Course Evaluation Data*.

Perhaps not surprisingly, we also noted that scholars on different sides of a particular issue often pick and choose particular studies to highlight and reference. While this is to be expected, we were dismayed and concerned by the apparent lack of objectivity related to this sort of “selectivity”. Frequently authors do not mention the specifics of methodology or the size and scope of a study, nor do they consider the generalizability of findings. This is problematic. For example, many authors continue to cite studies that have long been refuted, debunked or found to be methodologically unsound by the majority of scholars. This includes the so-called Dr. Fox study by Naftulin, Ware and Donnelly (1973) which is now widely viewed as invalid (Abrami, 2001; Ali & Sell, 1998). Some scholars have noted this when referencing it. However, others still cite it as proof that an instructor’s enthusiasm or expressiveness can result in higher ratings (Wright, 2006). Similarly, Wright (2006) cites the Williams and Ceci study (1997) also viewed by

most as methodologically unsound because it draws its conclusions from results for one small class, from one term; as such, the findings cannot be generalized.

One challenge to the generalizability of research findings is the diversity of course evaluation instruments, policies and processes, as well as the diversity of institutional and instructional contexts. These all vary significantly by, and sometimes within, institutions. As discussed in *Section 4.C.ii: Developing Evaluation Instruments*, the wording, order and combination of items, or even the scales used on questionnaires, can substantially affect the results received; therefore, studies conducted on one survey instrument may yield different conclusions than the same study performed on another. Similarly, teaching is such a complex and multi-faceted enterprise, with such a range of participants and external influences, that separating one variable from others is a significant challenge in any study.

Section 2

Context

2.A Evaluating Teaching in Higher Education

Moore and Kuol (2005) have argued that [g]iven that it is an almost universal phenomenon that research activity reaps more individual rewards than those associated with teaching, efforts to measure the teaching related dimensions of [faculty] performance, and to pay attention to those measures in the context of an individual's professional development helps to create more parity of esteem between the teaching and research components of the academic role" (p. 143). The quantifiability and comparability of most course evaluations makes the imprecise art of evaluating teaching seem more objective and manageable.

In Canada and the U.S. common means of evaluating teaching typically include course evaluations, letters from students and colleagues, in-class/peer evaluations, the receipt of teaching awards, course materials and texts and evidence of innovative strategies and practices. Each of these measures brings with it its own restrictions and limitations. This is why most institutions rely on more than one form of evidence to develop a complete understanding of a candidate's teaching contributions. However, course evaluations or student ratings are one of the most common tools used to assess classroom teaching (Wright, 2006; Seldin, 1999; Centra, 1979) and some believe the most heavily weighted (or over-weighted) for personnel decisions (Franklin, 2001). Student evaluations are also one of the most controversial and highly-debated measures. Nonetheless, they are still widely used. Many have argued that there is no other option that provides the same sort of quantifiable and comparable data (Abrami, 2001).

By a wide margin, course evaluations are used for summative, as opposed to formative, purposes (see *Section 2.D.iii Collecting and interpreting formative feedback*) that is, as a means to make personnel decisions (e.g. hiring, tenure, promotion, and annual review) based in part on a student's rating of an instructor's teaching effectiveness. The collected data, in particular the qualitative responses, are also used by instructors and teaching support offices to provide formative feedback intended to facilitate improved teaching and course development. Wright (2008) cautions against the use of instruments not specifically designed to provide formative feedback for this purpose, and that separate instruments should be designed to provide summative and formative feedback respectively.

Much has been written about the problems with course evaluations. Educational scholars have examined issues of bias, have identified concerns regarding their statistical reliability and have questioned their ability to accurately gauge the teaching effectiveness of faculty. In addition, some have argued that the feedback provided by course evaluations does not effectively promote change in faculty behaviour. However, a significant majority of researchers consider student evaluations to be a useful measure of

the instructional behaviours that contribute to teaching effectiveness (including Beran, Violato & Kline, 2007; Abrami, 2001; Schmelkin, Spencer & Gellman, 1997; Marsh, 1987).

2.B The Vocabulary of Student Course Evaluations

There are almost as many terms used to describe student course evaluations as there are articles about them; among the most common are "student evaluations," "course evaluations," "student ratings of instruction," and "student evaluations of teaching (SETs)." Each of these phrases has slightly different connotations, depending on whether they emphasize students, courses, ratings, or evaluation. Wright (2008) has suggested that the most appropriate term for end-of-course summative evaluations used primarily for personnel decisions (and not for teaching development) is "student ratings of instruction" because this most accurately reflects how the instrument is used. For further discussion of this terminology, please see *Section 6.B.i: Defining Teaching Vocabulary and Expectations*. Throughout this paper, we have used several of these terms interchangeably but have selected "student course evaluations" as our primary term because this is the phrase used most frequently at Canadian institutions.

2.C Faculty, Administrator and Student Perceptions of Course Evaluations

> "There are probably more misconceptions about student ratings than facts known about them, yet we do know quite a bit" (p. 3).
>
> Ory, J.C. (2001). Faculty thoughts and concerns about student ratings. In K.G. Lewis (Ed.), Techniques and strategies for interpreting student evaluations [Special issue]. *New Directions for Teaching and Learning*, *87*, 3-15.

Countless myths and misperceptions regarding course evaluations exist and inevitably influence faculty, university administrator and student perceptions. In spite of solid research to counter these assumptions, such beliefs persist and continue to spread. One only need raise the issue at a departmental meeting, faculty luncheon, or campus event to elicit a range of "anecdotal evidence" from various members of the university community. As Nasser and Fresko (2002) note, few extensive studies have been conducted on the attitudes and perceptions about course evaluation systems by those who use them and who are affected by them (particularly faculty, students and administrators). Studies that address these issues are typically small, capturing responses from a limited number of individuals; however, there is some consistency in prevailing attitudes and thus some generalizations can be made.

2.C.i Faculty Perceptions

Student course evaluations have been established as a source of anxiety for faculty (Hodges & Stanton, 2007; Ryan, Anderson & Birchler, 1980) and for some incite outright hostility (Franklin & Theall, 1989). Such attitudes are derived from persistent beliefs that evaluations are biased (Eiszler, 2002; Feldman, 1976), that students are not competent evaluators (Nasser & Fresko, 2002; Ryan, Anderson & Birchler, 1980) and that ratings are impacted by student grade expectations (Baldwin & Blattner, 2003). These issues and others have led both faculty and administrators to question the overall validity of student evaluations and their use and the potential misuse of data (Beran, Violato & Kline, 2007; Ory, 2001), particularly in relation to personnel decisions (Nasser & Fresko, 2002; Sproule, 2000; Ryan, Anderson & Birchler, 1980). However, as we will explore more fully in *Section 4: Reliability, Validity and Interpretation of Course Evaluation Data*, the research has frequently disproved many of these concerns. As Theall and Franklin (2000) have observed, "[f]aculty discomfort with ratings and shortfalls in good practice are signs of persistent disjuncture between the worlds of research and practice" (p. 95). These negative perceptions of evaluations can lead faculty to discount their importance and can hinder teaching and course development efforts. And as Aleamoni (1999) and Ory (2001) have argued, both faculty and administrators have continued to generate and perpetuate the mythology and misperceptions about course evaluations.

Anecdotal evidence combined with various empirical studies clearly demonstrates that many faculty still object to the use and are suspicious of student evaluation systems. Some have argued that a higher percentage of faculty possess negative rather than positive or neutral attitudes toward evaluation tools (Nasser & Fresko, 2002; Abrami, 2001; Theall & Franklin, 2001; Wachtel, 1998; Centra, 1993). Studies have also found that the use of course evaluation systems can decrease faculty morale (Ryan, Anderson & Birchler, 1980). However, the findings are mixed and faculty opinions vary widely (Wachtel, 1998; Schmelkin, Spencer & Gellman, 1997; Newport, 1996). A 2005 study of 357 faculty (Beran et al.), revealed that a majority of the instructors surveyed had generally positive views of course evaluations: 63% indicated they did not find them to be intrusive, 70% did not find them to be a waste of time and 82% did not think they were an inappropriate means of assessment. Moreover, Beran and colleagues (2007, 2005) found that more than half of the faculty surveyed believed that ratings data were being used appropriately by academic administrators.

2.C.ii Administrator Perceptions

Most studies have demonstrated that administrators, in general, have a positive attitude toward evaluation data and find it a useful source of information for personnel decisions (Campbell & Bozeman, 2008; Beran et al., 2005). This was the most common administrative use of evaluation data. For example, Beran et al. (2005) found that 82% of the administrators surveyed in their study use student evaluation ratings for summative purposes, particularly for decisions relating to promotion, tenure and merit.

While administrators may agree that these tools are effective, their attitudes are also subject to the pervasive misconceptions surrounding validity concerns. Theall and Franklin (2001), Abrami (2001) and others have noted that such misconceptions prevail due to a general lack of familiarity with the research on ratings validity or an unwillingness

to accept findings. In addition, the literature has also shown that administrators are influenced by their own approaches and attitudes about teaching and about individual instructors. McKeachie (1997) notes that an administrator's own stereotypes about teaching influences their judgments about teaching effectiveness: if a candidate does not conform to their existing stereotype about what makes a good instructor, they are viewed negatively.

In *Section 4.D.ii: Challenges to Interpretation and Use for Summative Purposes* we discuss the tendency of some administrators to overestimate the precision of evaluation results, particularly when comparing results between courses and instructors.

2.C.iii Student Perceptions

Research on student perceptions of course evaluations and their use of evaluation data is limited. Most of the studies have been small, drawing on samples from one institution (Campbell & Bozeman, 2008; Beran et al., 2005). From these, it would appear that students perceive the process of collecting student feedback as valid and useful. Moreover, they also believe that students can be and are effective evaluators of teaching. However, some studies demonstrate that students are not always aware of how institutions use collected data (Campbell & Bozeman, 2008; Beran et al., 2005; Wachtel, 1998), nor do they always understand the impact that ratings have on personnel decisions. Some studies find that students are skeptical that their input is used and reviewed (Wachtel, 1998). Furthermore, many students make little use of ratings data: in a study of 1,194 students and 35 alumni at one Canadian institution, Beran et al. (2005) discovered that 56% of students did not use ratings data at all. Of the 43% who indicated they had consulted them, less than one-third (31%) used them to select courses based on content and structure (e.g. assignments, workload, topics) and almost two-thirds used them to select courses based on the instructor (64%).

2.D Common Characteristics of Course Evaluations

In the process of researching course evaluations, several scholars have identified the common characteristics of course evaluation tools. Algozzine et al. (2004), for example, describe a typical evaluation based on their research on the development and use of course evaluation instruments:

> *The historical and traditional method of evaluating instruction in university classes is to have students provide feedback on "effectiveness" using a "cafeteria-style" rating scale.... Traditional "cafeteria-style" course evaluation systems have similar characteristics: (a) an instrument is developed, comprised of a series of open- and closed-ended questions about course content and teaching effectiveness; (b) at least one item addresses 'overall' effectiveness; (c) written comments about the course content and the effectiveness of the instructor are solicited; (d) anonymity of responses is assured and assumed; (e) responses are obtained at the end of the term in the absence of the instructor; (f) item and scale responses are summarized across instructors, departments, and colleges and evidence of "teaching*

> *effectiveness" used in making various professional development decisions; and (g) student (for example, GPA, academic year), course (required, graduate), and instructor (novice, experienced) differences largely are ignored in analysis and reporting of scores reflective of effectiveness* (p. 135).

The various items included on course evaluation forms assess different and separable aspects of an instructor's teaching behaviours and the course. Generally, students assess each of these individually, ranking some more positively than others (Beran, Violato & Kline, 2007).

As we will see in *Section 3: Current Policy and Practice in North America*, there are variations in format and practice across institutions. However, certain elements are almost universal. Course evaluation forms are most commonly distributed at the conclusion of a particular unit of instruction. They are almost always anonymous (or, less frequently, confidential) and most frequently incorporate both qualitative and quantitative responses. Quantitative questions ask students to respond on Likert and other rating scales, most commonly with five or seven points. At some institutions a standardized form is available for use within all courses, whereas at others, forms are developed at the divisional and, less frequently, at the departmental level. In some instances, standard questions are mandated for all faculty (in a division or institution-wide); in others, individual faculty members and/or departments can supplement these questions with ones particular to their programs or teaching activities. In general, faculty are removed from the process of collecting course evaluation data and typically are unable to access the ratings until the final grades for all students have been submitted. While traditionally course evaluations have been administered in-class and on paper (using a scannable form), recently a number of institutions have moved toward the implementation of online tools (see *Section 6.A.i: Online Evaluation Tools* for a more thorough discussion of this emerging trend).

Evaluations generally request specific feedback on measures of teaching effectiveness and on particular aspects of a course, as well as global rating questions and, frequently, a limited number of open-ended questions that seek qualitative written responses. Most evaluation instruments are designed to be employed in summative evaluation of teaching, but formative assessment is possible through alternative models of student evaluation instruments and through the diagnostic interpretation of results from the kinds of evaluations described above. (See *Section 4.D.i: Reporting of Evaluation Results* for a more detailed discussion of the various means by which institutions can effectively report evaluation data for summative and formative purposes.)

2.D.i Common measures of teaching effectiveness

Items on course evaluations seek information about course design and delivery and instructor behaviour. Cashin (1995) notes six elements that commonly appear on evaluations: 1) questions about course content; 2) questions about the instructor's communication skills; 3) questions about student-teacher interaction; 4) questions about course difficulty and workload; 5) questions about assessment practices in the course; and, 6) student self-assessment questions.

The different components of course evaluations also derive from research about student learning and about successful teaching behaviours. The perceived need to ask questions about multiple categories of teaching behaviour emerges from the belief that teaching effectiveness is multidimensional; that is, that instructors may excel in some elements of teaching and not in others (Algozzine et al., 2004; Marsh & Roche, 1997; Marsh 1987). In 1987, Marsh developed the Student Evaluations of Education Quality (SEEQ) evaluation instrument, which includes nine categories of questions about teaching behaviours that he argued should all be present in order to ensure that an evaluation is representative of teaching effectiveness: 1) learning/value; 2) instructor enthusiasm; 3) organization; 4) individual rapport; 5) group interaction; 6) breadth of coverage; 7) examinations/grading; 8) assignments/readings; and 9) workload/difficulty. Similar measures of teaching effectiveness have been identified by Braskamp and Ory (1994) and Centra (1993) and in the Individual Development and Educational Assessment (IDEA) evaluation system developed by R. Cashin at Kansas State University. These include course organization and planning, clarity/communication skills, teacher student interaction/rapport, course difficulty/workload, grading and examinations and student self-rated learning. Other studies, such as those by Feldman (1989), have identified as many as 28 categories of teaching behaviours. The challenges of identifying and defining those teaching activities and strategies that most contribute to student learning are discussed in *Section 4.C.i: Defining Effective Teaching*.

In the Canadian context, Harry Murray (1987) at the University of Western Ontario developed the Teaching Behaviours Inventory, which can be used to gather information from students on 60 instructor behaviours and teaching activities. These behaviours measure teaching activities believed to improve student learning and are grouped into nine categories: 1) clarity; 2) expression; 3) interaction; 4) organization; 5) pacing; 6) disclosure; 7) speech; 8) rapport; 9) teaching aids.

Examples of several of these instruments can be found in Appendix B: Sample Institutional Evaluation Instruments. Adapted versions of these instruments are in use at many institutions as discussed in *Section 4.C.ii: Developing Evaluation Instruments*.

2.D.ii Collecting and interpreting qualitative feedback

Scholars engaged in the evaluation of teaching have contended for several decades that assessment of teaching effectiveness is best conducted according to multiple, qualitative measures of teaching effectiveness in addition to student ratings (Lattuca & Domagal-Goldman, 2007; Ory, 2001; Arreola, 1983). Many of these qualitative means of assessing teaching (including portfolios or dossiers, self- and peer-assessment and written teaching narratives) are not conducted by students; indeed, as discussed in *Section 4.B: Students as Evaluators*, while students are effective at measuring in-class teaching behaviours and activities, they are not well-qualified to evaluate course content or teaching goals and other sources of information therefore need to be consulted.

However, arguments for the inclusion of qualitative sources also indicate the value of collecting such feedback from students on topics addressed in course evaluations. Indeed, Harper and Kuh (2007) note that qualitative means of assessment can often bring to light issues that cannot emerge through conventional quantitative means. For this reason, qualitative feedback from students is primarily conducted, evaluated and

used for formative, rather than summative, purposes (Franklin, 2001; Lewis, 2001). Frequently, this takes the form of mid-course evaluations similar in character to end-of-course evaluations (Lewis, 2001), or more informal, in-class assessment (Diamond, 2004). Mid-course evaluations are discussed in *Section 5.C.i: Ensuring Utility for Students* and *Section 6.A.iii: Increasing Use of Evaluations for Formative Purposes*.

Algozinne et al. (2004), however, note that most standard end-of-term course evaluation forms do include an opportunity for students to include written comments. In these cases, although qualitative data is collected, it is often not effectively interpreted, analyzed, or incorporated into summative evaluation procedures. The management and interpretation of written feedback is discussed in *Section 3.F.ii: Management of Written Comments* and *3.G.iii: Use of Written Comments*. One challenge is that student comments can be misleading or inaccurate; Hodges and Stanton (2007) argue that student confusion about their own learning processes can lead to conflicting or confusing comments on evaluations. Another challenge is the perceived increase in time and effort needed to assess written comments; several studies (Beran, Violato & Kline, 2007; Beran et al., 2005; Wagenaar, 1995) of the use of evaluations by instructors and administrators indicate that these groups rarely review written comments, preferring instead to use only what they perceive to be the more time-efficient global ratings. These authors suggest that training be offered to students, instructors and administrators about the value of written comments and on techniques for, respectively, writing and analyzing these comments effectively.

Some faculty place more trust in the qualitative responses to open-ended questions provided by students than in the quantitative ratings; others claim the opposite. However, studies have shown that there is a correlation between the qualitative and quantitative ratings (Cashin, 1995).

Abrami (2001) argues that qualitative measures should not be introduced into the summative evaluation of teaching because their reliability and validity cannot be easily assessed; Harper and Kuh (2007) argue that this concern, while not inaccurate, is not germane to the way qualitative information can and should be used in summative assessment.

2.D.iii Collecting and interpreting formative feedback

As noted by Beran, Violato and Kline (2007), Beran et al. (2005) and Wagenaar (1995), teaching evaluations are primarily used, by a wide majority, for summative purposes; that is, by administrators to support personnel decisions. Beran, Violato and Kline's (2007) study, in fact, demonstrates that though faculty believe teaching evaluations to be useful in assessing teaching, they rarely employ the results of their own evaluations in course or professional development decisions.

Most scholars attribute this relative absence of formative use of teaching evaluations to a lack of resources for interpreting evaluations and identifying teaching strategies that might address problems that emerge (Beran et al., 2005; Wagennar, 1995). This can limit the ability of evaluations to improve teaching; Ory (2001) and Marsh (2007) note that evaluations may lead to improved teaching only if their results are discussed with a colleague.

Nonetheless, Lewis (2001) and Ory (2001) note that to be most effective in improving teaching, assessment should be both continuous and formative and evaluated in the context of an instructor's personal goals for teaching improvement. They argue that if resources exist to assist in the interpretation and implementation of evaluation results, teaching evaluations can be extremely useful as a professional development tool.

Formative feedback may be conducted using traditional end-of-course evaluations or through alternative forms of evaluation. Aultman (2006) and Lewis (2001) advocate the use of early and mid-semester evaluations to gather formative feedback that can be acted upon immediately. Hodges and Stanton (2007) describe how written student comments can reveal information about aspects of the learning process that students do not understand and can therefore serve as an important course development tool.

Another kind of formative feedback that can emerge from more standard summative evaluations is the diagnostic evaluation. If the evaluations used are multi-dimensional, a report can be provided to instructors identifying their areas of strength and those that need improvement. Such a report can facilitate self-directed and institutional-supported teaching development (Crosson et al., 2006; Marsh & Roche, 1997). Such reports are further described in *Section 4.D.i: Reporting of Evaluation Results*.

2.E Common Uses for Course Evaluation Data

Moore and Kuol (2005) have found that student evaluation systems help to counter anecdotal information about teaching behaviours and effectiveness. They also assert that such tools provide another means to assess teaching and thus help to shrink the existing gap between the evaluation methods for teaching and research.

There are several common uses for course evaluation data: teaching improvement; personnel decisions; course selection (by students); and increasingly, in the compilation of teaching award nominations files.

Teaching improvement

Since the widespread use of evaluation began, researchers have argued that course evaluation data can effectively be used for the purpose of improving teaching and thereby student learning (Goldschmid, 1978). However, Marsh (2007) and Goldschmid (1978) have found that course evaluation data alone rarely bring about changes to teaching behaviours since many faculty are not trained in data analysis and are therefore less likely to have the necessary skills to interpret their ratings. Moreover, many faculty are not given the opportunity (voluntary or mandatory) to discuss their results with departmental chairs or deans and only some take advantage of the services and resources offered by campus teaching and learning support offices. As a result, the majority of faculty simply conduct a cursory review of the collected data and rarely attempt to make specific changes based on student feedback.

Research has demonstrated that when faculty are provided training or assistance and consultations with colleagues or faculty/educational developers, they make changes to

their teaching behaviours (Penny & Coe, 2004). To encourage change and positively influence teaching behaviours, Abrami (2001) has recommended more open communication regarding collected data and the interpretation of the results. Beran, Violato and Kline (2007) suggest that evaluations be "supplemented by complementary sources of information regarding instructional effectiveness" and argue that "all user groups, including administrators, faculty, and students should be aware" (p. 37) of the need for this supplemental information when using student ratings.

Personnel decisions

Researchers in the 1980s and 1990s regularly questioned the use of course evaluations for summative decisions. In part, these concerns stem from beliefs that ratings data were not being used effectively or equitably. However, the debate about the effective use of evaluation data for summative (and also formative) purposes also relates to the questions that guide these personnel decisions. In the last decade, attitudes have shifted and most scholars, among them Abrami (2001) and Algozzine et al. (2004), generally accept – and/or attest to – the validity of course evaluation ratings for these personnel decisions, including hiring, tenure and promotion.

Thirty years ago, research indicated that while faculty favoured the use of student evaluations for use in promotion and tenure decisions (Rich, 1976), university administrators were not regularly relying on them for such purposes (McKeachie & Lin, 1975). More recently, some studies have suggested that administrators are more likely than individual instructors to make use of course evaluation data (Beran et al., 2005) particularly for personnel decisions (Nasser & Fresko, 2002; Haskell, 1997; Schmelkin, Spencer & Gellman, 1997). Some studies have noted that it is unclear whether administrators are using the collected information appropriately (Abrami, 2001), or if it is being misinterpreted or misused as the only source of data about teaching (Franklin & Theall, 1989).

A recent study (Beran, Violato & Kline, 2007) at one research-intensive Canadian university found that administrators, in general, view student evaluations positively but do have some reservations regarding their effective use. Beran et al. (2005) found that administrators find course evaluation data useful for evaluating individual teaching (for making personnel decisions and recommendations for teaching awards), monitoring progress for the remediation of teaching problems, evaluation of teaching at the unit level and for curriculum planning. In this study, administrators indicated that the most useful questions on course evaluation tools were the global items that provided information on the overall quality of the course or the instructor. This corresponds to recommendations from other studies (d'Apollonia & Abrami, 1997) that global items be used and to findings from Cashin and Downey (1992) that indicate these are the most useful indicators of teaching effectiveness.

In spite of their usefulness for summative evaluation and personnel decisions, there is general consensus that course evaluation data should not be used in isolation but rather should be one of multiple indicators used to assess teaching (Beran, Violato & Kline, 2007; Ory, 2001).

Course selection by students

At some institutions course evaluation data are made available to students through publications such as “anti-calendars.” Anti-calendars typically provide summaries of evaluation data, along with selected comments from students. These documents are designed to be used by students for the purpose of course selection; some evidence suggests that their use for such purposes is limited (Beran et al., 2005). See *Section 3.F.iii: Publication of Results* and *5.C.i.: Ensuring Utility for Students* for an overview of current practice of and recommendations for this use of evaluation data.

Teaching awards

Course evaluation data are often a required element for teaching award nomination dossiers both internally (at departmental, divisional, or institutional levels) and externally (e.g. the Ontario Confederation of University Faculty Association, 3M National Teaching Fellowship). Here, the expectation is that candidates will demonstrate excellence in teaching within their discipline, for which course evaluations serve as one indicator. Moreover, since such data are regularly collected, candidates can normally demonstrate sustained excellence or provide comparable data to indicate their relative performance within their department, division, or institution.

Section 3

Current Policy and Practice in North America

3.A Introduction

As noted in *Section 1.A: Methodology*, this overview of current policy and practice at a selected number of institutions across North America is meant to provide some insight into a variety of evaluation instruments, processes and uses. The goal of this survey is not to identify the prevalence of particular practices, but rather to highlight the range and variation, as well as the commonalities, in the development, administration and interpretation of course evaluations at institutions that vary by mission, programmatic focus, size and jurisdiction.

After reviewing all available information, we organized practice and policy according to the categories outlined below. Not all categories were addressed in the information available from each institution and therefore only relevant information from each source is presented. The fact that, frequently, only incomplete information is available is itself important: while some schools (for example, Harvard) keep some information on websites accessible only to faculty, in many cases the information available to us is the same information that would be readily available to instructors seeking information about course evaluation policy and practice at their own institutions.

We noted, in particular, a significant absence of policies regarding, or information available to instructors and administrators providing guidance about, the interpretation of course evaluation results. A small number of institutions – for example, University of Michigan – provide a guidebook to facilitate and contextualize course evaluation results, but most institutional policies and information address only the process of conducting evaluations and disseminating the results. Information about interpreting evaluations is, however, essential to the appropriate use of course evaluations in the evaluation of teaching, particularly when this evaluation is for the purpose of hiring, tenure, or promotion decisions. Consequently, *Section 5.C.iii: Ensuring Utility for Administrators* discusses relevant recommendations for the provision of interpretive guidelines to instructors and administrators.

Table 1: Surveyed Institutions

Ontario	Colleges:	George Brown College, Sheridan College Institute of Technology and Advanced Learning, Humber College Institute of Technology and Advanced Learning, Seneca College of Applied Arts & Technology
	Universities:	University of Guelph, McMaster University, Queen's University, Ryerson University, Trent University, University of Ontario Institute of Technology (UOIT), University of Toronto (UofT), York University
Other Canada:	Brandon University (Manitoba), Dalhousie University (Nova Scotia), McGill University (Quebec), St. Francis Xavier University (SFX) (Nova Scotia), University of Alberta, University of British Columbia (UBC)	
United States:	Amherst College (Massachusetts), Harvard University (Massachusetts), University of Michigan, University of Minnesota	
Governance and organizational bodies:	Ontario Postsecondary Education Quality Assessment Board (PEQAB), California Postsecondary Education Commission, Ontario Public Service Employees Union (OPSEU, represents Ontario college faculty), Ontario Ministry of Training, Colleges and Universities	
Other organizations:	American Association of University Professors (AAUP), Canadian Association of University Teachers (CAUT)	

Please see Appendix C.5 for the list of institutional and organizational policies and documents referenced in this section.

3.B Course Evaluation Policies

3.B.i Prevalence and location of policies

Most institutions maintain course evaluation policies at the institutional level, which are frequently supplemented by divisional policies or procedures. Policies are commonly located in one of four governance or institutional bodies. The first is the faculty collective agreement or related document (e.g. Brandon, Queen's, Ryerson, SFX, Trent). When course evaluation policies are located in the collective agreement, they are usually discussed in the broader context of the evaluation of teaching for hiring, promotion and tenure. At some institutions, including Ryerson, a copy of the university-wide evaluation form is appended to the collective agreement. The second location is Human Resources. This is the case at Humber, whose faculty are part of the OPSEU collective agreement which does not specifically address course evaluations. Third, course evaluation polices are also frequently outlined in Senate (Guelph, McGill, McMaster, UBC, York) or Academic Council policies (Alberta, UOIT). Finally, course evaluation policies are found under the jurisdiction of the institutional office or centre for teaching development and support; such is the case at Dalhousie and Harvard. Michigan is unique: teaching evaluations are administered through its Office of Evaluations and Examinations, an office dedicated to administering and analyzing surveys and tests. At some institutions (e.g the UofT), we could not identify a formal university-wide policy, but in these cases

informal information about course evaluations could frequently be found in similar locations. Furthermore, where institution-wide policies were not present, divisional policies could be located (e.g. UofT Faculty of Arts & Science).

3.B.ii Focus and scope of policies

Policies primarily offer instructions about the administration and implementation of evaluations (e.g. the frequency with which evaluations are performed, the means by which courses to be evaluated are selected and whether and how student anonymity will be protected) and the storage and dissemination of results. Many policies also clearly specify which individuals (e.g. instructor, chair, dean) or bodies (departmental evaluation committee, tenure and promotion committee) have access to the data. A number of policy documents articulate the institution's goals or purpose in relation to the collection of course evaluation data (e.g. Alberta, McGill, UBC, UOIT, York). Some policies (e.g. those at Brandon, Trent and SFX) offer guidelines for the development or modification of evaluation instruments, while others specify very clearly the type of instrument to be used (Alberta, McGill, Ryerson, Queen's, UOIT) or even the number of questions to be included on the form (as at McGill where the maximum number is 25 with 4 mandated institution-wide items and up to 21 additional questions added by academic units). The UBC Policy on Student Evaluation of Teaching includes a section on the assignment of responsibilities, which details specific roles for students, administrators, faculties, departments and instructors.

Policies embedded within collective agreements focus primarily on how course evaluation data may be used in the evaluation of teaching for tenure and promotion. Where formal policies specifically addressing evaluations do not exist, the use of course evaluation data for this purpose is often outlined in other institutional documents, such as policies and procedures related to appointments and promotions (e.g. UofT). In general, policies or information located through offices dedicated to advancing teaching, testing, or student learning provided more thorough information to faculty and administrators about evaluation data and interpretation. These guides are discussed at more length in *Section 3.G.ii: Information Supplied with Evaluation Results* and *Section 5.C.iii: Ensuring Utility for Administrators*.

3.C Design and Approval of Evaluation Instruments

3.C.i Development and approval of evaluation instruments

The process of course evaluation instrument design varied widely in terms of the responsibility for developing questionnaires and the formality of the process for their approval. Across the institutions we surveyed, we found evidence of course evaluation instrument development processes at every level of administration (from the level of individual faculty as at Amherst to the Senate/Academic Council as at UBC and UOIT).

A number of institutions (e.g. Alberta, Dalhousie, Harvard, Humber, McGill, Michigan, Queen's, SFX, UOIT) have mandated the use of a common course evaluation instrument across the institution, whereas others delegate this authority to specific divisions/departments (e.g. Guelph, McMaster, Trent). These forms may be developed

through a teaching or evaluation office (e.g. Michigan, Dalhousie) or through an evaluations committee (e.g. Harvard) or may be determined through governance processes (e.g. Alberta). In some cases (e.g. Alberta, Dalhousie, Michigan, Queen's), the common instrument includes opportunities for faculty to include items selected or developed by the division, department, or individual. A number of institutions offer a bank of items from which faculty can select additional items to be added to the evaluation form (e.g. Queen's). In general, faculty are permitted and encouraged to conduct their own informal or supplementary evaluations in addition to those developed at the institutional level (e.g. Alberta, Queen's, UBC).

The collective agreement at SFX includes a requirement that any changes to the course evaluation instrument be approved by the faculty Senate, with a formal process to petition any changes. A similar stipulation can be found in the Queen's and Ryerson collective agreements. The Faculty Policy at Guelph delegates the design of the evaluation instrument to the department; however, it requires approval by at least two-thirds of the faculty within the unit before implementation.

Some institutions (e.g. Brandon, York) devolve evaluation design and approval to the level of the Faculty, Unit, or Centre. Evaluations must generally adhere to institution-wide policies for the administration, collection and dissemination of evaluation results and are generally approved by the Dean or Director of the Faculty, Unit, or Centre.

Amherst allows individual faculty members to develop their own evaluation instruments, often with approval from a department Chair or divisional Dean. The instrument may be voluntarily standardized at the department or even divisional level.

3.C.ii Questionnaire format and content

We located sample evaluation instruments from a range of institutions. Those from Alberta, Dalhousie, Harvard, Humber, Michigan, Queen's, Seneca and UOIT are used by all instructors (with the frequent exception of teaching assistants) across the institution. The following description of evaluation instruments draws on these examples. Evaluation instruments designed at the divisional, departmental, or individual level can be expected to be significantly more varied.

We found that the structure and content of course evaluation forms strongly parallels the typical evaluation instrument described by Algozzine et al. (2004) in *Section 2.D: Common Characteristics of Course Evaluations*. Most of the instruments primarily requested quantitative ratings and many provided space for additional qualitative comments from students. We identified several different scales ranging between four and seven points, including Likert scales, quality rating scales and frequency rating scales. All forms included at least one question that asked students for a general rating of the course or the instructor. All forms asked questions about course content. This included questions about assignments and, frequently, the relevance of material covered to other courses or to prospective future vocations. All forms also asked questions about teaching behaviours of the instructor. These almost always include questions about instructor enthusiasm for the material, about availability to students and about classroom atmosphere and engagement. Seneca notes that its form was developed through an adaptation of Harry Murray's (1987) Teaching Behaviours Inventory, a popular Canadian

teaching evaluation instrument (see Appendix A). Most forms included questions about the physical environment of the classroom. Several instruments included questions about the use of classroom technology.

In Fall 2007, the UBC introduced a new ratings system using six university-module items (UMI) for all course evaluations. Additional items can be added to evaluation forms by divisions or departments but the following six items are mandatory for use by all instructors:

1. The clarity of the instructor's expectations of learning.
2. The instructor's ability to communicate the course content effectively.
3. The instructor's ability to inspire interest in the subject.
4. The fairness of the instructor's assessment of learning (exams, essays, tests, etc.)
5. The instructor's concern for students' learning.
6. The overall quality of the instructor's teaching.

(These questions use a 5-point scale: (1) Very Poor; (2) Poor; (3) Adequate; (4) Good; and (5) Excellent.)

Several of the instruments contained more unusual elements. The Harvard form included both scaled questions and open-ended opportunities to provide written feedback for every topic addressed. Queen's, Dalhousie and Michigan each employ an instrument that includes a limited number of common questions with a larger number of questions that can be developed or selected by departments or by individual faculty members. The evaluation form at Queen's includes four mandated questions and allows for up to seven items to be selected by departments and a maximum of 10 (from a bank of 200) by the instructor. Dalhousie includes 10 common questions and two sections of five questions for which a question number and scale is provided, but for which the department and the individual faculty member can supply the questions. Michigan's instrument includes four common questions that must appear on each evaluation. One of these questions – "I had a strong desire to take this course" – is used primarily to contextualize the results received on the evaluation, as their office has found that higher responses to this question correspond to higher overall course ratings. The faculty member may choose whether or not to include a group of eight more questions designed by the Michigan Student Assembly, the results of which are published in an annual course guide for students. Faculty then select 18 additional questions (or 26 if they elect not to include the course guide questions) from a bank of over 200 questions on topics including student development; instructor effectiveness; writing, reading, laboratory and other assignments; course materials, including audiovisual materials; instructional computing; grading and examinations; and student responsibility. Instructors may also elect to include up to five open-ended questions about course content, material, assignments and instruction.

Please see Appendix C.2 for a detailed overview of the Michigan Teaching Questionnaires program and instrument, Appendix C.1 for McGill's pool of evaluation questions and Appendix B for additional examples of course evaluation instruments.

3.C.iii Review of evaluation instruments

Recent revisions to the evaluation instrument at the University of Minnesota provide an interesting (and exemplary) case study of the process of reviewing teaching evaluation

instruments. The FAQ (see Appendix C.4) provided for faculty to address questions about the revision process details the steps through which the instrument was reviewed and changed.

The FAQ notes that the form was revised because it "was not based on research about teaching and learning and had a number of items that were not helpful to instructors, administrators, or students" (p. 2), echoing the research reviewed in *Section 4.C.i: Defining Effective Teaching* which argues that validity and utility depends strongly on the ability for institutions to identify questions that reflect the goals and practice of teaching in their institution.

A committee developed and proposed new questions based on extensive research on teaching in higher education as well as in reference to existing instruments and piloted the new instrument in 50 courses. The new instrument was then put towards a vote and passed by the Faculty Senate.
UBC also recently revised its evaluation instrument and process. The new instrument was developed by a Student Evaluation of Teaching Committee and was approved by the University Senate. After the first round of evaluations using the new instrument, the results from each of the six institution-wide questions was reviewed for reliability and validity; overall, the questions were deemed valuable, though it was suggested that certain aspects of the questions' wording could be improved. The reviewers also recommended ongoing assessment of the instrument and improvements to online data storage and collection.

Other institutions, including Queen's and Ryerson, specify in their collective agreements that joint union-administration committees are to be established to review forms and approve any subsequent changes.

3.D Implementation Processes

3.D.i Method of delivery

Institutions conduct, analyze and disseminate the results of course evaluations either online, on paper, or through a combination of the two methods. While institutions that conduct or have explored the possibility of conducting evaluations online (that is, evaluation forms are delivered to students through email or course management systems and are completed on a computer) note that both response rates and overall evaluation ratings are lower (though formal research on this topic is mixed; see *Section 6.A.i: Online Evaluation Tools*) for online evaluations when compared to in-class evaluations (see, for example, the report of the University of Michigan Task Force on Online Evaluations & Placement Examinations), conducting evaluations online remains an attractive prospect: online evaluations save a significant amount of personnel time and, consequently, money. Changes to response rates or average ratings are not necessarily a problem if all evaluations are conducted online and if relevant contextualizing information is provided to faculty and administrators. For this reason, conducting some evaluations online and some in class is not advisable.

Though the presence of online methods of conducting evaluations is growing (we noted a number of schools whose evaluations were conducted entirely online, including UOIT and McGill; others offer a modular approach permitting several means of conducting evaluations, as at Guelph, Queen's, Trent and UBC; and several other schools noted the desire to explore the possibility or were piloting online delivery methods, as at Ryerson), the primary means of delivering course evaluations remains through in-class, paper evaluation forms. Typically, these forms are printed on scannable bubble sheets to facilitate analysis.

Most institutions conduct paper evaluations but conduct the analysis of evaluations, store evaluation data and, less frequently, disseminate the results of evaluations online or via computer.

3.D.ii Implementation guidelines or policies

Guidelines for the administration of course evaluation policies include the selection of courses or instructors to be evaluated and the process of printing, distributing and collecting evaluation forms. Institutions vary on the frequency and comprehensiveness with which they conduct evaluations. Most institutions evaluate each course every year. Less frequently, institutions select a portion of the courses taught by each faculty member. For example, Seneca evaluates three courses taught by each full-time faculty member; its evaluations policy states that these courses should be selected in coordination with faculty and should be representative of the range of types and levels of courses taught. Some institutions do not conduct evaluations, or in some cases do not disseminate results, for very small classes (e.g. McGill) due to reliability concerns (Cashin, 1995); at others, official policies prescribe (e.g. Alberta) alternate methods of evaluation for classes with low enrolment (under 10 students).

Evaluations are normally coordinated at the administrative level that maintains responsibility for course evaluations at that institution (this may be within a department or division, at the provostial level or within institutional registrarial or assessment and evaluation offices). In general, this body prints the forms and distributes or arranges for the distribution of forms to individual faculty members (usually in coordination with departmental administration).

Faculty are often given the responsibility of coordinating course evaluations in individual classes. Many institutions (e.g. McGill, UBC) provide information for faculty (usually in the form of a statement to be read in class) to communicate to students the process and importance of completing evaluations. Harvard includes a statement directed to students on the evaluation itself which reminds students that their responses will be kept anonymous and that student evaluations are read and taken seriously. In addition, the statement asks students to provide thorough and constructive comments and to avoid comments on their instructor's personal appearance or characteristics. Similar practices are in place at Alberta and Guelph.

Frequently, faculty then elicit a student volunteer (though in some cases, administrative support is offered) to distribute and collect the evaluation forms and return them to an administrative office, usually at the departmental level. These processes are designed to

protect student anonymity, to ensure that evaluations are processed uniformly and to ensure that evaluations are not seen by faculty until after grades have been submitted.

All institutions that we surveyed restrict faculty involvement in the evaluation process. Polices at Queen's and Alberta note that instructors are not permitted to distribute or collect the form and are to be absent from the room while students are completing the evaluation forms – practices common to all institutions reviewed.

While most institutions administer evaluations during the last several weeks of a course (either in-class or online), Guelph requires that students receive a copy of the form at the outset of the term. Guelph's policy documents also state that departments are to distribute an overview of related policies and procedures, as it pertains to the collection of evaluation data, to all students. As at other institutions, students complete the forms near the end of a course.

The Canadian Association of University Teachers (CAUT) has prepared a "Model Clause on the Evaluation of Teaching Performance" (see Appendix D.1) that they recommend be provided to faculty with institutional policies. Alberta, for example, has included a link to this statement in the materials that accompany their course evaluation policy and procedures.

3.E Analysis of Results

The analysis of course evaluation results is the process of collating, translating and synthesizing individual student responses. This analysis may be done by the administrative body responsible for course evaluations at that institution (for example, the Office of Evaluations and Examinations at Michigan or Test Scoring and Questionnaire Services at Alberta) or by external consultants (e.g. Seneca). Normally, this includes the calculation of response means for each question on the evaluation, as well as response means that can be used for comparative purposes. For example, at Seneca, means from each evaluation are reported in the context of means at the program, school, Faculty and institutional level. Written comments are most often typed into a computer file to ensure student anonymity; their management is further described below.

At Alberta, institutional policy details how evaluation results are to be analyzed and presented for distribution. The General Faculty Policy states that numerical summaries, detailing the response rates for each category and the median score to one decimal point for each evaluation item are to be distributed to the instructor, students, Chair and Director or Dean. In addition, numerical values which take into account and summarize skewed data and identify outliers from the general population, if they exist, are also required for all reports.

3.F Access to Results

3.F.i Who has access? To what?

At the institutions surveyed, faculty whose courses are being evaluated have full access to collected data. In general, this access is restricted until final marks have been submitted for all students enrolled in the specific course. Implementation guidelines and procedures detail the level of access for other members of the university community. In general, departmental chairs or unit heads, deans and tenure and promotion committees share the same level of access as individual faculty members; this is the case at Guelph, McMaster, Queen's and York. In some cases, this is restricted to the compulsory questions found on all evaluation forms and does not include data from supplementary optional questions added to the form by the instructor (as at Alberta, Ryerson and UBC). Faculty at McGill must grant permission for their ratings results to be made available to the broader university community, including students. This is similar to Trent, where the collective agreement states that evaluations remain confidential to the faculty member. At neither institution does this restrict the use of ratings data for tenure and promotion purposes.

There is some variation in practice in providing access to evaluation results for students. In some cases, institutional guidelines merely recommend, and do not require, that data be made available to students (e.g. UOIT and York). At others, summary reports for students are produced using data from specific evaluation modules (e.g. UBC). Several institutional policies make no mention of students in relation to data access and some restrict them to the viewing of summary results from the mandated institution-wide questions (e.g. Alberta).

At UofT, student associations in some divisions such as the Faculty of Arts & Science produce an Anti-Calendar with summarized data from undergraduate courses. Here, faculty may choose to deny publication of their results. Harvard and Michigan maintain similar systems; at Harvard all results are shared, while at Michigan, instructors may opt to include a set of eight questions in their evaluation specifically designed for inclusion in a course evaluation guide for students.

3.F.ii Management of written comments

While many course evaluation forms include both qualitative questions, requesting written responses from students to specific questions and space for general comments, institutional guidelines are not always explicit with regard to how this data is to be managed. In some cases, there was simply mention of the practice of collecting written comments; others outlined who had access to this information; and some dictated very clearly the processes for collecting, reporting and managing qualitative responses (e.g. Alberta and Guelph).

Alberta's General Faculty Policy states that written comments are to be typed to ensure student anonymity. Alternatively, students may wish to submit typed comments separately from the in-class/online evaluation process. At Guelph, written comments require a legible student signature (as outlined in both institutional policy and the accompanying Provostial Protocol document). If student comments are unsigned they

are only shared with the instructor. All course evaluation forms include a statement detailing this policy. At Queen's, the University Survey of Student Assessment of Teaching (USAT) form is anonymous, with the exception of a section that provides an opportunity for signed written comments from students. These policies may exist to allow for the inclusion of written comments in tenure and promotion materials, as collective agreements sometimes prohibit the use of anonymous, non-aggregate data (e.g. Brandon).

McMaster's Policy on the Encouragement of Teaching Excellence charges departments with consolidating ratings data into a report with tabulated numerical data and an evaluative summary of written comments.

Some institutions share written student comments with the instructor only and do not keep copies in a central file (e.g. Michigan, Queen's, Ryerson, SFX). McGill's policy indicates that written comments are confidential to the instructor and the chair of the department. Others store written comments with quantitative student ratings (e.g. Amherst, Harvard).

Many institutions without formal policies for the management of written comments nonetheless made recommendations in relation to potential uses for such material (see *Section 3.G.iii: Use of Written Comments* below).

3.F.iii Publication of results

As noted above, some institutions make it a practice to regularly publish (or report) course evaluation data (Alberta, McGill, Queen's, Ryerson, SFX, UBC), whereas others merely recommend that the results be disseminated (e.g. UOIT, York). The method of distribution varies, from printed digests or summary reports (Alberta, Harvard, McMaster, Queen's, UBC, UofT Faculty of Arts & Science) to online documents detailing institution-wide, divisional and departmental averages for specific questions (e.g. Ryerson). Publication of results does not imply widescale student use, as detailed above in *Section 2.C.iii: Student Perceptions of Course Evaluations*.

The CAUT "Policy on the Use of Anonymous Student Questionnaires in the Evaluation of Teaching" (2006) states that "[w]here/when student organizations conduct anonymous student surveys and publish the results in order to assist students in the selection of their courses, academic staff participation should be optional" (see Appendix D.2 for the full statement).

3.G Interpretation and Use of Results

3.G.i Summative and formative purposes

At the majority of institutions reviewed for this study, course evaluations are conducted at the end of a course, thereby collecting and providing summative data. This intention is clearly outlined in institutional documentation (e.g. Alberta, Guelph, McMaster, McGill, Queen's, Ryerson, Trent, UOIT). In some cases, policy documents further recommend that formative data also be collected through mid-course evaluations or other means.

This is the case at UBC and Alberta, which both advocate for modular or multi-faceted, ongoing teaching evaluation through a variety of means. Such a recommendation may also appear outside of formal policy in implementation guidelines, as at Ryerson.

In a recent review of its evaluation instrument (see *Section 3.C.iii: Review of Evaluation Instruments*), the University of Minnesota (2008) created a parallel mid-course version of its institutional evaluation form designed to provide formative feedback. They note that this mid-course instrument "includes both the core items from the end-of-semester form and also a number of written items designed to help instructors improve their teaching" (p. 3).

3.G.ii Information supplied with evaluation results

The majority of institutions surveyed make the raw data and summary reports available to faculty and department heads following the submission of final grades in a particular term/semester. However, there is considerable variety in the supplementary information that is provided to faculty and administrators who will be interpreting the data. For example, Michigan includes divisional means with evaluation summaries while others include instructor or departmental averages. At Queen's, reports are provided to faculty members and administrators with aggregated data for quantitative questions, along with the mean, standard deviation, frequency and number of eligible responses calculated. These reports also include graphical representations of data.

General guides outlining the implementation process have been produced at a number of institutions (e.g. Ryerson, McGill, SFX and UOIT). The Queen's Collective Agreement requires that the university provide all those charged with assessing and evaluating teaching performance with a clear explanation of statistical terms used in the evaluation process. The Queen's Office of the University Registrar, the unit responsible for the administration of student evaluations, has prepared a number of documents for users, including an FAQ that addresses how the system works and an information sheet for evaluation report users which details what the reports provide and how the data was analyzed.
Some institutions (McGill, Michigan) have developed guides that detail how to interpret evaluation results for personnel decisions and to improve teaching effectiveness. At York, the Senate Committee on Teaching and Learning has developed a guide to teaching assessment and evaluation which provides faculty and administrators with an overview of the various evaluation mechanisms, their benefits and limitations and advice on how to use them effectively and reflectively. The Teaching and Learning Services office at McGill has published a similar guide titled *Effective and Appropriate Use of Student Ratings of Instruction: Research based suggestions*.

At Alberta, all evaluation data is distributed to chairs, deans, tenure committees and students with a cautionary reminder about various forms of bias. This institution offers one of the most extensive documents to support those administering, interpreting and receiving evaluation data. A comprehensive 54-page manual includes excerpts from the General Faculty Policy pertaining to the evaluation of teaching and the use of the Universal Student Ratings of Instruction system (USRI) which outlines the purpose for evaluation, the instrument format and content and implementation procedures. The document also includes practical information for faculty regarding the administration of

student evaluations, copies of evaluation instruments and a lengthy catalogue of additional questions that may be added by the instructor. Sample copies of instructor and administrator reports are also provided for review along with brief guidelines for reading and interpreting these documents.

Please see Appendix C.3 for examples of the University of Alberta evaluation data reports.

3.G.iii Use of written comments

While many institutional policies refrain from making recommendations regarding how written comments are to be managed, some (Harvard, Ryerson), note that such feedback can be used for teaching award nominations, or included in teaching dossiers and tenure and promotion files. However, some institutions explicitly state that written comments should not be used for personnel decisions (e.g. McGill). Brandon prohibits the use of anonymous information in materials used in tenure and promotion decisions, effectively preventing the use of comments from anonymous student evaluations. SFX permits the use of written comments in tenure and promotion materials only when the faculty member has granted permission for the inclusion of these materials. At other institutions, guidelines for administrators note that while such data can be effective, they caution users about potential bias and limitations of such material (Alberta).

3.G.iv Tenure, promotion and merit

Summary data from course evaluations are regularly used for the purpose of tenure, promotion and annual merit review. This use is articulated in institutional (or divisional) course evaluation policies and in those documents detailing procedures and policies related to tenure, promotion and annual review. All of the institutions surveyed highlight the use of course evaluation data for this purpose.

All institutions, however, note that course evaluations alone should not be the basis for formal evaluation of teaching, and some mandate that evaluations be accompanied by relevant, contextualizing information (often in the form of a teaching dossier; see *Section 3.G.vi: Other Evidence of Teaching Effectiveness*). At SFX, course evaluations may only be considered in formal teaching evaluations if course evaluations over the past three years display a consistent pattern. Course evaluations must be presented in the context of additional relevant information about the course, including its place in the curriculum, course size, information about course material and delivery and the instructor's other teaching duties.

3.G.v Teaching awards

Although rarely mentioned in formal policy, most institutions surveyed note that course evaluation data could potentially be used when compiling teaching award nominations. This is a common requirement for both internal and external teaching awards.

At Harvard, course evaluation results are used to award the Harvard University Certificate of Distinction in Teaching to teaching fellows, teaching assistants, preceptors

and lecturers. At UofT, course evaluation data form one part of the evidence in the nomination dossier for the President's Teaching Award.

3.G.vi Other evidence of teaching effectiveness

In general, we found teaching dossiers (or portfolios) to be the most common form of additional evidence recommended by institutions (e.g. Guelph, McGill, Queen's, Ryerson, Trent, UBC). The most common elements found in a dossier include a teaching philosophy statement, information on pedagogical strategies used inside and outside the classroom, representative course materials, sample student work and evidence of teaching awards, professional development, mentorship and research on teaching and learning (Seldin, 1999).

Peer evaluation is also often suggested (e.g. Guelph) as are other forms of in-class evaluation (e.g. Alberta, Amherst) such as on-site observations by colleagues of faculty developers.

3.H Relationship of Course Evaluations to Accountability Measures

Course evaluation results are sometimes used as an element of larger jurisdictional accountability measures. For example, the Ontario Postsecondary Educational Quality Assessment Board, which accredits degree programs at Ontario colleges, requires the assurance that programs will regularly review teaching through means including student evaluations, but does not review student evaluations directly as part of its assessment program. In its Multi-Year Accountability Agreement with the Ontario Ministry of Training, Colleges and Universities, Sheridan commits to particular levels of student satisfaction with courses and curriculum as measured on its student feedback instrument.

The inclusion of student evaluations in accountability mechanisms is by no means universal, however. California's Postsecondary Education Commission Accountability System, for example, does not request the results of student evaluations.

Section 4

Reliability, Validity and Interpretation of Course Evaluation Data

4.A Introduction to Reliability and Validity

There is general and long-standing agreement in the literature that course evaluation instruments can be, and most often are, reliable tools because they provide consistent and stable measures for specific items (e.g. an instructor's organizational skills or relative workload) (see for example, Abrami, 2001; Theall & Franklin, 2001; Wachtel, 1998; Goldschmid, 1978). This is particularly true when the tool has been carefully constructed and psychometrically tested before use (Centra, 1993; Aleamoni, 1987; Marsh, 1984). Moreover, their reliability is further confirmed by the fact that scores generally represent averages of evaluations collected from a number of students in a given class (Hoyt & Pallett, 1999). Marsh and Roche (1997) and Marsh (1987) have studied the reliability of course evaluation tools by examining the level of agreement on particular items from students in the same course. They have argued that "inter-rater agreement" is an indicator of reliability; however, they note that the reliability factor decreases slightly in smaller classes. Reliability has also been examined through multi-section testing, demonstrating consistency in results in multiple course sections (Ory & Ryan, 2001).

Although most researchers may agree that student evaluations of teaching are reliable tools, there is somewhat less consensus regarding their overall validity: the degree to which the tool accurately measures specific items (e.g. instructor availability) or provides a general rating of the course or instructor. At times during the past 40 years, there has been agreement on some aspects relating to validity (such as the effect of particular course, student and instructor characteristics; see *Section 4.E.i: Overview of Studied Variables* for more on these variables), although conclusions have not remained constant and subsequent studies often discount earlier findings based on methodological grounds (e.g. sample size).

In addition to the variables that may affect evaluation results, we found in our research that many scholars identify additional threats to validity. In particular, validity is strongly determined by the development of appropriate questions, scales and implementation procedures: further, validity is also conditional on the appropriate use and interpretation of evaluation data.

A useful historical overview of the research from 1975-1995 by Greenwald (1997) notes that the majority of publications produced during this period indicate that course evaluations are valid. In a 1997 special issue of *American Psychologist* focusing on course evaluations, the contributors to the volume (among them McKeachie, Greenwald and Marsh & Roche) agreed that student course evaluations are the "single most valid source on teaching effectiveness" (McKeachie, 1997, p. 1218). Those who have found

course evaluations to be valid (Abrami, 2001; Ali & Sell, 1998; Abrami, d'Apollonia and Cohen, 1990; Marsh, 1987) have shown that ratings data can be correlated to other evidence of teaching effectiveness such as evaluations from colleagues or trained faculty development personnel (Ory & Ryan, 2001; Ali & Sell, 1998; Wachtel, 1998).

While it is rare to find current research that outright dismisses course evaluations due to validity concerns, disagreement continues to persist in relation to the validity of particular aspects of evaluations in relation to the range of variables that may impact ratings results (see *Section 4.E.i: Overview of Studied Variables* below).

4.B Students as Evaluators

The fundamental question regarding the validity of student course evaluations is whether students can, in fact, accurately evaluate teaching. As noted in *Section 2.C.i: Faculty Perceptions of Course Evaluations*, one of the primary concerns identified by faculty about course evaluations is a fear that students are not reliable assessors of teaching behaviours or courses. The research both assuages and validates this concern. Agreement regarding the competency of students as evaluators can be traced back to the literature from the 1970s (Goldschmid, 1978). Several studies demonstrate that students are reliable and effective at evaluating teaching behaviours (for example, presentation, clarity, organization and active learning techniques), the amount they have learned, the ease or difficulty of their learning experience in the course, the workload in the course and the validity and value of the assessment used in the course (Nasser & Fresko, 2002; Theall & Franklin, 2001; Ory & Ryan, 2001, Wachtel, 1998; Wagenaar, 1995). Scriven (1995) has argued that students are "in a unique position to rate their own increased knowledge and comprehension as well as changed motivation toward the subject taught. As students, they are also in a good position to judge such matters as whether tests covered all the material of the course" (p. 2).

Indeed, Theall and Franklin (2001) argue that, on these issues, students may in fact be more qualified than expert or peer assessors to rate their instruction; they write that "peers and administrators are generally more knowledgeable of the content and thus cannot necessarily empathize with the views of students who may be having problems" (p. 48).

Many studies agree that other elements commonly found on evaluations are more difficult for students to assess. These include the level, amount and accuracy of course content and an instructor's knowledge of, or competency in, his or her discipline (Coren, 2001; Theall & Franklin, 2001; Green, Calderon & Reider, 1998; Cashin, 1998; Ali & Sell, 1998; d'Appolonia & Abrami, 1997; Calderon et al., 1996). Such factors cannot be accurately assessed by students due to their limited experience and knowledge of a particular discipline. Ory and Ryan (2001) state that "the one instructional dimension we do not believe students, especially undergraduates, should be asked to evaluate is course content" (p. 38). It has also been suggested that students are unable to evaluate instructor grading practices and methods of delivery, appropriateness of selected readings and whether instructors present any bias in their delivery of course content

(Hoyt & Pallett, 1999; Keig & Waggoner, 1994; Cashin, 1988; Cohen & McKeachie, 1980).

See *Section 4.E.ii: Validity Testing* for information about means of validating student responses.

4.C External Validity: Creating the instrument

Ory (2001) and Theall and Franklin (2001) note that, for evaluations to be valid measures of teaching effectiveness, the questions on the evaluation instrument must reflect both 1) the ways in which the evaluations are used for formative or summative evaluation of teaching and 2) the current pedagogical and instructional goals of the institution. Ory and Ryan (2001) also note the importance of ensuring that evaluation questions match only and all of that which the evaluation is attempting to measure; they identify:

> *[C]onstruct underrepresentation and construct-irrelevant variance as two significant threats to validity. As stated, construct underrepresentation occurs if an assessment is too narrow or fails to include important dimensions of the construct, and construct-irrelevant variance exists if an assessment is too broad and contains excess reliable variance that affects responses in a manner irrelevant to the interpreted construct* (p. 33).

The centrality of valid questions to ensuring valid evaluations cannot be overstated. The task of identifying and developing relevant questions is described below.
Design of the instrument also plays an important role in ensuring validity. Scriven (1995) suggests that the validity of evaluations may also be affected by the length of the form, while Sedlmeier (2006) discusses the effect of particular rating scales on evaluation results. These issues are further discussed in *Section 4.C.ii: Developing Evaluation Instruments*.

4.C.i Defining effective teaching

Definitions

Ory and Ryan (2001) write that "to make valid inferences about student ratings of instruction, the rating items must be relevant to and representative of the processes, strategies, and knowledge domain of teaching quality" (p. 32). For course evaluations to be valid measures of teaching effectiveness, not only must the questions reflect those aspects of teaching identified as effective, but the very definition of effective teaching must be identified and agreed upon. Defining effective teaching behaviour is difficult, however: Ory and Ryan argue that no "universal set of characteristics of effective teachers and courses that should be used as a target…appears to exist" (p. 32).

Several evaluation instruments have attempted to identify specific teaching behaviours or means of structuring and presenting course material that contribute to effective teaching. For example, the developers of the SEEQ (Marsh, 1987), the Teaching Behaviours Inventory (Murray, 1987) and IDEA (Cashin, 1992) each reviewed research on student

learning and surveyed students and faculty to identify characteristics of effective teaching (see Appendix A). These items may be further validated through comparisons to other measures of student learning (for example, academic performance). However, there are few means beyond logical analysis (Marsh, 1987) to ensure that these characteristics of effective teaching are representative, comprehensive, or generalizable. Wachtel (1998) notes that several scholars have warned that other means of assessing the validity of student evaluations "presume[…] a consensus which does not yet exist. That is, how can we evaluate teaching effectiveness adequately if we cannot even agree on what constitutes effective teaching?" (p. 193).

Indeed, several scholars have warned of negative effects from delineating specific characteristics of effective teaching. McKeachie (1997) draws on Scriven (1981) to argue that "no ratings of teaching style (e.g. enthusiasm, organization, warmth) should be used, because teaching effectiveness can be achieved in many ways. Using characteristics that generally have positive correlations with effectiveness penalizes the teacher who is effective despite less than top scores on one or more of the dimensions usually associated with effectiveness" (pp. 1218-1219). McKeachie also notes that "faculty members and administrators have stereotypes about what good teaching involves" and that "teachers who do not conform to the stereotype [according to the elements of teaching identified on evaluations] are likely to be judged to be ineffective despite other evidence of effectiveness" (p. 1219).

Ory and Ryan (2001) categorize the institutional effects of evaluations into "intended" and "unintended" consequences. Among the unintended consequences is the possibility that instructors will use items on the ratings form to shape their teaching or courses. If the items on an evaluation form do not reflect institutional priorities or means of effective teaching, these evaluations may have a negative impact on teaching at the institution.

Neumann's (2001) study of disciplinary differences in teaching also has important implications for the definition of effective teaching for course evaluations. Neumann argues that learning goals and teaching styles vary significantly by discipline:

> *Donald (1983) noted that hard pure fields had tightly structured courses with highly related concepts and principles. Soft pure fields had open course structures and were loosely organised. In considering educational goals, Braxton (1995) found the hard disciplines place greater importance on student career preparation and emphasise cognitive goals such as learning facts, principles and concepts. Soft areas place greater importance on broad general knowledge, on student character development and on effective thinking skills such as critical thinking. Hativa (1997) found that soft pure fields placed greater importance on creativity of thinking and oral and written expression, while hard pure and hard applied fields placed strong emphasis on ability to apply methods and principles* (p. 138).

These disciplinary differences could strongly affect the ways in which evaluations are constructed and their validity assessed, as well as the degree to which they accurately reflect teaching effectiveness according to the standards of a particular field. In general,

however, Neumann notes that “in most cases, the evaluation instruments employed are generic, implying that teaching across disciplines is the same” (p. 143).

In contrast to these concerns, however, Wachtel (1998) argues that “students and faculty generally agree on what are the components of effective teaching and their relative importance” (p. 192) and that questions developed from these components can accurately reflect student learning and teaching effectiveness.

Global vs. multidimensional measures of teaching effectiveness

Most attempts to identify particular characteristics of effective teaching stem from a belief that teaching should be measured according to multiple aspects or categories of teaching activity – for example, approachability, enthusiasm, or organization – rather than by questions that seek a broad, global response to the course or the instructor in general. This argument has been advanced most notably by Marsh (1987) and Marsh and Roche (1997). These multiple measures stem from the proposition that teaching is multidimensional – that is, that instructors might excel in one aspect of teaching but not in another. Ory (2001) notes that while a particular set of effective teaching behaviours cannot easily be identified, results from research on student perceptions of effective teaching “support the notion that students view instructional quality as multidimensional” (p. 10).

As these researchers argue, in addition to more accurately depicting effective teaching, multidimensional evaluations can be valuable as diagnostic tools to provide formative feedback (Marsh & Roche, 1997) as they can help instructors identify areas of strength and weakness. Furthermore, Marsh and Roche argue that multidimensional feedback is important to continued research on teaching evaluations. They note several examples of cases in which a variable that might have been identified as a potential source of bias could actually, through evaluating student responses based on a particular dimension of teaching, be shown to have a positive effect on student learning. In other words, “an effect that has been interpreted as ‘bias’ to SETs [can be] more appropriately interpreted as support for their validity with respect to one component of effective teaching” (p. 1193).

McKeachie (1997) concurs with Marsh’s grouping of teaching behaviours into dimensions, noting that this assists with reading and interpreting data and therefore is more likely to lead to improvement. However, he notes uncertainty over the number of dimensions that should be reported on for summative purposes: “should a score representing a weighted summary of the factors be represented (as Marsh and Roche [1997] argue), or should one simply use results of one or more overall ratings of teaching effectiveness (as contended by d’Apollonia & Abrami, 1997)?” (p. 1218).

Many researchers, however, argue that global rating questions are equally, if not more, valuable for summative assessment of teaching than multidimensional measures (Algozinne et al., 2004; Abrami, 2001; Cashin, 1995, 1992). In their study of the ways in which evaluations are used by faculty and administrators, Beran, Violato and Kline (2007) note that global measures are frequently the only ones assessed in formal evaluation processes (see *Section 2.E: Common Uses for Course Evaluation Data*). Abrami (2001) and Cashin (1995, 1992) note that there is a strong correlation between global questions and other measures of teaching effectiveness. When a limited number of results are

reviewed, global questions can accurately serve as a proxy for more complex measurements and therefore in these cases, a true global question, rather than a limited selection of multidimensional measures, is more representative of teaching effectiveness. McKeachie (1997) agrees that such global ratings are valuable for summative evaluation of teaching even if they are not particularly valuable for providing formative feedback. Marsh and Roche (1997), however, disagree with this perspective, arguing that "global or overall ratings cannot adequately represent the multidimensionality of teaching. They also may be more susceptible to context, mood, and other potential biases than are specific items that are more closely tied to actual teaching behaviors" (p. 1188).

In a recent revision to its evaluation instrument (see *Section 3.C.iii: Review of Evaluation Instruments*), the University of Minnesota decided to eliminate its global question, "How would you rate the instructor's overall teaching ability?" The committee charged with revising the instrument argued that this item was too often the only score evaluated in summative teaching assessment, that students have difficulty responding to the question, that the item is not diagnostic and that global questions such as these do not correlate with ratings on questions that review specific teaching characteristics.

There is little discussion in the current literature regarding the particular phrasing of global questions. Scriven (1995) has noted that many summative evaluations ask the wrong global questions. He cites the following common mistakes: questions that require comparisons between teachers; questions that ask students if they would recommend the course to others; and questions that ask students if a course is the "best" they have ever had. Unfortunately, Scriven does not elaborate on why such phrasing is undesirable.

4.C.ii Developing evaluation instruments

Question selection

As Marsh and Roche (1997) argue, the selection of evaluation questions is an essential factor in ensuring that evaluations are valid measures of teaching effectiveness and that:

> *[T]he validity and usefulness of SET information depend on the content and the coverage of the items. Poorly worded or inappropriate items will not provide useful information, whereas scores averaged across an ill-defined assortment of items offer no basis for knowing what is being measured* (p. 1187).

This is both for reasons related to the ways in which students respond to questions (see *Section 4.C.ii: Developing Evaluation Instruments*), the relationship between evaluation questions and those teaching characteristics deemed important or effective in a particular context (see *Section 4.C.i: Defining Effective Teaching*) and the range of questions students can accurately answer (see *Section 4.B: Students as Evaluators*). Despite these important considerations, however, evaluation items are often selected with less care than might be expected. Marsh and Roche (1997) write that "in practice, most instruments are based on a mixture of logical and pragmatic considerations, occasionally including some psychometric evidence such as reliability or factor analysis" (p. 1187). Ory and Ryan (2001) note that "many of the [course evaluation] forms used today have been developed from other existing forms without much thought to theory or construct

domains" (p. 32). As *Section 3.C: Design and Approval of Evaluation Instruments* demonstrates, evaluation development and approval policies and practices vary significantly from institution to institution. Imprecise question selection and instrument development therefore remains a significant barrier to evaluation validity.

Psychometric testing

The wording, order and scale used in questions can themselves have a significant effect on ratings. Consequently, an important element to ensuring the validity of evaluation forms is psychometric testing. As noted above, Ory and Ryan (2001) argue that many institutions develop evaluations using questions that are simply adapted from existing forms. Although these original forms – for example, the question pool developed by the University of Michigan or the IDEA Student Ratings of Instruction system – have undergone extensive psychometric testing, the adapted evaluations that Ory and Ryan describe have not, and may not retain the validity of the originals. As Marsh and Roche (1997) note, "'homemade' [student evaluation of teaching] surveys constructed by lecturers or committees are rarely evaluated in relation to rigorous psychometric consideration and revised accordingly" (p. 1188). Franklin (2001) identifies common problems with such homemade surveys, including double-barreled questions, "overly complex or ambiguous items," or "poorly scaled response options" (p. 89).

Ory and Ryan also write that little is known about the process by which students respond to evaluation questions and whether students respond to rating scales consistently. They note, for example, that there is no research to identify whether "students respond to items by comparing the instructor's performance to that of other instructors or to some idealized standard" (p. 33). Similarly, little research is available to demonstrate how students interpret individual points on rating scales, and that "we need to determine if there is a proper fit between the meaning of the scale for students and its intended meaning" (p. 34). Finally, they note that the ways in which students respond to evaluation scales may vary by demographic factors including age, academic year and cultural background. Other studies demonstrate similar threats to evaluation validity: Greenwald (1997) notes that depending on how a form is constructed, students may provide the same, or similar, rating for all items; Sedlmeier (2006) demonstrates that the order and scale used in quantitative student ratings affect the outcome of the evaluation. Coren (2001) discusses the "halo effect": the notion that when viewing some aspects of an individual in a positive light, there is a tendency to view everything a person says or does in the same light, thereby offering less confidence that ratings of individual items reflect specific strengths and weaknesses. The halo effect also amplifies negative views.

Instrument review

Determining the optimal frequency with which evaluations are revised is a matter of striking a balance between ensuring that evaluation items reflect current pedagogical and institutional practice and priorities and ensuring the evaluation items are selected and evaluated carefully enough that they meet the construct and psychometric validity criteria described above. Ory and Ryan (2001) caution against the use of outdated evaluation questions. They argue that:

> *[F]or example, many colleges and universities are now encouraging faculty to use computer technology in their teaching. Have the rating forms used on these campuses been modified to include technology*

> *items? The value implications of student ratings, whether intended or unintended, may be that the rating content defines dimensions of teaching that are valued and supported by the institution* (p. 38).

Theall and Franklin (2001) recommend that "when institutional or programmatic changes are made, [institutions should] review the evaluation system and adapt it as needed" but emphasize that institutions should "seek expert advice and assistance when necessary" (p. 53) in order to meet another of their recommendations: to "adhere to rigorous psychometric and measurement principles and practices" (p. 52).

4.D External Validity: Reporting and interpreting evaluation results

d'Apollonia and Abrami (1997) note that "[m]any experts in faculty evaluation consider that the validity of summative evaluations based on student ratings is threatened by inappropriate data collection, analysis, reporting, and interpretation" (p. 1203, citing works by Arreola, 1995; Theall, 1994; and Franklin & Theall, 1990).

Similarly, they argue that the "specific questions used, the use of global versus factor scores, the possible biasing of variables, and so forth, are relatively minor problems. The major validity problem is in the use of the ratings by personnel committees and administrators" (p. 1222). Franklin and Theall (1989) come to similar conclusions. The appropriate presentation of evaluation data and the appropriate and trained interpretation of that data is essential, even central, to the validity of evaluations.

4.D.i Reporting of evaluation results

Selecting scores and score composites to report

As noted in *Section 2.D: Common Characteristics of Course Evaluations* and *Section 3.C.ii: Questionnaire Format and Content,* many ratings instruments include an array of items focusing on aspects of the course (content, subject matter, workload) and the instructor (availability, timeliness in returning graded work, clarity regarding expectations) combined with a set of global questions (asking students to rank the course overall and the overall effectiveness of the instructor). With a well-constructed form, results from each of the questions can provide valuable input to faculty, administrators and students. However, the way in which both global and multidimensional items are reported to faculty and administrators can affect the validity of the interpretations derived from that data.

Abrami (2001) notes that ratings data from individual items on evaluation forms can and should be used for formative purposes to improve teaching effectiveness. However, providing faculty with reports from each individual item could prove to be overwhelming and difficult to evaluate. In response, Algozzine et al. (2004) have suggested that items be grouped by category of teaching behaviour or course elements and that faculty receive only category scores, rather than scores for each item.

As discussed in *Section 4.C.i: Defining Effective Teaching*, teaching is a multi-faceted activity and, as such, some scholars (e.g. Marsh) stress the need for a multidimensional evaluation form to fully assess teaching effectiveness. However, Abrami (2001) has argued that while teaching is a multi-faceted activity, summative decisions about teaching effectiveness are not, and that administrators benefit most from a single score representing a broad and comprehensive definition of teaching effectiveness. Scholars disagree, however, about whether such a single score should be derived from a) a broad question asking students to rate a course or instructor in general (a position advocated by Abrami, 2001), or b) a score calculated by averaging several dimensions of teaching, weighted according to institutional priorities (proposed by d'Apollonia & Abrami, 1997).

Cashin and Downey (1992) agree that scores of individual dimensions are of little use to administrators for the purpose of summative evaluation. Their study of data from 17,183 courses representing 105 institutions found that short, economical forms (e.g. with global questions) were able to capture much of the information needed for summative purposes. However, more detailed questions and results can assist with formative evaluations.

Data Presentation

Some researchers have raised concerns about how statistical summaries of ratings data are presented to faculty and have noted that many who are charged with interpreting the data are not armed with the information or skills to do so effectively. At many institutions, both faculty and administrators are given summary reports which may or may not include information on statistical deviations, details on how the data were prepared, or guides for interpretation and use. As Abrami (2001) and Theall and Franklin (2001) have argued, without these sorts of information, administrators may be making inaccurate, and possibly inappropriate, personnel decisions. To address some of these concerns, Abrami (2001) recommends that normative data be displayed in a visual format (chart, graph) – particularly for those with little or no experience with statistics.

4.D.ii Challenges to interpretation and use for summative purposes

Statistical value or evaluation data

Scriven (1995) highlights a range of common errors related to the use of course ratings data, including the use of scores without regard to distribution; treating small differences as important; and using evaluation data as the primary tool in summative or formative evaluation. Abrami (2001) has raised concerns about the misinterpretation and misuse of evaluation data for personnel decisions. He cautions administrators not to over-emphasize small ratings differences, particularly if they are not well-versed in statistical analysis. Similarly, McKeachie (1997), d'Appolonia and Abrami (1997) and Wagenaar (1995) caution administrators from overestimating the precision of evaluation results and recommend that, rather than using raw scores reported to one or more decimal points for interpretive purposes, administrators should classify scores in one of three broad categories: exceptional, adequate, or unacceptable. These broad categories would mitigate any variation or bias introduced by disciplinary or course characteristics in order to allow for fair comparison between courses or instructors. Furthermore, McKeachie notes that these broad categories would better reflect the ways in which teaching evaluations are used for summative purposes.

Even if such categories are not implemented, administrators should ensure that they can articulate a meaningful distinction between the possible levels of the ratings they review (e.g. between a 3.5 and a 3.6 on a 5-point scale) before using those scores for formal evaluation purposes. McKeachie (1997) also argues that the "presentation of numerical means or medians (often to two decimal places) leads to making decisions based on small numerical differences – differences that are unlikely to distinguish between competent and incompetent teachers" (p. 1223). Administrators should therefore not be given information that is more precise than it is meaningful.

Administrator awareness of research and statistics

Theall and Franklin (2001) suggest that a major challenge to the validity of student ratings is the minimal facility many administrators have in interpreting the results they receive and the lack of training available to them to improve these skills. Menges (2000) concurs, writing that "a great many individuals in the assessment area would assert that no matter how valid and reliable the instrument is, consumers can and do misuse the results from it" (p. 8). Franklin (2001) warns those working with course evaluations:

> *[not to] assume that those who will examine these ratings have the necessary skills and knowledge to use them within the guidelines recommended by ratings experts. … In one multi-institutional study, more than half of the faculty using ratings of the colleagues could not answer basic questions about the common statistics that appear on typical ratings reports, such as means and standard deviations* (p. 86).

Wachtel (1998) similarly cautions that "faculty and administrators have little knowledge of existing research in this area and therefore may … engage in some kind of abuse (for example, according too much significance to the last decimal place in a class average score)" (p. 193).

Franklin and Theall (1989) note, however, that it should not necessarily be the responsibility of administrators and faculty members to develop these statistical skills. They note that "because ratings exist in larger systems, we cannot reasonably expect every end user to be a statistician or have the psychometric skills to evaluate his/her own skill at interpreting ratings" (p. 21). They instead recommended that the users of ratings are provided with "guidelines, warnings, interpretive statements, and comments" to contextualize ratings and guide interpretation (p. 21). Our review of current practices at select North American institutions revealed that seemingly little information is provided to university/college administrators to assist them in the interpretation of evaluation data. We uncovered only a handful of examples of institutions developing or offering training materials or handbooks for this purpose.

Theall and Franklin (2000) and Abrami (2001) have raised concerns about the expertise of those reviewing, interpreting and making decisions based on ratings data, noting that academic administrators are rarely well-versed in the research nor are they trained to effectively interpret evaluation data from their own institutions: a fact that may negatively impact personnel decisions. As a result, Abrami (2001) stresses the importance of increasing "the expertise of individuals involved in decision-making" by reforming the "reporting system and guiding the decision-making process" (p. 64) and provides

recommendations for institutions as to how they may improve judgments about teaching effectiveness when using evaluation data.

McKeachie (1997) and others have strongly recommended that institutions improve efforts to assist students to become better evaluators of teaching and to better train administrators in the interpretation and use of ratings data for personnel decisions. Similarly, Abrami (2001) suggests that faculty distrust of and concern regarding the use of ratings data for promotion and tenure decisions can be addressed/alleviated by reforming institutional reporting structures and ensuring transparency in the decision-making process. If evaluation processes are not standardized across an institution, division, or department, valid comparisons between instructors cannot be made (Hoyt & Pallett, 1999).

Using evaluations for comparative purposes

In order to improve the decision-making process, Abrami (2001) recommends that institutions determine their evaluation strategy in advance, suggesting either norm-referenced or criterion-referenced evaluations. In norm-referenced evaluation systems, individual faculty are compared to an appropriate grouping of other faculty (e.g. based on course type and/or discipline). If evaluation data are to be used for formative purposes, it can be useful for an individual faculty member to know where he or she sits within his or her department, or to understand how their ratings compare to previously taught sections of the course. The alternative to norm-referencing is criterion-referencing, where a standard for performance is set (with or independent of triangulating measures of teaching effectiveness) and instructor performance is compared against this standard (Abrami, 2001).

Scholars disagree whether course evaluations should be subject to comparison. While McKeachie (1997) believes in the validity and usefulness of evaluation results for summative purposes, he is more concerned about their use to make comparisons. He argues that administrators wrongly use ratings data to make comparisons using numerical means or medians. He notes that:

> *Comparisons of ratings in different classes are dubious not only because of between-classes differences in the students but also because of differences in goals, teaching methods, content and a myriad of other variables. Moreover, as I suggested earlier, comparisons are not needed for personnel decisions. To the degree that student ratings enter into such decisions, faculty members can be reliably allocated to three or four categories by simply looking at the distribution of student ratings: How many students rated the teachers as very good or excellent? How many students were dissatisfied?* (p. 1222)

This concern is echoed by Algozinne et al. (2004) and Zabaleta (2007). McKeachie goes on to suggest, however, that such comparisons can be made if only broad ratings are used (see *Section 4.D.i: Reporting of Evaluation Results*).

Cashin (1990) disagrees, arguing that “without comparative data it is not possible to meaningfully interpret student rating data” (p. 2). Cashin’s argument in favour of comparisons, however, suggests that comparisons often act as a proxy or substitute for

careful consideration or review of the evaluation instrument or ratings scales. If institutions knew more about what evaluation questions were asking, or how students respond to evaluations, such comparisons may not be necessary.

For institutions that do compare evaluation results between instructors, Abrami (2001) suggests several means by which the statistical and conceptual errors that emerge from norm-referenced evaluations may be mitigated, including the addition of margins of error and the visual representation of evaluation results (see *Section 4.D.i: Reporting of Evaluation Results*).

Comparing courses

Several scholars argue that evaluation results will better represent the teaching effectiveness of an instructor if possible variations in evaluations due to course characteristics are mitigated by developing an average rating across multiple courses. Abrami (2001) suggests that each instructor should identify a set of courses that balances lower- and upper-level courses, and elective and required courses. Abrami also argues that:

> *[s]ince summative decisions are often based on a collection of [course evaluations], the mean, variance, and sample size for an individual faculty member should be combined from several courses… Individual course results may be more useful for formative purposes, whereas combined course results are more useful for summative purposes* (p. 72).

Franklin (2001) agrees that "averaged results from comparable courses taken over several semesters are likely to be considerably more reliable for comparisons than those from single courses" (p. 92). Franklin further notes that "the number of courses required to construct 'average' results increases as the class size decreases. Generally, five or more courses are recommended in most cases, although very small classes certainly need more. For example, courses with as few as five students may need twenty sections for comparison" (p. 92).

Abrami nonetheless cautions that clear policies to determine which courses will be included or excluded in an overall rating should be developed and should be implemented equitably and consistently.

Comparing instructors

The research suggests that because of disciplinary differences in teaching styles and goals, teaching evaluations – if compared at all – should be compared only between instructors in the same or similar disciplines. Any comparisons between instructors should also provide ample opportunities for contextualization of the data, and should ensure that the courses being compared share similar characteristics (see *Section 4.E.i: Overview of Studied Variables* for a description of course characteristics that might affect evaluation results).

Neumann (2001) highlights the different definitions of effective teaching in different disciplines. As an alternative to conducting evaluations generically across diverse departments, she notes the work of other scholars who propose "the development of

discipline-specific teaching evaluation instruments" or "the development of a number of instruments which reflect the variety of teaching philosophies suited to the diversity of disciplines" (p. 143). She also refers to her previous work which "highlights how rating results from generic instruments can be used by universities in a manner that recognizes disciplinary variation" (p. 143).

Even at the department level, variations by instructor may affect the validity of comparative data (Addison, Best & Warrington, 2006). For example, Theall and Franklin (2001) discuss that a particular example of gender bias could be explained by the fact that most of the required, lower-level courses in the department were taught by women. What appeared to be a gender bias in evaluations was actually a reflection of the fact that particular instructors taught courses with particular characteristics. This reinforces the need to ensure that data are presented from a representative or equitable selection of courses.

4.E Internal Validity: the influence of variables on evaluation results

4.E.i Overview of studied variables

A great deal of attention has been paid in the research to the wide range of factors that may or may not impact the validity of student evaluation data (Zabaleta, 2007; Addison et al., 2006; Algozzine et al., 2004; Ory & Ryan, 2001; Theall & Franklin, 2001; Ali & Sell, 1998; Wachtel, 1998; Cashin, 1995, 1988). Information in this section is primarily drawn from these reviews (see Appendix E for some examples of summaries on research on potentially biasing variables from these reviews).

The variables discussed in the literature fall into four categories: administrative conditions, course characteristics, instructor characteristics and student characteristics. The chart below details the specific factors that fall under each of these categories.

It should be noted that any effect on overall ratings from any of these particular variables, even when statistically significant, is almost always very small – often changing the ratings by less than one-tenth of 1%. Because of this, even if these variables do have an effect on evaluation outcomes, validity can almost always be maintained by reporting scores to no more than one decimal place or as part of a broad category, as described in *Section 4.D.i: Reporting of Evaluation Results*.

Other strategies for managing variables and ensuring that they do not impact overall validity are discussed in *Section 5.C: Ensuring Utility*.

Table 2: Researched Variables

Category	Variable Items
Administrative conditions	Timing of evaluations
	Instructions to students
	Anonymity
	Presence of instructor
	Purpose
Course variables/characteristics (those that cannot be controlled by the instructor)	Class size
	Time of day
	Elective/Required course
	Workload/Difficulty
	Course level
	Discipline
Instructor variables/characteristics	Age
	Research productivity
	Race
	Personality/popularity
	Expressiveness
	Rank and Experience
	Gender
Student variables/characteristics	Age
	Gender
	Year of study
	GPA
	Personality
	Gender
	Motivation
	Attendance
	Grades

Overall, the research into these variables is overwhelming and inconsistent in quality and scope. A select number of recent and comprehensive reviews of this research provide a fair summary of previous studies. These include Algozzine et al. (2004), Ory and Ryan (2001), Wachtel (1998), Ali and Sell (1998) and Marsh and Roche (1997). In general, no variables have been found to have a substantial effect (e.g. something that would alter the ratings beyond the second decimal place) on ratings, except for expected grades. Some studies (cf. William & Ceci, 1997 and the "Dr. Fox" study) have identified factors that appear to reflect bias (e.g. presentation skills, instructor enthusiasm or personality); however, these studies have been largely discounted either for methodological reasons

or because these factors may actually measure improved teaching. In discussing the long list of variables that have been shown to influence student ratings to varying degrees, Algozzine et al. (2004) have argued that they cannot be viewed as biasing variables unless they alter ratings without measuring differences in teaching effectiveness. Similarly, d'Apollonia and Abrami (1997) conclude that even though administrative, course and instructor characteristics may influence ratings, they do not result in this definition of bias and therefore do not reflect invalidity in ratings. Cashin (1988) and Marsh (1984) have also argued that the only variables that can possibly introduce bias are those that are "not a function of the instructor's teaching effectiveness" (Cashin, 1988, p. 3) – for example, class size. Cashin goes on to note that these variables "may impact teaching effectiveness, but instructors should not be faulted if they are less effective teaching large classes of unmotivated students than their colleagues are with small classes of motivated students" and that such factors should "be controlled for by using appropriate comparative data" (p. 3).

As noted, the wide range of variables have been thoroughly examined and re-examined in the literature. The scope of this study does not permit us to provide a comprehensive review of all of the researched variables; instead, the following highlights some of the variables that are either more contentious and are actively debated in the literature or those that have resulted in particularly interesting findings.

Administrative conditions

Timing of evaluations: In general, the timing of evaluations has demonstrated no significant impact on evaluation ratings (Wachtel, 1998). There is some evidence to show that when evaluations are completed during final exams, results are lower (Ory, 2001); therefore, most scholars recommend that evaluations be administered before final exams and the submission of final grades (d'Apollonia & Abrami, 1997).

Articulating evaluation goals and providing instructions to students: Stating the purpose of evaluations (e.g. noting that they will be used for personnel decisions) may positively impact results (Algozzine et al., 2004; Cashin, 1995); however, the results on this variable have been mixed. Cashin (1995) suggests that this can be controlled through the use of standardized instructions. Fox (2006) has noted that ratings can be improved when instructors request more critical feedback from their students.

Anonymity: Students' concerns regarding potential academic repercussions appear to increase when they are asked to sign evaluation questionnaires; thus, signed ratings tend to be higher. Therefore, most scholars recommend that they remain anonymous (Cashin, 1995) while some have suggested they instead be confidential (the institution, but not the instructor, would be able to identify who completed the evaluation) to encourage and ensure that students provide responsible evaluations and to allow for future follow-up (Wright, 2006).

Presence of instructor during administration of evaluations: Ratings appear to be higher when an instructor is present during their administration; however, this can be controlled by ensuring that the instructor leaves the room while students complete the forms (Cashin, 1995). Algozzine et al. (2004) notes that instructor

presence does not significantly impact validity unless this practice is combined with non-anonymous ratings.

Course characteristics

Class size: Although some studies have found smaller classes often receive slightly higher evaluation ratings (Algozzine et al., 2004; Williams & Ory, 1992; Centra & Creech, 1976), the correlation between class size and ratings is statistically insignificant and is therefore not viewed as having any impact on validity (Marsh & Roche, 1997; d'Apollonia & Abrami, 1997; Aleamoni, 1997; McKeachie, 1997; Cashin, 1995, 1988; Marsh, 1987). McKeachie (1997) notes that there is evidence to suggest that faculty teach better in smaller classes, which would make any effect on ratings a sign not of bias but an accurate reflection of teaching effectiveness. However, because instructors may not have much agency over class size, care should be taken to either contextualize class size in evaluation data reports or to make sure that instructors whose results are being compared also have comparable average class sizes.

Elective/required: Students frequently rank electives somewhat more positively than required courses; however, this has not been found to have a significant impact on ratings (Algozzine et al., 2004; Cashin, 1995, 1988). The status of an instructor's courses as required or elective should be managed similarly to class size in summative reporting of evaluation results.

Workload/course difficulty: Although many faculty believe that harder courses or higher workload results in lower evaluations, this has not been supported by the research which has produced inconsistent results (Marsh, 1987). "Easy" courses are not guaranteed higher evaluations. Additionally, some studies have shown that difficult courses and/or those with a higher workload receive more positive evaluations (Cashin, 1988).

Course Level: Research findings have suggested that the level of the course can impact ratings (Algozzine et al., 2004; Marsh, 1997; Cashin, 1988, 1995; Aleamoni & Hexner, 1980) with some evidence demonstrating that higher level courses may receive higher ratings. Again, this information must be contextualized in evaluation data reporting.

Discipline: Some studies have shown that particular disciplines receive higher ratings (with the most positive being received in the humanities, followed by the social sciences and then the natural sciences (Johnson, 2003; Neumann, 2001; Ory, 2001; Wachtel, 1998; Cashin, 1990). This reflects disciplinary differences in teaching styles and goals rather than a source of bias. Neumann (2001) and Cashin (1995, 1988) caution that comparisons across disciplines may therefore not be accurate.

Instructor characteristics

Personality/popularity: Ali and Sell's (1998) review of the literature on the popularity or personality of an instructor shows that there is general agreement that this has insignificant impact on evaluation results. Two studies, one published by Naftulin, Ware and Donnelly (1973), also known as the "Dr. Fox"

study, and another by Williams and Ceci (1997) concluded that instructor enthusiasm can impact evaluations. However, both of these findings have been widely refuted on methodological grounds by most scholars in recent years (Abrami, 2001; Kulik, 2001; Marsh & Dunkin, 1992). Abrami (2001) and Theall and Franklin (2001) have argued that there is no research to substantiate the claim that popularity or personality results in higher ratings, and Ory (2001) argues that "personality" may actually measure teaching behaviours, such as enthusiasm, that may in fact influence teaching effectiveness.

Expressiveness: The research on instructor expressiveness, like that surrounding personality and popularity, is complicated and sometimes unsound. Some studies have established clearly that expressiveness tends to enhance learning and therefore cannot be considered a biasing factor (Cashin, 1995).

Rank and experience: d'Apollonia & Abrami (1997) and Arreola (2000) find that these variables do not significantly affect evaluation results. Marsh (2001) found that experience does not lead to improved ratings and may in fact have a negative relationship with teaching effectiveness.

Gender: In general, studies relating to gender have produced inconclusive results, but most have shown that this variable has little or no impact on evaluations (Algozzine et al., 2004; Theall & Franklin, 2001; Marsh & Roche, 1997; Cashin, 1995; Arreola, 2000; Aleamoni & Hexner, 1980).

Student characteristics

Gender: There is some evidence to suggest that students tend to rate instructors of the same sex slightly higher (Ory, 2001). This is only significant in disciplines with substantial gender imbalances, in which case this factor may usefully be contextualized when data is presented.

Motivation: Student motivation or prior interest in the course may impact ratings, resulting in higher evaluations (Cashin, 1988/95). The University of Michigan uses a question about motivation to contextualize ratings data.

Attendance: A recent study which surveyed over 9,000 Israeli college students found that there was a positive relationship between high attendance rates and positive course ratings. In general, this was not viewed as a biasing variable because greater attendance leads to improved learning (Davidovitch & Soen, 2006). It should be noted that this has been the only full-scale study examining this issue that we located.

Grades: Expectations, Inflation and Leniency – Myth or Reality?

Perhaps the most controversial variable discussed in the research is the grades-ratings relationship. Do students' expectations regarding their final grade impact their ratings of an instructor's teaching effectiveness? A recent study by Baldwin and Blattner (2003) found that 40% of faculty believe this to be true. This question has received a great deal of attention from the research and is still a matter of much debate. Aleamoni (1999) has identified 37 studies that revealed correlations between expected/received grades and

positive ratings and 24 studies that found no significant relationship. (For a recent review of the literature on the grades-leniency hypothesis, see Gump, 2007).

Some studies have found a relationship between positive evaluations and grades. This correlation has been interpreted by some as a clear indication that grading leniency can result in improved evaluations (Wachtel, 1998). Greenwald and Gillmore (1997) have argued that since student expectations regarding final grades impacts their evaluation of an instructor, ratings should be statistically adjusted to correct for this factor. Abrami (2001) and others have refuted this claim, arguing that the impact is not substantial. Abrami argues that neither lenient nor harsh grading practices impact course ratings in any statistically meaningful way. Similarly, Marsh (1987) and Marsh and Roche (1997) have argued that while grade expectations may reveal a level of bias, the impact on ratings is weak and relatively unsubstantial.

McKeachie (1997) and others have expressed concerns about Greenwald and Gillmore's conclusions of their 1997 study on grading leniency, suggesting that their argument is flawed. In a re-examination of Greenwald and Gillmore's data sets, Marsh and Roche (2000) found that higher evaluations were given to those courses and instructors with higher workloads.

Heckert et al. (2006) review some of the studies on the grades-evaluation relationship, noting the conflicting opinions in the literature. Their particular study tested the grading leniency hypothesis in a study of 463 students by examining the impact of two variables: class difficulty and student effort. Heckert and colleagues found that higher evaluations were given to courses in which the difficulty level met students' expectations. In addition, evaluations were also positive when students indicated they had expended more effort than anticipated. Overall, this study concluded that more demanding instructors received higher evaluations and therefore refuted the grading leniency hypothesis and the notion that faculty could "buy" better evaluations with higher grades.

Wachtel (1998) and others (Marsh & Dunkin, 1992; Murray, 1987) have suggested that a positive correlation between expected grades and instructor ratings might simply be evidence of student learning: students both expect higher grades and rate faculty more positively when they have had a positive classroom experience. Alternatively, Chambers and Schmitt (2002) posit a comparison process model to explain the relationship between grade expectations and evaluations. In this theory, students base their grade expectations on experiences in other courses (workload, effort and final grade). If the comparison is positive they produce positive ratings; if, however, it is negative this will be reflected in their evaluation of the instructor. Addison et al. (2006) refute this hypothesis, pointing to the results of their small study which concluded that grade expectations are also influenced by pre-conceived notions of whether or not a course will be hard or easy. Their survey of students indicated that those who found the course more difficult than originally expected rated the course less favourably, while those who found the course easier than anticipated ranked it more positively. Addison and colleagues also concluded that the effect of perceived difficulty was independent of the grade students earned in a class, thus indicating that faculty grading practices have a limited impact on evaluation results.

In reviewing the research that focuses on the grading-leniency hypothesis, Gump (2007) questions the generalizability of the results from these studies which are often contradictory. In particular, he points to such concerns as study methodology, applicability of results beyond a particular institution (i.e. the ability to replicate findings) and differences in the use and definitions of key terms used in the research (e.g. bias, workload).

4.E.ii Validity testing

Ory and Ryan (2001) identify five primary means through which the validity of course evaluations have been assessed: multisection, multitrait-multimethod, bias, laboratory and dimensionality studies. Ory and Ryan argue that only the first three methods have contributed to an understanding of the validity of course evaluations. They dismiss laboratory studies as an appropriate means of assessing validity because of the artificial environment in which they are conducted. They also note that dimensionality assessments – studies that attempt "to identify a 'common' set of factors underlying the construct being measured by student ratings of instruction" – have not been able to "identify a single set of dimensions and merely support the notion that students view instructional quality as multidimensional" (p. 31) (as discussed in *Section 4.C.i: Defining Effective Teaching*).

Importantly, Ory and Ryan (2001) note that most of the tools used to assess the validity of student ratings have successfully focused on the degree to which evaluations match other means of teaching effectiveness and on identifying any external influences on ratings. Studies, however, tend not to evaluate the ways in which ratings are interpreted and put to use by students, faculty and administrators. As these elements can significantly affect the validity of ratings instruments even if the items on the instruments themselves have been carefully tested as described in *Section 4.C.ii: Developing Evaluation Instruments*.

Multisection validity studies

As described by Greenwald (1997) and Ory and Ryan (2001), a common means of assessing the validity of evaluations are multisection studies. These studies compare the academic performance of students in different sections of the same course and compare this academic performance with evaluation ratings. In general, Ory and Ryan note, "multisection validity studies have shown substantial correlations with student achievement as measured by examination performance" (p. 30). However, these studies have been criticized for two reasons: first, because they must assess courses with multiple sections, they generally evaluate only lower-level courses. Ory and Ryan (2001) have argued that the learning goals in these courses are different from those in upper-level courses and therefore that the conclusions drawn from these studies cannot be generalized to evaluations in upper-level courses. Second, assessment in these large courses and appropriate to these studies is often limited to multiple-choice tests that may not measure a wide range of learning objectives, again suggesting that the generalizability of these studies may be limited.

Multitrait-multimethod studies

These studies compare student ratings with other means of evaluating teaching, including alumni surveys, evaluations by colleagues and self-ratings. These studies may

also include multiple means of assessment (for example, content analysis of a teaching dossier or focus groups). Ory and Ryan (2001) argue that these studies have generally shown substantial correlation between the evaluations received through these multiple means.

Bias studies

These studies use factor analysis to identify any external or environmental influences on student ratings. Ory and Ryan (2001) note that "numerous studies have been conducted to determine relationships (or lack thereof) between ratings and a wide range of potential influences" but that "the research literature reveals few, if any, potentially biasing influences on the rating process" (p. 31). They also note that the results of these studies are themselves not always valid or conclusive.

Means of validating student responses

The accuracy of student ratings is generally assessed through the comparison of student ratings with other measures of teaching effectiveness, particularly student academic performance (Ory & Ryan, 2001; Theall & Franklin, 2001; Wachtel 1998). Such research often correlates final grades (as an indicator of student learning) with evaluation results. Some scholars have argued that high evaluation scores are indicative of student learning; however, as noted in *Section 4.E.i: Overview of Studied Variables*, others have suggested that high scores may be a result of some other factor (e.g. lenient grading on the part of the instructor). This model of assessing the accuracy of student ratings has been criticized by a small number of researchers because these studies, for methodological reasons, have focused only on lower-level courses that rely on standardized assessment (Ory & Ryan, 2001). Abrami's (2001) review of the research concludes that there is ample empirical evidence to demonstrate that course evaluation data can and do indicate learning.

Faculty also frequently express concern that students are easily manipulated into providing higher ratings through grade inflation or particularly charismatic instructors (Theall & Franklin, 2001). A large number of studies on these issues have been conducted and are described above in *Section 4.E.i: Overview of Studied Variables*. In general this research shows that while evaluation results may appear to demonstrate that students reward lenient and personable instructors, the actual relationship between these factors and evaluation ratings is substantially more complex and that, in general, the accuracy of student ratings is upheld.

Student responses are also verified through comparisons with the ratings of other assessors: Ory and Ryan (2001) describe how "research has detected high positive correlations between student ratings and alumni ratings … and moderate positive correlations between student overall ratings and self-ratings … and peer ratings" (p. 36). Similarly, Murray (1987) has found that student ratings of instruction are comparable to those made by trained observers. Arreola (2000) Aleamoni (1987) and others have shown that student evaluations are consistent and stable and correlate with colleague ratings/peer observations.

According to Nasser and Fresko (2002), Ory (2001) and Remedios and Lieberman (2008), a common faculty concern about the validity of student ratings of instruction is the ability of students to accurately assess the value of an educational experience before putting their knowledge from the course to use in other courses or in their careers.

However, several studies comparing alumni ratings with student ratings indicate that a student's assessment of a course does not change substantially over time (Ory & Ryan, 2001; Theall & Franklin, 2001).

Contextual validity

Theall and Franklin (2000) have introduced a range of context-based variables that may impact validity that have not yet been fully explored in the literature. These variables include changing instructional practices, changing student populations, changing faculty needs, changing institutional priorities, changing technology and data requirements and changing faculty development and evaluation practices. Most evaluation forms were developed when lecture-based teaching was the norm. However, in recent years, teaching practices have shifted to include collaborative learning techniques, active and problem-based learning and increased use of academic technology. Existing evaluation instruments may no longer accurately or adequately assess these new teaching and learning contexts (see *Section 3.C.ii: Questionnaire Format and Content*). Some institutions may already be addressing this concern through the use of customizable forms that allow faculty to select appropriate items while other institutions have responded by extensively revising their evaluation instruments.

A large portion of the research on course evaluations was conducted on a population of students that is no longer representative of today's undergraduates. These demographic shifts (in age, ethnicity and socio-economic status) may impact student attitudes toward teaching effectiveness and consequently the ratings they give instructors and courses (Theall & Franklin, 2000). Given these contextual changes and the fact that most research to-date has not adequately considered them, Theall and Franklin (2000) raise concerns about making generalizations regarding course evaluation systems based on the current scholarship.

Potential sources of validity for student course evaluations include:

- The positive and statistically significant correlation of ratings with student learning;
- The unique position and qualifications of the students in rating their own increased knowledge and comprehension;
- The unique position of the students in rating changed motivation toward the subject matter taught and to a changed general attitude toward further learning in the subject area;
- The unique position of students in rating observable matters of fact relevant to competent teaching (e.g. punctuality of the instructor);
- The unique position of the students in identifying the regular presence of teaching style indicators (e.g. enthusiasm, encouragement of students); and,
- Students are in the best position to judge whether tests covered course content.

[adapted from Scriven, M. (1995). Student ratings offer useful input to teacher evaluations. *Practical Assessment, Research & Evaluation*, *4*(7), 4-5.]

Section 5

Implementing Effective Evaluation Measures: Recommendations from the Research

5.A Introduction

As discussed in *Section 4: Reliability, Validity and Interpretation,* a substantial element of evaluation validity is the policies and practices surrounding the creation, administration and interpretation of evaluations. The recommendations below detail actions and policies an institution, division, or department may wish to implement to ensure the validity and utility of evaluation. The recommendations complement and draw on several useful articles which provide a series of recommendations to institutions to implement valid and equitable course evaluations. These include Moore and Kuol (2005), Franklin (2001), Ory and Ryan (2001), Theall and Franklin (2001) and Cashin (1990).

Collected in Appendix F are a number of guidelines for good evaluation practice drawn from current research.

5.B Ensuring Validity

Research on student evaluations identifies several recommendations to ensure that course evaluations can provide valid data for formative and summative evaluation of teaching:

Set clear evaluation goals, including clear definitions of what constitutes effective teaching at your institution and ensure that questions reflect these goals

Section 4.C.i: Defining Effective Teaching describes the importance of ensuring that evaluation questions match institutional teaching priorities and provide adequate information to make the kinds of summative assessments for which the instruments are being used. The identification of teaching measures to be evaluated and the development of evaluation questions should be viewed as an opportunity to encourage an institution-wide discussion about teaching goals and evaluation uses. To ensure that questions can provide meaningful feedback to instructors and can be used in the summative evaluation of teaching, the questions that are ultimately selected should measure aspects of teaching that reflect these conclusions.

Design and test instruments according to rigorous theoretical and psychometric standards

The development of evaluation instruments should be a serious and substantial process involving many members of the institutional community. Questions should

be selected carefully according to well-developed theoretical and research-based constructs. Scales must be logical and clearly explained. Instruments should be approved by an appropriate committee or governance body through a transparent and consultative process. Approved instruments should be evaluated by experts in survey construction and continuously investigated through institutional research (See *Section 5.C.iv: Ensuring Utility for Institutions*). If an institution cannot devote the time or expense to developing a rigorous in-house instrument, it may wish to consider licensing a validated instrument from another institution.

Establish appropriate and standardized policies and processes for the administration of course evaluations

Clear and consistent policies and processes must be developed to ensure that the ratings collected are not compromised. This includes ensuring that:

Policy and practice about the administration of evaluations is standardized at the administrative level at which comparison between instructors or courses (if employed) is made

Many threats to validity are introduced through inconsistent administration of evaluations. This might include issues such as instructor presence during evaluations, inconsistent evaluation forms, or conducting some evaluations online and others on paper, among others. By ensuring that policies about the administration and reporting of evaluations are equitable and are applied consistently, institutions can make dramatic strides towards improving evaluation validity.

Each course achieves an appropriate response rate

Cashin (1990) recommends collecting feedback from at least 10 students and at least two-thirds of the class, whichever is higher. As described in *Section 4.D.i: Reporting of Evaluation Results*, to further ensure that evaluation results are representative, several scholars suggest averaging some or all of an instructor's evaluations to ensure that the responses collected provide an accurate representation of their teaching.

The anonymity of student responses is protected

There is little data to suggest that anonymous responses are any more or less accurate or valid than non-anonymous student responses.

Wright (2006) has argued that anonymous ratings absolve students of responsibility for their statements and opinions, and that "[w]ith no possibility for follow-up, students need not think through their decision" (p. 419). Wright notes that anonymous evaluations are intended to ensure that students are not reprimanded by faculty for negative comments. However, he argues that while the intentions behind protecting student anonymity may be positive, such a system effectively places more trust in students than faculty. He further raises concerns that students may use evaluations to vent anger or disappointment regarding low grades (notably, Wright does not point to any specific studies to support this theory).

Wright also suggests that anonymity may encourage abuse of evaluation instruments and the process of administration, hypothesizing that "students could enter the room and fill out evaluations who were not even in the class" (p. 419). To address this problem, Wright recommends that evaluations be confidential, with names stripped from the data before being viewed by faculty, so that students can be tracked by the administration to allow for follow up (e.g. to investigate an extremely high or low ranking or to identify variables that contribute to high or low rankings).

However, research does indicate that students may be uncomfortable providing non-anonymous data and that non-anonymous student responses yield somewhat higher ratings (Wachtel, 1998). Consequently, policies protecting anonymity should be applied consistently and uniformly as there is much to lose by jeopardizing the already minimal student trust of the evaluation system. Practice should also ensure that students understand that and how their anonymity will be protected.

An appropriate amount of data is distributed to appropriate populations and that appropriate and consistent policies for access to and storage of data is developed

Students, faculty and instructors each benefit from and require different data derived from course evaluations. Wachtel (1998) argues that students deserve to see the result of their input in the form of publicly distributed evaluation results. Many institutions who do share evaluation results publicly (see *Section 3.F.iii: Publication of Results*) choose to highlight a small number of global questions to distribute to students to assist with course selection. A number of institutions publish evaluation results, primarily to provide students with information to assist in the course selection process. See *Section 5.C.i: Ensuring Utility for Students* for further recommendations about sharing results with students.

Administrators should receive appropriate individual and comparative data that matches how they will use evaluation data. Administrators who are not providing diagnostic or formative feedback may require only data from the summative global survey items (see *Section 4.D.ii: Challenges to Interpretation and Use for Summative Purposes*). Instructors may receive further results that can be used for formative purposes. The data that administrators receive should match their facility with statistical and data analysis. Evaluation results should be accompanied by any additional information necessary to adequately contextualize the data (for example, interpretive guides, comparative means, or written narratives by faculty members; see *Section 5.C.iii: Ensuring Utility for Administrators*).

Individual faculty members should have access to all course evaluation data collected about their teaching, including anonymized student written comments. Instructors should also be provided with appropriate data summaries (see *Section 4.D.i: Reporting of Evaluation Results*) that help to contextualize the data they receive.

Institutions should maintain centralized records of teaching evaluations (see *Section 3.F.i: Who has Access? To What?* for examples of how evaluation data is maintained at several institutions). Originals should be retained for a limited amount of time but long enough to verify any contested results. Processed data should be retained confidentially by departments, divisions, or in a centralized database for as long as they may be used by instructors and institutions.

5.C Ensuring Utility

5.C.i For students

For evaluations to be accurate, students must be given enough information to adequately provide useful and appropriate responses. Consequently, policies and practice about course evaluations must address means by which an institution can:

Provide sufficient information to students about the administration and use of evaluations

Ory (2001) cites studies that show students provide more constructive, thorough, accurate and positive evaluations when they have been educated about the goals and uses of course evaluations (though Wachtel (1998) argues that studies on this variable are inconclusive). This occurs because students generally complete evaluations only at the end of the course and do not have an opportunity to see any effect from their efforts. Beran and colleagues (2007, 2005) and Wachtel (1998) note that, consequently, students often feel that their evaluation results are not reviewed and that their suggestions are not implemented. Students also occasionally feel that their anonymity is not protected when the process of data collection and storage is not properly explained, particularly with online course evaluations (which frequently request some form of authentication even if results are stored only in aggregate). As noted in *Section 2.C.iii: Student Perceptions of Course Evaluations*, this can affect evaluation results.

Students should be provided with thorough information about the uses of evaluations for teaching development and assessment and the role of teaching evaluations in career progression, hiring and the tenure process and about evaluation data storage and access. Instructors may also discuss any ways in which they have made changes to courses or to their teaching based on previous evaluations.

Svinicki (2001) suggests several ways in which instructors can discuss evaluations and help students understand what kinds of responses are most helpful to instructors and administrators.

Provide students with access to appropriate evaluation results

The question of whether aggregated evaluation results should be shared with students is surprisingly complex. As noted in *Section 4.D.i: Reporting of Evaluation Results*, Wachtel (1998) argues that students, having contributed to the teaching assessment process, deserve to see the results of their input. However, several studies (as reviewed in Wachtel, 1998) have suggested that an instructor's "reputation" (which may be derived from published evaluation results) can influence

student responses on future evaluation iterations. None of these effects, however, indicate that the validity of student evaluations are compromised by sharing results with students; rather, they simply indicate that evaluation results must be shared consistently so that any influence on evaluation results is consistent if evaluation results are being compared between courses or instructors.

Several schools seem to balance these considerations by providing access to aggregate data for a limited number of evaluation questions (see *Section 3.F.iii: Publication of Results*). These questions are generally broad, global questions that may have limited influence on student expectations about particular instructor traits.

In a study of 1,229 students, Beran et al. (2005) found that 52% of students had never consulted or used course evaluation ratings (primarily because they were unaware of their existence), while 47% reported using them to select courses and/or instructors. These results suggest that better publication and improved access to evaluation results may be necessary even at institutions that make results available to students.

Offer students other means to provide feedback

Because of the importance and value of helping students understand the evaluation process and the impact of the feedback they provide and to counteract student skepticism about evaluations, mid-course evaluations can significantly improve students' faith in evaluations (Wachtel, 1998) and ability to provide useful feedback. Mid-course evaluations, particularly when instructors discuss the results of evaluations with their students, help them understand how their feedback is interpreted and incorporated into changes to the course or to an instructors' teaching improving their perception of the value and utility of evaluations and leading them to provide more constructive feedback (Svinicki, 2001). Lewis (2001) also shows that conducting mid-course evaluations can improve ratings on end-of-course evaluations, as students become more able evaluators and more engaged in the course.

A number of authors provide guidance on conducting mid-course evaluations, including Lewis (2001) and Felder (1993).

5.C.ii For instructors

Several studies have concluded that a majority of faculty view student evaluations of teaching negatively or even with hostility (Nasser & Fresko, 2002; Abrami, 2001; Theall & Franklin, 2001; Centra, 1993). However, Beran et al. (2005) found that most faculty viewed ratings systems positively but that few faculty actually used the results to make changes to their courses or to their teaching. This is supported by the findings of other researchers whose studies indicate that ratings data often have little impact on teaching effectiveness or performance (Campbell & Bozeman, 2008, Marsh, 2007; Centra, 1998) particularly when they are provided without the benefit of consultation. With this in mind, institutions should therefore:

Request an accompanying narrative from faculty

As we saw in *Section 3.G.iv: Tenure, Promotion and Merit*, faculty are regularly asked to provide summary data (often in teaching dossiers) for promotion and tenure and for annual merit reviews. However, Ory (2000) notes that ideally, "[a]ssessment is more than counting, measuring, recording or accounting. It promotes teaching evaluation not as a scientific endeavour, with absolute truth as its goals, but rather as a form of argument where the faculty use their data to make a case for their teaching" (p. 17). Franklin (2001) suggests that this understanding of the use of evaluation data can be facilitated if faculty are given the opportunity to contextualize their ratings results with a narrative that highlights particular aspects of the course (e.g. experimental assessment techniques) that may clarify particular evaluation results. Faculty may also contextualize results within their ongoing teaching development, highlighting areas of improvement or changes made to the course or teaching methods as a result of previous evaluations. Franklin argues that such a narrative will "improve the odds that reviewers will consider your students' opinions in the full context of the complex factors that shaped them" (p. 85) and will help reviewers avoid common misinterpretations and misuses of data.

Use evaluation data as a means of providing formative feedback

Evaluation results, particularly those derived from instruments that measure specific teaching behaviours or elements of the course, can provide valuable diagnostic feedback of an instructor's or of a course's particular areas of strength and weakness. Qualitative feedback, in the form of written responses to open-ended evaluation questions, can also provide useful and specific information (see *Section 2.D.ii: Collecting and Interpreting Qualitative Feedback*).

There is ample evidence, however (as discussed in *Section 2.E: Common Uses for Course Evaluation Data*), that simply reviewing evaluation results is not enough to lead to improved teaching. For this, consultation on evaluation results (described below) is necessary.

See *Section 6.A.iii: Increasing Use of Evaluations for Formative Purposes* for suggested adaptations to evaluation instruments to ensure their utility for formative evaluation of teaching.

Encourage and provide the infrastructure for consultation on teaching evaluations

As we have seen, there is evidence to suggest that access to diagnostic data has substantially more impact when combined with consultations (with faculty development personnel or department heads). Lang and Kersting (2007) and Marsh (2007) study the impact of student ratings feedback on teaching improvement efforts and conclude that when evaluation data is not accompanied by some form of consultation the long-term effect is minimal. Their studies, conducted over four semesters and 13 years, respectively, demonstrated that while evaluation data alone may have an immediate positive impact on instructors, this is not sustained over time and in fact decreases fairly rapidly.

Hodges and Stanton (2007) note that when faculty receive assistance in analyzing evaluation results, they are more likely to view evaluations more positively and "as part of a scholarly approach to teaching" which can in turn "form the basis for

effective changes in our teaching approach,, and may inform our thinking about curricular issues as well" (p. 280). Moore and Kuol (2005) suggest a range of practical strategies for faculty in reviewing their ratings data aimed at helping them to manage their reactions and focusing their efforts on using the evaluations to improve their teaching performance (see table at the end of this section).

Penny and Coe (2006) have identified a number of strategies that faculty developers or colleagues can use to ensure effective consultation. These include: actively involving the faculty member in the process; using multiple sources of information (ratings, in-class observations); providing opportunities for faculty to interact with their peers; allowing sufficient time for dialogue and interaction (between the consultant and faculty member); using instructor self-ratings; using high quality feedback information; examining and understanding the faculty member's approach to teaching (e.g. philosophy and pedagogical strategies); and the setting of improvement goals for the faculty member.

Provide an opportunity for instructors to receive individualized assessment

Wright (2006) argues that "[i]t is frequently the case that all faculty are evaluated in the same fashion, whether they have been teaching for one or 15 years" (p. 420). He suggests instead that evaluation systems be adapted to reflect faculty rank. Beginning instructors may receive more comprehensive feedback and may be evaluated on a number of teaching measures, while tenured and very experienced instructors may benefit from more targeted feedback that reflects their individual teaching goals.

Hoyt and Pallett (1999) have outlined a comprehensive evaluation schedule for institutions, with suggested procedures for first-year instructors and particular groups of faculty (such as non-tenured and tenured faculty). For those faculty in their first year of appointment, they recommend that student evaluations be conducted for all courses along with at least one formative review from a colleague, thereby allowing department chairs to assess any areas for improvement quickly. For those heading toward tenure, they recommend that student ratings be collected for all courses at least twice in a five-year period (once early in their appointment and the other for their most recent teaching activity). In addition, formative ratings should be collected for one or two courses each year up to the tenure year.

Provide faculty with information about evaluation data collection and use

There is clear evidence to indicate that institutions are not doing enough to inform and educate faculty about policies and procedures relating to the collection of evaluation data. More specifically, there is inconsistent and often limited effort to ensure that faculty members understand how data are collected, analyzed and reported. Reviews of institutional materials along with results from surveys of university/college administrators reveal that those responsible for personnel decisions (be it for annual merit, promotion, or tenure) are not regularly ensuring transparency in the processes related to the administration of course evaluations. As Abrami (2001), Kulik (2001) and others have shown, educating faculty about course evaluations helps to debunk longstanding myths and misconceptions and alleviate fears about how data may be used by administrators. In addition, faculty who have a

better understanding of institutional expectations are more likely to seek out information and assistance in improving their teaching effectiveness.

5.C.iii For administrators

Administrators are the primary users of ratings data and require substantial training and support in order to effectively implement and interpret evaluations. To assist administrators with these tasks, institutions should:

Use evaluation data for summative purposes

Beran et al. (2005) found that a high majority of administrators (84% in a study of 52) find course evaluations to be a useful source of information (though their subsequent 2007 study found that only 31% believed evaluations were a valid indicator of teaching quality) particularly for personnel decisions. Beran, Violato and Kline (2007) found general agreement among administrators that evaluation data could be effectively used to determine the quality of teaching, to allocate merit and to reward teaching excellence. Evaluation data is valuable and valid enough that administrators can be confident in using it to make summative assessments of teaching effectiveness, with several caveats.

As noted earlier, d'Apollonia and Abrami (1997) and McKeachie (1997) have argued that though ratings can provide useful information about teaching, they should only be used by administrators to make "crude judgments." They agree that for summative purposes, tenure and promotion committees do not need to categorize teaching performance beyond defining it as exceptional, adequate, or unacceptable.

In the study conducted by Beran, Violato and Kline (2007), a significant portion of administrators (23%) felt ratings should be contextualized and supported by supplementary information. The use of evaluation data for summative purposes should also address the recommendations made below about the presentation of evaluation data and for the education of data users.

Educate and train administrators

Several studies referenced in this review indicate that while administrators may use student evaluation data for various purposes (chief among them personnel decisions), administrators are not familiar with the research on evaluation validity and best practices. Abrami (2001), Theall and Franklin (2001), Beran, Violato and Kline (2007), Beran et al. (2005) and others show that administrators often lack general understanding of how best to interpret and apply data from ratings. This is a cause for concern given that evaluation data regularly informs personnel decisions. While it is unreasonable to expect that administrators attain a thorough understanding of this vast field of higher education scholarship, they would benefit from a basic knowledge of the key issues as they pertain to the particular ways in which they use such data. This does not necessarily require detailed knowledge of statistical analysis, but it does require a basic understanding of how the tool works and what it does and does not measure. Understanding the limitations of a particular instrument is key. Education about the statistical value of evaluation data, possible external influences on evaluation results and effective means of managing and interpreting data, including appropriate comparative measures, would help ensure that when data is

used for summative purposes, decisions are fair and equitable. Please see *Section 4.E.i: Overview of Studied Variables* for a list of issues that should be addressed when training administrators in evaluation data use.

Present data so that it can be easily and accurately interpreted

Many researchers have commented on the need for concise, but useful, information for administrators regarding how best to relay evaluation data to others. Abrami (2001) suggests that the power of presentation should not be underestimated. He notes that visual representations, such as charts or graphs, can positively impact a reviewer's ability to interpret the information. For some, merely reporting averages is not enough: a chart or graph with comparators (e.g. departmental averages) can help to clarify an individual's scores and his/her place in relation to colleagues or in comparison to previous years of teaching (e.g. tracking changes over time). See *Section 4.D.ii: Challenges to Interpretation and Use for Summative Purposes* for more information about comparing evaluation data between courses or instructors. Reports should not present more data, or data in greater detail, than administrators need for their particular evaluation activities. Ory and Ryan (2001) and Theall and Franklin (2001) note that administrators should only be given data to a level of specificity (e.g. decimal places) that matches the level of specificity at which they are able to identify meaningful statistical and conceptual distinctions.

To assist administrators in making effective decisions using ratings data, Hoyt and Pallett (1999) recommend that the available data for all courses taught by a faculty member be presented and that the evaluation be based on a cumulative record of the instructors' teaching effectiveness (with a minimum of six courses).

Include appropriate supplementary evidence with evaluation data

To make valid judgments about teaching effectiveness, Cashin (1988) recommends using multiple sources of data. This might include self-reports or reflective narratives, information on course objectives, sample teaching materials, grading schemes, details on changes made to courses and evidence of scholarship on teaching and of professional development activities (much of which typically forms the basis for a teaching portfolio or dossier) (Hoyt & Pallett, 1999; Seldin, 1993, 1999). These data can triangulate and contextualize student ratings and can address elements of the course or teaching strategies that are not evident in evaluation data.

Some experts disagree about the value of using these other measures of teaching effectiveness for summative evaluations. In a review of the literature on peer evaluation of teaching, Bernstein (2008) notes that while many advocate for the use of collegial, in-class observations of teaching for formative purposes (to improve teaching effectiveness) most scholars caution against using informal observations for summative review because of their relative lack of validity. To improve this form of assessment, DeZure (1999) has recommended multiple observations by more than one trained individual and the use of a valid evaluation form.

While agreeing that additional evaluation measures (e.g. peer evaluations, course materials, etc.) may be used to supplement or complement ratings data, Abrami (2001) cautions that these are "less psychometrically sound" than evaluation

instruments and should not be used instead of formal end-of-term ratings forms (p. 65).

5.C.iv For institutions

Test and review instruments when institutional priorities or teaching practices change

Evaluation instruments should be regularly tested and reviewed by the institutions using them. As teaching methods change, students, faculty and administrators change, and as institutional policies change, ratings forms may need to be revised and updated. Institutions should ensure that the instrument is effectively measuring the specific items that are of interest to them, their faculty and students. Evaluation researchers do not recommend annual overhauls to ratings forms, or even annual minor revisions to the tool. This can negatively affect the ability of the instrument to contribute to longitudinal reviews at the institutional or faculty level. (For example, it may have a negative impact on a faculty member if items on an evaluation form or its scale is altered several times in the years leading up to tenure or promotion. If data is not clearly presented to indicate these changes, reviewers may misinterpret the results.)

Moreover, as higher education evolves, evaluation instruments should be reviewed, and if necessary revised, to address the changing contexts of postsecondary teaching. This may include shifts in pedagogical practices and student demographics or changes in a faculty member's assessment needs (for formative purposes), in institutional accountability measures, in technology, in faculty development practices and in evaluation research (Theall & Franklin, 2000).

Conduct self-studies and internal research

An institution should consider conducting internal research on its evaluation system. This could involve reviewing the instrument or surveying the community about the tool's utility or about their attitudes toward it. Ory (2001) notes that the effect of course, student and instructor influences might vary from institution to institution, and that different institutions or even divisions may find that they need to control for particular variables in order to produce evaluation data that can be accurately compared across courses or instructors. Research on one institution and on one instrument might not necessarily be entirely generalizable and should be validated by institutional research.

Establish policy frameworks for the collection, administration and use of student course evaluation systems

In *Section 3 Current Policy and Practice in North America* we reviewed a range of policies and practices from several dozen postsecondary institutions demonstrating the variations in policies across North America. In general, we noted that most institutions have developed policies regarding the collection, administration and use of student course evaluation systems. However, to our surprise, we uncovered several institutions that regularly use ratings systems but appeared to lack formal policy frameworks addressing these key issues. To ensure consistency, transparency and clarity, such frameworks should be adopted.

CAUT recommends that faculty and their representative associations be involved in the development of these policies. The CAUT "Policy on the Use of Anonymous Student Questionnaires in the Evaluation of Teaching" (2006) states that:

> *Any procedure initiated by the administration or the senior academic body to evaluate teaching performance, including any proposal to employ anonymous student questionnaires, should have the agreement of, or have been negotiated with the academic staff association, and should be incorporated in the collective agreement or faculty handbook* (see Appendix D.2).

Similar sentiments are echoed in the American Association of University Professors (AAUP) "Statement on Teaching Evaluation" (see Appendix D.3).

Establish clear administrative practices

A number of studies, including those that have surveyed users of evaluations, have recommended that institutions improve processes and practices related to the administration of course evaluation systems.

For institutions considering online evaluation systems (see *Section 6.A.i: Online Evaluation Tools* for a discussion of this emerging trend), Sorenson and Reiner (2003) provide a useful list of considerations, including how best to introduce organizational change, anticipate and address objections, assess readiness, educate users, create a convenient and secure system and promote collaboration and ownership.

Articulate evaluation goals and purpose

Noting that their study uncovered some ambiguity regarding the purpose of evaluations, Campbell and Bozeman (2008) recommend that institutions define and clearly articulate their statement of purpose for conducting evaluations and refine their administrative procedures to reflect these goals.

Develop educational materials and support networks for users

Franklin and Theall (1989) have shown that the less an individual knows about course evaluations, the more likely they are to question their usefulness as indicators of teaching effectiveness. They and others have also demonstrated that awareness about student evaluations is low and highly variable.

As noted above, students, faculty and administrators could benefit greatly from education on and training in the use of evaluation ratings systems – a responsibility that should fall to the institution (or a delegated authority).

It is highly recommended that institutions work to improve the education of those using and interpreting evaluation systems (Theall & Franklin, 2000). Moreover, a great deal of the literature calls on institutions to do more than simply provide summary reports of ratings (Theall & Franklin, 2000).

Theall and Franklin (2000) argue that evaluation and faculty development practices are “inextricably connected” as “good evaluation requires the definition of the characteristics and performance to be considered and the commitment of institutions and the individuals within them to use the best possible evidence accurately and fairly to make decisions” (p. 103).

Franklin and Theall (1989) have noted that guides or handbooks on course evaluation systems can be an important source of information for those reviewing, receiving, reporting on, interpreting and making decisions based upon ratings data. Such guides might include the following: a description of effective instruments; recommendations for administrative procedures, including implementation practices and policies; and methods for analysis, reporting and interpretation.

Individual Strategies for Analyzing Student Feedback

1. Control your defence mechanisms.
2. Analyze the source of your students’ reactions in a way that sheds light on any issues and problems that have been identified.
3. Work hard not to under-react or over-react to information that you receive via evaluation feedback.
4. Divide the issues raised by students into actionable and non-actionable categories.
5. Communicate with students before and after their provision of feedback.
6. Do not make the simplistic assumption that all positive responses are related to good teaching and all negative responses are related to bad teaching.
7. Remember that small changes can have big effects.
8. Develop a teaching enhancement strategy that takes into account the evaluation feedback (145-6).

Moore, S., & Kuol, N. (2005). A punitive tool or a valuable resource? Using student evaluations to enhance your teaching. In G. O’Neill, S. Moore, & B. McMulline (Eds)., *Emerging issues in the practice of university learning and teaching* (pp. 141-148). Dublin: All Ireland Society for Higher Education.

Section 6

Emerging Trends, Existing Gaps and Suggestions for Further Research

6.A Emerging Trends

The research and current practices at North American institutions reveal a number of new directions. We have selected what seem to us to be the areas most poised to drive changes in the administration and use of evaluations in the relatively near future.

6.A.i Online evaluation tools

The movement toward offering online or computer-based course evaluations began approximately 20 years ago, with more widespread adoption taking place over the past decade. The research in this area is still emerging and the debate regarding particular implementation-related issues is still burgeoning.

As we noted in *Section 3.D.i: Method of Delivery*, many North American institutions have begun to administer course evaluations online. Some have comprehensively adopted this method, while others offer both online and hard copy options and a few continue to use only the printed questionnaire format. A 2003 study of 256 American institutions revealed that 10% of institutions reported using online tools as their primary means of conducting course evaluations, while 78% indicated they used scannable paper forms. The remaining 12% used non-scannable paper forms (Hoffman, 2003).

The process for administering online evaluations varies, with some institutions providing time during class to complete the survey (e.g. in a computer lab) and others asking students to do so on their own time. Typically, a web address is provided for students through which they can access the evaluation instrument for their particular course or courses. The web link may be made available in-class or sent to students via a learning management system (such as Blackboard or WebCT). A unique and secure log-in code is usually provided or students may be asked to use their own student identification number.

While many attest to the range of advantages in using an online system, others are less convinced. Moreover, there are particular disadvantages that have been identified (e.g. low response rates) that have yet to be overcome. Many scholars have suggested how best to address these problems and institutions themselves have tested a variety of methods; however, none of these are yet widely accepted, nor are any foolproof.

The following is a summary of the key issues discussed in the research:

Cost-effectiveness and efficiency

The online administration of evaluations can be significantly advantageous, particularly in relation to cost, both monetary (printing forms) and in staff time (distributing, collecting, scanning, typing comments and storing data) (Donmeyer et al., 2004; Bothell & Henderson, 2003; Johnson, 2003; Sorenson & Reiner, 2003).

Since they are not typically conducted during class time, online evaluations do not use up time that could be used for teaching purposes (Donmeyer et al., 2004; Sorenson & Reiner, 2003).

Moreover, through online collection, data can be processed more quickly than that from paper forms (Donmeyer et al., 2004; Sorenson & Reiner, 2003) and more extensive, higher quality and customized reports can be produced (Llewellyn, 2003).

Student anonymity

Since student handwriting will not appear in online evaluations, some have argued that they are more capable of ensuring student anonymity (Donmeyer et al., 2004). However, many scholars have noted that anonymity remains a concern for students even when using online evaluation systems (Avery et al., 2006; Reid, 2001). In part, this may relate to a belief that log-in codes can be matched to individual students, a fear that increases when students access surveys with their unique student identification numbers. As noted above in *Section 4.E.i: Overview of Studied Variables*, anonymity is not a factor that significantly impact ratings results; however, it is still generally recommended that institutions do their best to ensure anonymity, and this pertains to the online environment as well. To address this, some universities and colleges have contracted external companies to collect and analyze data and prepare reports.

Instructor variables: faculty influence

The literature has also focused on some of the variables addressed in relation to traditional paper evaluations. Donmeyer et al. (2004) argue that "online evaluations are less susceptible to faculty influence than the in-class evaluations" suggesting that, with paper evaluations, instructors may do something on the day that evaluation forms are administered that could result in higher ratings (p. 612). They further assert that the mere presence of the instructor could impact evaluation scores. However, as we saw in the discussion of variables in *Section 4.E.i: Overview of Studied Variables*, instructor presence does not significantly impact evaluation ratings.

Administering evaluation forms: improving student responses

When administered online and completed outside of class, students are not restricted by the amount of time provided at the end of a class meeting. Many tools allow students to return to the survey to include additional information or edit comments before final submission. Some studies have found that qualitative responses provided in online forms are more extensive (Donmeyer et al., 2004; Sorenson & Reiner, 2003).

Survey flexibility

Donmeyer et al.(2004) and Sorenson and Reiner (2003) have argued that instructors are afforded more flexibility with online evaluations as they are able to customize questionnaires through the addition of items related to their individual course or teaching style. However, this option is also often available to faculty using paper forms.

Response rates

One of the primary concerns addressed in the literature relates to response rate (see Avery et al., 2006 for a review of the related research). Many institutions that have adopted online evaluation systems have witnessed a significant decrease in participation, often decreasing by half (or more) of that obtained with paper forms. Avery et al. (2006) note that some studies have reported response rates as low as 43% (compared to 61-82% for paper forms); their own study revealed similar findings. Ryerson's recent online pilot resulted in a 38% response rate, compared to their normal range of 50-60% (Faculty Course Survey, FAQ). Faculty themselves have raised this issue and it has affected their willingness to adopt online evaluation (Donmeyer et al., 2004). Researchers have suggested that, in part, low response rates reflect a concern for anonymity (Avery et al., 2006; Donmeyer et al., 2004) but also may be impacted by the requirement that students complete evaluations on their own time (Sorenson & Reiner, 2003). Other causes may relate to technical problems (Sorenson & Reiner, 2003).

Incentives have been used at some institutions to encourage student responses. These range from small grade incentives, to early release of final marks, to raffle prizes. The Ryerson Faculty Course Survey FAQ advises faculty against the use of such incentives, noting that bonus marks are a form of coercion, are not appropriate for non-academic work and would require that evaluations not be anonymized.

Donmeyer et al. (2004) found that by offering a small grade incentive, response rates could be increased equaling that obtained through in-class administration. The results from this study revealed that the grade incentive did not significantly bias ratings; however, given that this was a rather small study, it is difficult to generalize these findings. Further research should be conducted to test the impact of incentives of all types.

Instead of using incentives to motivate students, some institutions have relied on repeated messaging efforts to ensure acceptable response rates. Evidence suggests that this is an effective means of improving participation (Donmeyer et al., 2004).

Reliability and validity of results

As Avery et al. (2006) note, "little is known about the comparability of evaluation results obtained through Web-based collection mechanisms with those obtained through traditional paper forms," particularly in relation to the impact on mean course evaluation scores (p. 22). With this in mind, they caution administrators who are considering implementing an online system, since evaluation data is

used for summative purposes. Studies have examined the validity and reliability of online evaluation systems, testing whether or not the means of administration impacts the overall scores. Some findings suggest that ratings are generally consistent and that any variations are statistically insignificant (Avery et al., 2006; Donmeyer et al., 2004). Avery et al. (2006) did find that individual survey items received higher scores on the online forms than on the paper forms. In contrast, Hardy (2003) argues that the research is inconclusive and that online scores "may be lower or higher or the same" (p. 33).

Further research in a number of areas related to the online delivery of course evaluations is still needed. Additional and more extensive studies regarding response rates could prove useful, especially for those institutions considering a move to an online tool. Such research should address online response rates in comparison to those for paper evaluations and should also investigate the various types of incentives being offered, particularly in relation to bias. Anecdotal evidence suggests that faculty and administrators believe that online evaluations are only completed by those students who either "love" or "hate" an instructor. This may prove to be a misconception, but requires further consideration. Avery et al. (2006) have suggested that environmental factors such as the impact of peer influence on responses, or of distractions when completing the survey (at home, or in a public place) should also be studied. In addition, further investigation into the content and structure of the evaluation form is also required to determine whether or not online delivery demands any changes.

For more information about some institutions currently using online evaluation systems, visit: http://onset.byu.edu.

6.A.ii Connecting evaluation data to accountability measures and competency-based learning outcomes

Student and program assessment and evaluation in higher education has, for several years, been moving towards a competency-based assessment model. This mode of evaluation focuses on the measurement of pre-determined outcomes or objectives and the identification of authentic means of assessment whenever possible (that is, assessment measures that ask students to demonstrate the outcome directly, rather than using proxies such as course grades) (the Association of American Colleges & Universities is a strong proponent of this model of student and program evaluations as detailed in their 2007 report, *College Learning for the New Global Century*).
As discussed in *Section 3.H: Relationship of Course Evaluations to Accountability Measures*, course evaluations and other means of assessing student learning and the student experience are beginning to be incorporated into institutional and system-level assessment, accountability and planning processes. Such exercises are still rudimentary, primarily because course evaluations are not necessarily designed to measure specific program outcomes and generally do not offer means of authentic assessment.

For the trend towards incorporating course evaluations in institutional and system-level accountability and evaluation to continue effectively, course evaluation instruments will need to be modified to reflect program outcomes more directly. This parallels the argument proffered by Ory and Ryan (2001) and Theall and Franklin (2001) that

questions on evaluations should reflect institutional priorities. Hoyt and Pallet (1999) argue similarly that when evaluations are to be used primarily for summative purposes, instruments should focus on measuring identified outcomes (e.g. how successfully were the objectives of the course addressed?).

6.A.iii Increasing use of evaluations for formative purposes

As Ory noted in 2000, the evaluation of teaching continues to become more multi-faceted and formative. Marsh's (2007) study demonstrates that simply sharing the results of summative evaluations with instructors does little to improve teaching. Instead, fairly intensive consultation processes are required in order to see substantial and sustained improvements to teaching.

Several scholars (Abrami, 2001; McKeachie, 1997) have, however, also noted that course evaluations are not primarily designed for use in formative evaluation, although some instruments can have some diagnostic utility. By contrast, mid-course evaluations allow for the individualized and often qualitative feedback that can be most beneficial to instructors hoping for information about how to improve their teaching. Mid-course evaluations are therefore becoming an increasingly popular tool (Aultman, 2006), though because of their relative novelty, little research yet exists about best practices in their development and administration or about means to best incorporate their results in teaching improvement and development activities.

6.A.iv Contextualization of evaluation data for summative evaluation of teaching

Over the past 20 or so years, the research demonstrates a growing interest in issues related to improving evaluation practices. On the one hand this has included the introduction and exploration of new assessment tools such as teaching dossiers or portfolios and the use of peer or professional in-class reviews; on the other, it has also involved greater attention to institutional practices and policies.

In the field of faculty development, there has been a movement to address some of the concerns regarding the use and interpretation of course evaluation data for personnel decisions. As part of their regular practice, educational developers work with individual faculty members to "de-code" numerical and qualitative data, both for formative and summative purposes. Through consultation, they also assist faculty in making appropriate changes to their teaching strategies or course design (e.g. grading scheme, assignment design) in response to student comments and ratings. And some, like Franklin (2001) have recommended that faculty develop narratives to supplement and contextualize evaluation data for institutional evaluators, such as tenure and promotion committees. This process aids administrators when reviewing course evaluation data and is beneficial to the individual faculty member. Moreover, she asserts that "narratives can help your reviewers gain a fuller understanding of ratings as a valuable but imperfect measure of teaching effectiveness and therefore help them avoid common misinterpretations and misuses of data that can adversely affect their evaluation of your teaching" (p. 85).

As Franklin (2000) notes, numbers only reveal part of the story; a contextualizing narrative can speak to the multi-dimensionality of teaching and can address specific aspects of the course or the instructor's teaching style that a digest or summary of results

cannot. She recommends that faculty review research on the variables (see *Section 4.E.i: Overview of Studied Variables*) that may impact ratings to varying degrees and address these, as appropriate, in their narrative. For example, some evidence indicates elective courses receive higher ratings; therefore, instructors whose full teaching complement consists of required courses may wish to make note of this for evaluators.

6.B Existing Gaps and Suggestions for Further Research

6.B.i Defining teaching vocabulary and expectations

Theall and Franklin (2000) highlight the need to develop a "reliable and extensive common vocabulary to describe important postsecondary phenomena" (p. 104), including vocabulary for the evaluation of teaching, as well as the changing nature of postsecondary teaching (e.g. new instructional practices). They assert that this is "essential to any valid generalizing of ratings findings" (p. 104). Without a universal understanding of the essential terms used to discuss teaching and student learning our inability to reach a consensus about what constitutes effective teaching will persist.

As Wright (2008) suggests, the vocabulary used to discuss course evaluations must accurately express the ways in which the instruments are used. Although this paper reflects current practice, Wright suggests that current practice should change to ensure that students, faculty and institutions understand their respective roles in the process of evaluating teaching: students rate instruction, while administrators, institutions and faculty participate in evaluation of teaching, based in part on the results of student ratings. In particular, if evaluations strictly ask students to rate instruction as opposed to course content the name selected for the instrument should reflect this focus and should recognize that the title by which the instrument is referred can influence how it is used. Current practice on this issue is not consistent and future research may usefully be directed at identifying a common and meaningful terminology for use across post-secondary sectors in Canada.

6.B.ii Understanding evaluation users

Few studies on evaluation users – administrators, faculty and students – have been conducted. Those that exist are typically small and institution-based. Given differences in institutional mandate, disciplinary focus and culture, it is difficult to map these findings onto the broader higher education sector. These factors may impact individual perspectives and attitudes and their use of evaluation data. Additional research on a larger scale would certainly contribute to our understanding of evaluation users; however, this should not exclude further institutional studies which will continue to inform university and college administrators, faculty and students within the context of their own institutional culture. Moreover, such studies can also consider the specific evaluation tools used within the institution. Further research on evaluation users should be conducted in both contexts.

Theall and Franklin (2000) call for further research on the needs of the various users of course evaluation data, noting that the task of interpreting results varies by purpose. As

they suggest, interpretation goes beyond simply being able to unpack the numbers, but frequently involves an ability to provide further consultation on how to translate ratings into actual improvement in teaching methods. Similarly, Menges (2000) recommends further research into how administrators use student evaluation data for personnel decisions. McKeachie (1997) has suggested that focused observational research on the decision-making process would be helpful to enable a more thorough understanding of how ratings are used. For this he suggests that researchers attend tenure and promotion committee meetings to observe how ratings data is actually used by reviewers.

As Schmelkin, Spencer and Gellman (1997) have noted, we also require a better understanding of how faculty perceive and use evaluation data. To date, the findings have been mixed. Schmelkin and colleagues' study of 400 faculty found that, contrary to anecdotal evidence, faculty were not overly resistant to the use of course evaluation data for either formative or summative purposes, whereas opinion pieces and discussions on web sites and listservs suggests otherwise.

Of all those involved in the gathering and use of evaluation ratings, the least studied remains the student. McKeachie (1997) has advocated for attention to be paid to the way in which students understand the evaluation process and the manner in which they complete ratings forms. From this, he argues, we can move toward better educating students to become more sophisticated evaluators.

6.B.iii Educating evaluation users

There is a general and oft-repeated call to better educate all those who use course evaluation forms and related data. For students, this means ensuring that they understand the purpose and subsequent use of evaluations. For faculty, this means addressing the persistent myths that jeopardize the esteem and consequent utility of evaluations for teaching development. It also means helping them to identify useful contextualizing data that can be employed in their own interpretation of results and provided to administrators to aid in summative evaluation. For administrators, this means developing and providing training to be better evaluators which includes information about reporting and interpreting statistical data. And for institutions, this means ensuring that comprehensive policies are developed and implemented equitably.

To achieve these goals, further research is needed. To improve training and educational materials, we must have a better understanding of the users, including their attitudes and perceptions towards evaluation tools. Moreover, we also require more information regarding the knowledge users possess about course evaluation research and institutional processes and practice. And of course, a more thorough understanding of how ratings data are used for formative or summative purposes is also necessary. Research in these areas has been limited. Further inquiry will be of great benefit to the current scholarship, but more importantly it will enable future development of grounded and more effective recommendations regarding the creation of instruments, the administration and implementation of evaluation systems and the use and interpretation of ratings data.

6.B.iv Evaluating graduate student teaching assistants and instructors

There is a growing trend in postsecondary education to provide professional development opportunities and training for graduate students. Numerous professional development resources have been developed for graduate teaching assistants, and offices dedicated to providing training and support have proliferated across university campuses (see, for example, the Teaching Assistants' Training Program at the UofT or the Preparing Future Faculty program in the United States). Given that graduate student teaching assistants (TAs) have fairly extensive and direct contact with undergraduate students (through tutorials, labs, office hours and grading responsibilities) the quality of their teaching should be evaluated. While some institutions have developed mechanisms to evaluate graduate TAs, many institutions do not engage in this practice at all. In the few that do, procedures vary across departments and divisions. Moreover, student ratings can assist with professional development and the academic job search. An increasing number of institutions are requiring teaching dossiers (or portfolios) from job candidates; evidence of teaching experience and expertise forms an essential part of this document.

At present, there is limited research relating to the evaluation of graduate students as TAs. One area that requires particular consideration is the type of ratings forms and the scope of questions to be asked. While existing evaluation forms may provide opportunities to survey students on their experience with TAs, questions should be tailored to specifically address the range of activities and teaching behaviours particular to TAs. Instruments geared toward faculty, which frequently ask questions about the structure of the course, selected readings and assignments and tests are generally not appropriate for the evaluation of TAs, as they would not normally have any involvement with these aspects of a course.
Please also see Appendix G for a summary of recommendations for future research drawn from Greenwald (1997).

Section 7

Concluding Remarks

In spite of the fact that there are now thousands of articles devoted to the topic of student course evaluations, there is still much research to be done. Within this vast body of scholarship particular attention has been paid to issues related to validity and reliability: in fact, a significant majority of the studies and literature reviews have focused on these areas and continue to do so. However, as we have demonstrated in this review, the reliability and internal validity of course evaluations are now widely accepted by numerous scholars as evidenced by scores of grounded empirical evidence. It is perhaps now time to turn our attention toward some of the other issues that have received repeated calls for further consideration. These include:

- Improving information for and education of evaluation users and tested results;
- Developing and testing effective means of reporting results and tools for interpretation (in relation to user needs);
- Ensuring faculty and student commitment to the evaluation process; and,
- Regular review of evaluation instruments based on institutional needs and goals and in relation to current research findings.

While researchers must refocus their scholarship about course evaluation validity away from the investigation of individual survey items and towards these broader issues of survey design, implementation and interpretation, institutions must also adapt their view and use of evaluations. Evaluations must be designed to carefully match their institutional context and be accompanied by substantial institutional support. Policies must be comprehensive and equitable. Education for evaluation users – students, faculty and administrators – must dispel myths and misperceptions and improve skill and transparency. Evaluations must be accompanied by ongoing dialogue and support mechanisms, including consultation, to ensure that they contribute to the support and improvement of teaching within the institution.

Our research has clearly identified that evaluations are valuable and important tools for the assessment of teaching – but only if they are developed and supported with the understanding that validity is determined by much more than simply the ways students respond to individual items on a survey.

Works Cited

Abrami, P.C. (2001). Improving judgments about teaching effectiveness using teacher rating forms. In M. Theall, P.C. Abrami, and L.A. Mets (Eds.). The student ratings debate: Are they valid? How can we best use them? [Special issue]. *New Directions for Institutional Research 109*, 59-87.

Abrami, P.C., d'Apollonia, S., & Cohen, P.A. (1990). Validity of student ratings of instruction: What we know and what we do not know. *Journal of Educational Psychology*, *82*(2), 219-231.

Addison, W.E., Best, J., & Warrington, J.D. (2006). Students' perceptions of course difficulty and their ratings of the instructor. *College Student Journal*, *40*(2), 409-416.

Aleamoni, L.M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, *13*(2), 153-166.

Aleamoni, L.M. (1997). Issues in linking instructional improvement research to faculty development in higher education, *Journal of Personnel Evaluation in Education*, *11*(1), 31-37.

Aleamoni, L. M. (1987). Typical faculty concerns about student evaluation of teaching. *New Directions for Teaching and Learning*, *31*, 25-31.

Aleamoni, L. M. & Hexner, P. Z. (1980). A review of the research on student evaluation and a report on the effect of different sets of instructions on student course and instructor evaluation. *Instructional Science*, *9*(1), 67-84.

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

Ali, D.L & Sell Y. (1998) *Issues regarding the reliability, validity and utility of student ratings of instruction: A survey of research findings*. Calgary: University of Calgary APC Implementation Task Force on Student Ratings of Instruction.

Arreola, R.A. (2000). *Developing a comprehensive faculty evaluation system* (2nd ed.). Bolton, MA: Anker.

Arreola, R.A. (1995). *Developing a comprehensive faculty evaluation system*. Bolton, MA: Anker.

Arreola, R. A. (1983). Establishing successful faculty evaluation and development programs. In A. Smith (Ed.). Evaluating faculty and staff [Special issue]. *New Directions for Community Colleges*, *41*, 83-93.

Aultman, L.P. (2006). An unexpected benefit of formative student evaluations. *College Teaching*, *54*(3), 251.

Avery, R.J., Bryan, W.K., Mathios, A., Kang, H., & Bell, D. (2006). Electronic course evaluations: Does an online delivery system influence student evaluations? *Journal of Economic Education*, *37*(1), 21-37.

Baldwin, T. & Blattner, N. (2003). Guarding against potential bias in student evaluations: What every faculty member needs to know. *College Teaching*, *51*(1), 27-32.

Beran, T., Violato, C., & Kline, D. (2007). What's the 'use' of student ratings of instruction for administrators? One university's experience. *Canadian Journal of Higher Education*, *17*(1), 27-43.

Beran, T., Violato, C., Kline, D., & Frideres, J. (2005). The utility of student ratings of instruction for students, faculty, and administrators: A "consequential validity" study. *Canadian Journal of Higher Education*, *35*(2), 49-70.

Bernstein, D.J. (2008). Peer review and evaluation of the intellectual work of teaching, *Change*, *40*(2), 48-51.

Bothell, T.W., & Henderson, T. (2003). Do online ratings of instruction make $ense? In D.L. Sorenson & T.D. Johnson (Eds.), Online student ratings of instruction [Special issue]. *New Directions for Teaching and Learning, 96*, 69-80.

Braskamp, L.A. & Ory, J.C. (1994). *Assessing faculty work: Enhancing individual and institutional performance*. San Francisco: Jossey-Bass.

Calderon, T.G., Gabbin, A.L., & Green, B.P. (1996). *Report of the committee on promoting evaluating effective teaching*. Harrisonburg, VA: James Madison University.

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

Cashin, W.E. (1995). *Student ratings of teaching: The research revisited* (IDEA Paper No. 32). Manhattan, KS: Kansas State University Center for Faculty Evaluation and Development.

Cashin, W.E. (1992). Student ratings: The need for comparative data. *Instructional Evaluation and Faculty Development*, *12*(2), 1-6.

Cashin, W.E. (1990). Students do rate different academic fields differently. In Theall, M. & Franklin, J. (Eds.), Student ratings of instruction: Issues for improving practice [Special issue]. *New Directions for Teaching and Learning*, *43*, 113-121.

Cashin, W.E. (1989). *Defining and evaluating college teaching* (IDEA Paper No. 21). Manhattan, KS: Kansas State University Center for Faculty Evaluation and Development.

Cashin, W.E. (1988). *Student ratings of teaching: A summary of the research* (IDEA Paper No. 20). Manhattan, KS: Kansas State University Center for Faculty Evaluation and Development.

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

Centra, J.A. & Gaubatz, N.B. (1998, April). *Is there a gender bias in student evaluation of teaching?* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

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

Centra, J.A. (1979). *Determining faculty effectiveness: Assessing teaching, research, and service for personnel decisions and improvement.* San Francisco: Jossey-Bass.

Centra, J.A., & Creech, R.F. (1976). T*he relationship between student, teacher and course characteristics and student ratings of teacher effectiveness*. Princeton, N.J.: Educational Testing Service.

Chambers, B.A., & Schmitt, N. (2002). Inequity in the performance evaluation process: How you rate me affects how I rate you. *Journal of Personnel Evaluation in Education*, *16*(2), 103-112.

Cohen, P.A., & McKeachie, W.J. (1980). The role of colleagues in the evaluation of teaching. *Improving College and University Teaching*, *28*(4), 147-154.

Coren, S. (2001). Are course evaluations a threat to academic freedom? In S.E. Kahn & D. Pavlich (Eds.), *Academic Freedom and the Inclusive University* (pp. 104-117). Vancouver: University of British Columbia Press.

Crosson, A.C., Boston, M., Levison, A., Matsumura, L.C., Matsumura, L.C., Resnick, L.B., et al. (2006). *Beyond summative evaluation: The instructional quality assessment as a professional development tool* (CSE Technical Report 691). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

d'Apollonia, S., & Abrami, P.C. (1997). Navigating student ratings of instruction. *American Psychologist*, *52*(11), 1198-1208.

Davidovitch, N., & Soen, D. (2006). Class attendance and students' evaluation of their college instructors. *College Student Journal*, *40*(3), 691-703.

DeZure, D. (1999). Evaluating reaching through peer classroom observation. In P. Seldin (Ed.), *Changing practices in evaluating teaching. A practical guide to improved faculty performance and promotion/tenure decisions* (pp. 70-96). Bolton, MA: Anker.

Diamond, M. R. (2004). The usefulness of structured mid-term feedback as a catalyst for change in higher education classes. *Active Learning in Higher Education 5*(3), 217-231.

Donmeyer, C.J., Baum, P., Hanna, R.W. & Chapman, K.S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: Their effects on response rates and evaluations. *Assessment & Evaluation in Higher Education*, *29*(5), 611-623.

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

Felder, R. (1993). What do they know anyway? 2: Making evaluations effective. *Chemical Engineering Education, 27*(1), 28-29.

Feldman, K.A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education*, *30*(2), 137-194.

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

Fox, C.R. (2006). The availability heuristic in the classroom: How soliciting more criticism can boost your course ratings. *Judgment and Decision Making*, *1*(1), 86-90.

Franklin, J. (2001). Interpreting the numbers: Using a narrative to help others read student evaluations of your teaching accurately. In K.G. Lewis (Ed.), Techniques and strategies for interpreting student evaluations [Special issue]. *New Directions for Teaching and Learning*, *87*, 85-100.

Franklin, J., & Theall, M. (1990). Communicating student ratings to decision makers: Design for good practice. In M. Theall & J. Franklin (Eds.), Student ratings of instruction: Issues for improving practice [Special issue]. *New Directions for Teaching and Learning*, *43*, 75-93.

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

Goldschmid, M.L. (1978). The evaluation and improvement of teaching in higher education. *Higher Education*, *7*(2), 221-245.

Green, B.P., Calderon, T.G., & Reider, B.P. (1998). A content analysis of teaching evaluation instruments used in accounting departments. *Issues in Accounting Education*, *13*(1), 15-30.

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.

Gump, S.E. (2007). Student evaluations of teaching effectiveness and the leniency hypothesis: A literature review. *Educational Research Quarterly*, *30*(3), 55-68

Hardy, N. (2003). Online ratings: Fact and fiction. In D.L. Sorenson & T.D. Johnson (Eds.), Online student ratings of instruction [Special issue]. *New Directions for Teaching and Learning*, *96*, 31-38.

Harper, S.R., & Kuh, G. (2007). Myths and misconceptions about using qualitative methods in assessment. In S.R. Harper & S.D. Museus (Eds.), Using qualitative methods in institutional assessment [Special issue]. *New Directions for Institutional Research*, *136*, 5-14.

Haskell, R.E. (1997). Academic freedom, tenure, and the student evaluation of faculty: Galloping polls in the 21st century. *Education Policy Analysis*, *5*(6).

Heckert, T.M., Latier, A., Ringwald-Burton, A., & Drazen, C. (2006). Relations among student effort, perceived class difficulty appropriateness, and student evaluations of teaching: Is it possible to "buy" better evaluations through lenient grading? *College Student Journal*, *40*(3), 588-596.

Hodges, L.C., & Stanton, K. (2007). Translating comments on student evaluations into the language of learning. *Innovative Higher Education*, *31*, 279-286

Hoffman, K.M. (2003). Online course evaluations and reporting in higher education. In D.L. Sorenson & T.D. Johnson (Eds.), Online student ratings of instruction [Special issue]. *New Directions for Teaching and Learning*, *96*, 25-30.

Hoyt, D.P., & Pallett, W.H. (1999). *Appraising teaching effectiveness: Beyond student ratings* (IDEA Paper No.36). Manhattan, KS: Kansas State University Center for Faculty Evaluation and Development.

Johnson, T.D. (2003). Online student ratings: Will students respond? In D.L. Sorenson & T.D. Johnson (Eds.), Online student ratings of instruction [Special issue]. *New Directions for Teaching and Learning*, *96*, 49-60.

Keig, L., & Waggoner, M.D. (1994). Collaborative peer review: The role of faculty in improving college teaching. *ASHE-ERIC Higher Education Report No. 2*. Washington, DC: George Washington University.

Kulik, J.A. (2001). Student ratings: Validity, utility, and controversy. In M. Theall, P.C. Abrami, & L.A. Mets (Eds.), The student ratings debate: Are they valid? How can we best use them? [Special issue]. *New Directions for Institutional Research*, *109*, 9-25.

Lang, J.W.B., & Kersting, M. (2007). Regular feedback from student ratings of instruction: Do college teachers improve their ratings in the long run? *Instructional Science*, *35*, 187-205.

Lattuca, L.R., & Domagal-Goldman, J.M. (2007). Using qualitative methods to assess teaching effectiveness. In S.R. Harper & S.D. Museus (Eds.), Using qualitative methods in institutional assessment [Special issue]. *New Directions for Institutional Research*, *136*, 81-93.

Lewis, K.G. (2001). Using midsemester student feedback and responding to it. In K.G. Lewis (Ed.), Techniques and strategies for interpreting student evaluations [Special issue]. *New Directions for Teaching and Learning*, *87*, 33-44.

Llewellyn, D.C. (2003). Online reporting of results of online student ratings. In D.L. Sorenson & T.D. Johnson (Eds.), Online student ratings of instruction [Special issue]. *New Directions for Teaching and Learning*, *96*, 61-68.

Marsh, H.W. (2007). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology*, *99*(4), 775-790.

Marsh, H.W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, *11*(3), 253-388.

Marsh, H.W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, *76*(5), 707-754.

Marsh, H.W., & Dunkin, M.J. (1992). Students' evaluations of university teaching: A multidimensional approach. In J.C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 8) (pp. 143-223). New York: Agathon Press.

Marsh, H.W., & Roche, L (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology*, *92*(1), 202-228.

Marsh, H.W., & Roche, L. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias and utility. *American Psychologist*, *52*(11), 1187-1197.

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist*, *52*(11), 1218-1225.

McKeachie, W.J., & Lin, W.G. (1975). Multiple discriminant analysis of student ratings of college teachers. *Journal of Educational Research*, *68*(8), 300-305.

Menges, R.J. (2000). Shortcomings of research on evaluating and improving teaching in higher education. In K.E. Ryan (Ed.), Evaluating teaching in higher education: A vision for the future [Special issue]. *New Directions for Teaching and Learning*, *83*, 5-11

Moore, S., & Kuol, N. (2005). A punitive tool or a valuable resource? Using student evaluations to enhance your teaching. In G. O'Neill, S. Moore, & B. McMulline (Eds)., *Emerging issues in the practice of university learning and teaching* (pp. 141-148). Dublin: All Ireland Society for Higher Education.

Murray, H.G. (1987). Acquiring student feedback that improves instruction. In Weimer, M.G. (Ed.), Teaching large classes well [Special issue]. *New Directions for Teaching and Learning*, *32*, 85-96.

Naftulin, D.H., Ware, J.E., & Donnelly, F.A. (1973), The Dr. Fox lecture: A paradigm of educational seduction. *Journal of Medical Education*, *48*, 630-635.

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

Neumann, R. (2001). Disciplinary differences and university teaching. *Studies in Higher Education*, *26*(2), 135-146.

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

Ory, J.C. (2001). Faculty thoughts and concerns about student ratings. In K.G. Lewis (Ed.), Techniques and strategies for interpreting student evaluations [Special issue]. *New Directions for Teaching and Learning*, *87*, 3-15.

Ory, J.C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? In M. Theall, P.C Abrami, & L.A. Mets (Eds.), The student ratings debate: Are they valid? How can we best use them? [Special issue]. *New Directions for Institutional Research*, *109*, 27-44.

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

Reid, I.C. (2001). Reflections on using the Internet for the evaluation of course delivery. *The Internet and Higher Education*, *4*(1), 61-75.

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.

Rich, H.E (1976). Attitudes of college and university faculty toward the use of student evaluations. *Educational Research Quarterly*, *1*, 17-28

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

Schmelkin, L.P., Spencer, K.J. & Gellman, E.S. (1997). Faculty perspectives on course and teacher evaluations. *Research in Higher Education*, *38*(5), 575-592.

Scriven, M. (1981). *Evaluation thesaurus* (3rd ed.). Pt. Reyes, CA: Edgepress.

Scriven, M. (1995). Student ratings offer useful input to teacher evaluations. *Practical Assessment, Research & Evaluation*, *4*(7).

Sedlmeier, P. (2006). The role of scales in student ratings. *Learning and Instruction*, *16*(5), 401-415.

Seldin, P. (1999). *Changing practices in evaluating reaching. A practical guide to improved faculty performance for promotion/tenure decisions*. Bolton, MA: Anker.

Seldin, P. (1993). *Successful use of teaching portfolios.* Bolton, MA: Anker.

Sorenson, D.L. & Reiner, C. (2003). Charting the uncharted seas of online student ratings of instruction. In D.L. Sorenson & T.D. Johnson (Eds.), Online student ratings of instruction [Special issue]. *New Directions for Teaching and Learning*, *96*, 1-24.

Sproule, R. (2000). Student evaluation of teaching: A methodological critique of evaluation practices. *Education Policy Analysis*, *8*(50).

Theall, M. (2002). *Student ratings: Myths vs. research evidence*. Retrieved June 26, 2008, from https://studentratings.byu.edu/info/faculty/myths.asp.

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? In M. Theall, P.C Abrami, & L.A. Mets (Eds.), The student ratings debate: Are they valid? How can we best use them? [Special issue]. *New Directions for Institutional Research, 109*, 45-56.

Theall, M., & Franklin, J. (2000). Creating responsive student ratings systems to improve evaluation practice. In K.E. Ryan (Ed.), Evaluating teaching in higher education: A vision for the future [Special issue]. *New Directions for Teaching and Learning*, *83*, 95-107.

Theall, M. (1994). What's wrong with faculty evaluation: A debate on the state of practice. *Instructional Evaluation and Faculty Development*, *14*, 27-34

Wagenaar, T.C. (1995). Student evaluation of teaching: Some cautions and suggestions. *Teaching Sociology*, *23*(1), 64-68.

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

Williams, W.M. & Ceci, S.J. (1997). How'm I doing? Problems with student ratings of instructors and courses. *Change*, *29*(5), 13-23.

Williams, R and Ory, J.C. (1992). *A further look at class size, discipline differences and student ratings*. Urbana-Champaign: Office of Instructional Resources, University of Illinois at Urbana-Champaign.

Wright, R.E. (2006). Student evaluations of faculty: Concerns raised in the literature, and possible solutions. *College Student Journal*, *40*(2), 417-422.

Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education*, *12*(1), 55-76.