Instructional Development Committee of Council Reports

Framework for Student Evaluation of Teaching at the University of Saskatchewan

Approved by University Council January 29, 2004

This framework document should be read in conjunction with the University Council's document entitled *Principles of Evaluation of Teaching at the University of Saskatchewan* approved in March 2002 and the document entitled *Framework for Peer Evaluation of Teaching at the University of Saskatchewan: Best Practices* approved in June 2003. Outcomes from student evaluations are one component of a larger evaluation process used for personnel and career decisions. It is important to consider multiple indicators of teaching effectiveness for major career decisions. Student-based evaluations of teaching are important but should be considered in the context of other performance indicators. This document addresses student evaluations of teaching as part of an overall evaluation process to help improve teaching, assess teaching effectiveness and establish good evaluation practices. The Framework is intended to be a guide for departments and college to assess their evaluation practices. Units are encouraged to continue using practices that are consistent with the *Principles* and *Framework* documents and to incorporate good practices in their evaluation processes over time.

Philosophy

One of the goals of the University, as set out in *A Framework for Planning at the University of Saskatchewan,* is to improve the quality of instructional programs. The *Framework* document states that the University must be governed by considerations of quality and accountability. "A university that is quality conscious will be accountable to its students, its alumni and the people of the Province" (p. 5). Strengthening the teaching evaluation processes over time will demonstrate the University's concern for quality instruction. By making the evaluation of teaching a regular process of our teaching activities, the University will be more accountable to students and teachers alike. As the University strives for excellence based on international standards, it is important to gather information about our outstanding contributions to teaching.

The Principles of Evaluation of Teaching at the University of Saskatchewan states, "the evaluation of teaching at the University of Saskatchewan may serve several functions. Most importantly, teaching evaluations are to be used to assist faculty with the development and improvement of instruction. Data collected from teaching evaluations can also serve a summative function to assist with collegial and administrative decisions" In the *Principles* document, University Council approved a recommendation that all teachers be evaluated at least once every three years and that teachers under consideration for renewal of probation, tenure or promotion should have their teaching evaluated on an annual basis. Thus, student evaluation of teaching is mandatory.

The University of Saskatchewan Standards for Promotion and Tenure declares that "good teaching is expected of all faculty and evaluation of teaching . . . requires more than classroom

performance. Candidates will be expected to demonstrate mastery of their subject area(s) or discipline(s), to make thorough preparation for their classes, to communicate effectively with their students, to show a willingness to respond to students' questions and concerns, and to exhibit fairness in evaluating students . . . faculty are expected to remain committed to improving/enhancing their teaching performance and to remedy problems identified with their teaching. As faculty progress through the ranks, they will be expected to extend their knowledge of their field(s) or discipline(s), i.e. with respect to classes, currency of the material presented, and new teaching methods".

The University of Saskatchewan appreciates and recognizes the commitment of sessional lecturers to good teaching. Student evaluation of teaching should be one of the criteria collected for consideration of right of first refusal for sessional lecturers. Evaluations by students must be consistent with the procedures set out in Articles 14 - Right of First Refusal and 18 - Formal Teaching Assessment of the Collective Agreement between the University of Saskatchewan and CUPE 3287. Unless there are demonstrable reasons for developing and using an alternative evaluation procedure for sessional lecturers, instruments for student evaluation of teaching should be applied consistently to the teaching of all teachers within an academic unit.

The University of Saskatchewan Students' Union (USSU) and the Graduate Students' Association (GSA) recognize the importance of high quality instruction. To this end, both organizations support endeavors to improve instruction and provide better evaluation systems for teachers. The student organizations believe that teachers should be accountable for their teaching and want to ensure that the student voice is heard in the evaluation process. In November of 2002, the University Students Council (USC) adopted and endorsed a statement on evaluations adopted by the Canadian Academic Round Table (CART). The basic principle of the statement is that evaluations should be mandatory and accessible. The discussions held at CART and USC focused on the student belief that evaluations are an important part of the academic world. As students at the national and local level are experiencing rising tuition fees, they want to know that the evaluations they complete will improve the classroom experience. The students are encouraged by the high quality of instruction at the University of Saskatchewan, as evidenced by the nominations and annual honours for the University of Saskatchewan Students' Union Teaching Excellence Awards. Student concerns, however, must be taken seriously and their views should make a difference. Although progress is continuing on student evaluation of teaching , it is the goal of the USSU to advocate for a teaching evaluation process that is mandatory for teachers and where results are accessible to all students. Accessibility means sharing information with students about teaching evaluations completed by students. The *Principles of Evaluation of Teaching at the University of Saskatchewan* states that "summative information from evaluations by students should be made available, on a restricted basis, to all admitted and registered students at the University of Saskatchewan (2002, p. 6). Certainly there is an abundance of information already collected through student evaluations of teaching. It may not be possible to share outcomes of evaluation processes electronically at this point but departments and colleges are encouraged to make summative results from student-based evaluations available to students in some form. Ultimately, students want electronic access to aggregated, summative data from student evaluations of teaching at the University of Saskatchewan.

Thus, commitment to high quality instruction and enhancement of instruction is the responsibility of all engaged in instructional activities. Teachers should strive to achieve excellence in teaching and incorporate best practices into their teaching to improve student learning.

Approved Course Evaluations as distinct from Evaluations of Teaching

Approved course evaluations are those that are approved for the use by the faculty of a department (or college in the case of non-departmentalized college) in committee within guidelines established by the college, and are properly validated instruments of evaluation. For many units on campus, course evaluations have been developed to gather information about student satisfaction with particular courses and to gather information about how, and how much, students are learning within a particular academic program. Course evaluations can also provide information about course delivery and modes of instruction. Information collected from course evaluations may be useful to teachers if the data provides information about teaching effectiveness. Departments and colleges can use course evaluations to determine the efficacy of course structures, new pedagogical approaches and academic programs. It is recognized that students may provide feedback on teaching through other processes such as course evaluations and systematic program evaluations. Teaching evaluations are different, however. Teaching evaluations measure teaching effectiveness. It is important to identify a number of dimensions of teaching for the evaluation process. See page ten of this *Framework* for further details. It is important for a faculty to discuss the issue of asking global questions (i.e., one question for each dimension) as compared to asking a number of questions about a particular factor or dimension. For example, "[s]tudies comparing student achievement test scores with teaching evaluations found significant correlations between the amount of learning and quality of teaching in factors such as 'teacher preparation and organization' and 'clarity and understandableness'" (Lawall, 1998, p. 7). For academic units that do not have approved teaching evaluations, sample evaluations are provided for consideration in appendices A, B and C. The issue of validity is addressed in appendices D and E.

Formative and Summative Evaluation by Students

Evaluation of teaching can be both formative and summative. Evaluations are expected to provide feedback to improve teaching (formative) and to provide information for use in administrative or collegial decisions (summative). A summative evaluation is used for decision-making but may also contain formative components to assist the teacher with the improvement of instruction. Teaching evaluations can simultaneously help teachers improve their teaching while also providing information used in decisions such as renewal of probation, tenure, promotion and right of first refusal. "Faculty must be continually engaged in discussing teaching in order both to nurture new teachers into the community of teacher-scholars and to render the process of making personnel decisions (who gets hired, who gets tenured, who gets merit pay, and the like) more open and more informed by reasoned decisions that consider teaching seriously. The idea is then in the spirit of both continuous quality improvement and the practice of self-regulation within professions" (Van Note Chism, 1999, p. 6).

a) Formative Evaluation

Information gathered from the evaluation of teaching may be used for formative purposes to assist with instructional development and improvement. Formative evaluation "describes activities that are to provide teachers with information that they can use to improve their teaching. The information is intended for their personal uses rather than for public inspection . . . The information should be rich in detail so that teachers can obtain clear insights on the nature of their teaching strengths and weaknesses" (Van Note Chism, 1999, p. 3.).

b) Summative Evaluation

For summative purposes, evaluation of teaching is associated with collegial decision-making processes including tenure, promotion and salary review, right of first refusal and for review of academic programs. "Summative evaluation of teaching focuses on information needed to make a personnel decision . . . Consequently, the information is for public inspection . . . it is often more general and comparative in nature than data for formative evaluation" (Van Note Chism, 1999, p. 3).

Evaluation of teaching performance should be based on a series of evaluations of a candidate's teaching performance and teaching materials over a period of time and based on multiple sources of data (i.e., evaluations by students, peer evaluation and self-evaluation, etc.). Career-related decisions should not be based solely on teaching evaluations by students. However, the University and the teacher recognize their respective accountabilities to students by including students in the evaluation process. Comparative information can be used to demonstrate that improvements have occurred over time or that information gathered from students has been used to enhance learning techniques and approaches in the classroom. To dismiss the perception that student evaluations are based on the popularity of the teacher, it is important to develop multidimensional questionnaires that measure a number of dimensions. "While certain personality types may be more naturally adapted to the conditions of certain classes, effective teaching in any situation is composed of learnable skills and behaviours" (Lawall, 1998, p. 6). The measurement of these factors or dimensions is essential to evaluate teaching effectiveness and to gather information about how to improve teaching. Serial evaluations using different sources of information should mitigate anomalies. Feedback gathered from a series of evaluations, from both students and peers, should provide information that will permit the teacher to improve over time.

In any proper evaluation process, it is important to ensure that feedback is provided to participants, including students. Department and colleges are encouraged to make summative information from student evaluation of teaching available to students through department or college offices, in a department or college resource room, via electronic means or through a local branch of the University Libraries. It is also recommended that teachers who gather formative feedback advise those students how their views and comments will be taken into account in a particular course or what the effect will be on future instruction.

c) Summative and Formative Evaluations as One Procedure

Teaching evaluations can simultaneously help teachers improve their teaching while also providing information used as part of decisions such as probation renewal, tenure and promotion. Formative and summative teaching evaluations do not have to be mutually exclusive activities. Well-designed, student-based teaching evaluation questionnaires can serve both functions. This point will be demonstrated later in a discussion of how teaching evaluation procedures could be established by an academic unit.

Non-Identifying and Confidential Evaluations

It is important to understand the difference between non-identifying student evaluations and confidential materials submitted by students. Consistent with the *Principles* document, "confidential evaluations are signed and may include qualitative comments from students." Where there is the opportunity for students to provide written comments, the comments must be aggregated and summarized to protect the anonymity of students. For formative purposes, the summarized comments should be provided to the teacher. For summative purposes, the summarized qualitative data should be analyzed and interpreted. Additional confidential material (e.g., testimonials and other letters signed by students) may be gathered for summative purposes. Anonymous (non-identifying) evaluations are unacceptable unless the instruments are approved teaching evaluations. Information from such student evaluations must be aggregated or summarized and exclude anonymous remarks by students. For colleges that used approved non-identifying (unsigned) evaluations, care must be taken to ensure that only students registered in the course respond to the evaluation and the data must be aggregated and summarized before use; anonymous remarks made by students must be excluded for any career-related decisions. Departments and colleges can choose to adopt confidential (signed) or anonymous (non-identifying) evaluation instruments in so far as the instruments protect the anonymity of students and provide valid and reliable conclusions on teaching effectiveness.

Students must be advised that their feedback will be used for career decisions involving the teachers and that teachers will only have access to the aggregated and summarized data. Students must be advised that committees involved with career decisions will have access to the aggregated and summarized data and, on request, may have access to the raw data.

While "signed" evaluations may become obsolete in the future due to changing technologies, nonetheless, new evaluation systems should still adhere to the principles set out in the *Principles of Evaluation of Teaching*, and be compatible with relevant sections of the appropriate collective agreements as well as with *The University of Saskatchewan Standards for Promotion and Tenure*.

Frequency

It is important that evaluations be conducted serially so as to provide a reasonable sampling of evidence over a time period. Departments and colleges should determine the frequency of evaluations for those who have achieved tenure, permanent or continuing status and those who have been promoted to the highest rank at the University. It is recommended that student

evaluations be conducted annually for teachers in probationary appointments and every three years after achieving tenure/promotion to full professor; sessionals may opt to have regular student evaluations following the achievement of the right of first refusal. The recommendation regarding frequency should be viewed as a minimum. If a teacher requests additional teaching evaluations, every effort should be made to accommodate such a request. Colleges are encouraged to administer evaluation of teaching by students for on campus and off-campus courses.

The *Framework* appreciates that student fatigue may become a factor when administering lengthy questionnaires towards the latter part of the academic term. This is especially true for team-taught courses or courses with multiple teachers associated with clinical rotations. Students should be advised that their opinions are important and data collected will be used for career decisions for teachers. Longer questionnaires may provide more information that can be of benefit to teachers and to committees charged with career-related decisions.

Raw data gathered from student evaluations should be stored for two or three years to ensure that the data can be reviewed if concerns arise concerning the analysis or accuracy of the information. Summarized and aggregated data should be kept in the personnel file in the Dean's Office for the duration of the career of the teacher at the University of Saskatchewan. Departments and colleges are encouraged to develop a norm base of evaluation outcomes for comparative purposes and to assist with the interpretation of evaluation results over time.

Best Practices

It is important that evaluation of teaching rely on "multiple sources of evidence from multiple parties" (Centra, 1993a; Braskamp & Ory, 1994 as cited by Van Note Chism, 1999, p. 7). Student evaluation of teaching should be conducted serially and information gathered should be collected from various course levels.

Departments and colleges are encouraged to continue using evaluation instruments that have provided reliable data for a long period of time. A norm base on student evaluation of teaching will permit a department or college to interpret evaluation outcomes relative to the base information. Caution is advised when changing questionnaires and approaches to evaluation, especially student rating systems. The design and development of a valid, reliable instrument to be used to measure teaching performance "is a technical task requiring professional expertise in statistics and psychological measurement. It is a common fallacy among educators that all that is required to develop a questionnaire is to sit down and write a set of questions" (Arreola, 2000, p. 94). If students can be identified, the questionnaire may need approval from the appropriate ethics committee. Departments and colleges should ensure that they develop an interpretation of student ratings and a weighting of the criteria used to determine teaching effectiveness (Van Note Chism, 1999, p. 37). Teachers should be provided with information about how the data will be interpreted and weighted for decision-making purposes.

This *Framework* document makes no value judgment on qualitative and quantitative evaluation systems. However, some argue that open-ended questions often "produce a colorful array of responses in the students' own words but provide very little representation (consensus)

information for the instructor to use in formative evaluation . . . The use of closed-ended . . . items . . . can provide accurate counts on the types of responses to each item. The most acceptable and frequently used approach is to use a combination of closed-ended and open-ended items" (Arreola, 2000, p. 95).

Teachers should not administer student evaluation instruments. The task should be given to a designated individual who will distribute the evaluations to the students and give them information about the how the material will be used; this person could be a student-volunteer, student representative or an administrator for the academic unit. Proper instruction and training should be given to individuals charged with administering student questionnaires. These individuals should be advised about the department's or college's evaluation philosophy and the proper protocols for gathering information. Students should be given enough time to complete the questionnaire and give feedback. It is recommended that evaluation instruments be administered at the beginning of a class. The optimum time period for administering a student evaluation during the term is after the deadline when students are permitted to withdraw from the course and prior to the last two weeks of the term.

A best practice is to include a statement on the instrument outlining the importance of teaching evaluations and providing information on how the information will be used. If the outcomes are made accessible to students, they should be advised how to access the information and when the aggregated information might be available to them. As stated earlier, departments and colleges are encouraged to make summative information from the student evaluations of teaching available to students.

Information from the evaluations should be transmitted to the department or college for interpretation and use. The Gwenna Moss Teaching and Learning Centre shall not be involved in the data collection and analysis for summative evaluations. Colleges and departments should have clear understandings of the relative weightings given to peer evaluations, evaluations by students and self-evaluations. It is good practice to use a series of results for a particular course over a reasonable period of time and compare results between teachers teaching the same or similar courses. This comparative data may provide information that can be used to establish teaching evaluation norms and averages in a particular unit. The longitudinal data can be used for discussions concerning improvements to curricula and pedagogy, and can assist with the identification of evaluation anomalies. Serial evaluations can also protect teachers from anomalous courses that produce skewed results (e.g., where a new technology or approach to instruction is attempted and proves unsuccessful) and cohorts of students that consistently differentiate from the norm. Faculty must be provided with comparative data and interpretative material to use the information gathered from teaching evaluations by students. As well, it is important to have follow-up processes and systems in place to support teachers who want to improve their teaching. Subsequent improvements should reflect favorably on the teacher's commitment and performance.

Teaching Evaluations Viewed as a Systematic Process

The decision to adopt a teaching evaluation questionnaire, while clearly an important step, represents only the beginning of a continuing process. Academic units may decide to design a

new questionnaire that specifically addresses the dimensions of teaching quality viewed by that unit to be important. If a new questionnaire is adopted, the unit should gather data designed to assess the reliability and validity of the instrument. This is not an easy task. Since the "unit of analysis" is the individual course taught by a teacher, it is often necessary to gather evaluation data for several years. Statistically valid conclusions regarding the psychometric adequacy of a teaching evaluation questionnaire require relatively large samples of data.

If a unit decides to adopt an existing questionnaire that has demonstrated reliability and validity in some other unit, a principle called "validity generalization" may be applicable. Validity generalization refers to the principle that a questionnaire with demonstrable reliability and validity in one academic setting can be assumed to have reliability and validity in similar settings for other academic units. If an existing questionnaire is adopted by another unit with minor modifications, it should not be necessary to totally replicate earlier studies of psychometric adequacy. See Appendix E for further information about validity generalization.

An example of this principle being applied at the University of Saskatchewan is the teaching evaluation questionnaires used by the College of Law and the Department of Psychology. The College of Law questionnaire was formally adopted in May 1992 after four years of use in Law classes. The data gathered in those four years was used to demonstrate the reliability and validity of the questionnaire. Based on analyses of these data, as reported to the College of Law faculty, the Law questionnaire was formally approved as a reliable and valid instrument.

In adapting the questionnaire for use in the Department of Psychology, an additional dimension examining students' views of the appropriateness of grading evaluation procedures was added (this was absent from the Law questionnaire as most Law courses involved grades based solely on a final examination). Assessment of the questionnaire as adapted by Psychology involved an in-depth examination of the reliability and validity of the new "Evaluation Procedures" dimension, but did not involve an independent replication of the analyses previously conducted for the sections of the questionnaire used in Law. (The teaching evaluation questionnaire eventually adopted by the Department of Psychology is attached in Appendix B.)

Once a questionnaire is adopted by an academic unit based on demonstrable evidence in support of reliability and validity, it is then necessary for that unit to begin the continuing process of developing comparative norms. Comparative norms are developed through a year-by-year accumulation of evaluative results for teachers across all evaluated courses. In the Department of Psychology for example, all teachers are encouraged to conduct teaching evaluations in all courses taught. This eventually allows the development of a database showing how teachers' ratings have evolved over the period since the questionnaire was adopted.

As sufficient evaluation data accumulates over a period of several years, it then becomes possible to compare a particular teacher's ratings in a particular course with the ratings given to all teachers of similar courses over the years. In Psychology, these comparisons are primarily based on the use of percentiles. Comparisons are only made where sufficient data have accumulated across years to ensure confidentiality. In Psychology, courses are compared based on: (1) 1st year, 2nd year, 3rd year and 4th year courses; and (2) groupings based on the content

areas of courses such as developmental psychology, social psychology, cognitive psychology and research methodology.

The questionnaires used in Law and Psychology serve both formative and summative evaluation purposes. Questions detail specific behaviours related to teaching activities. Teachers are encouraged to examine the student ratings received for such questions with a view to continuing those behaviours that are positively viewed, and to improving upon those behaviours seen by students as less positive.

The questionnaires also provide information used in a summative fashion. Five separate and relatively independent dimensions of instructor behaviour are scored by summing student responses across questions. These dimensions are: (1) instructor knowledge; (2) quality of course organization; (3) quality of course presentation; (4) instructor receptiveness to interactions with students; and (5) fairness of course evaluation procedures. A "Total Teaching Effectiveness" score obtained by summing across these five dimensions is also reported. These scores, combined with normative data showing how the scores compare with ratings for other teachers, are used as part of the information considered to make a judgment regarding how a teacher "exceeds", "meets", or "fails to meet" the standard associated with the teaching activities as required by the standards for renewal, tenure and promotion.

Sample Evaluations

Several sample evaluations are provided in the attached appendices. Appendix A contains an approach to formative evaluation known as Stop, Start and Continue. It is recommended for teachers who want to gather formative information mid-way through the term on how to improve a course or to gather information on a new approach or technique for instruction. The information is intended for improvement of teaching performance and for teachers who want to be more responsive to the immediate learning needs of students.

Appendix B is a summative evaluation instrument from the Department of Psychology. Appendix C is a summative evaluation questionnaire for seminar courses in the College of Law. It is recognized that these sample evaluation instruments will not serve the needs of all disciplines and types of courses offered. Appendix A will serve the needs of individual teachers committed to instructional development. Appendices B and C are examples of questionnaires for a lecture format course or a seminar course. There are other course formats, including courses with clinical rotations, laboratory-based courses, independent study courses, graduate courses, performing and studio art courses and other teaching activities, where these samples will not apply. Further exemptions to summative evaluations are outlined in Appendix C of the University Council document *Principles of Evaluation of Teaching at the University of Saskatchewan*

Units that have reliable student questionnaires and evaluation instruments should continue to use them. Units who wish to strengthen their own questionnaires may wish to review these samples. Caution is advised for units that want to adopt aspects of these sample evaluations. Departments or colleges that choose to adopt shorter questionnaires are advised that there are trade-offs with respect to the length of the questionnaire (number of questions used to measure certain indicators

or dimensions of teaching effectiveness) and the reliability of evaluation results. Departments and colleges should ensure that the information that is derived from evaluation instruments is reliable and provides useful information on teaching effectiveness. Issues related to reliability and validity of student evaluations are presented more fully in appendices D and E of this document.

Recommendations for Best Practices:

1. All teachers be evaluated at least once every three years and that teachers under consideration for renewal of probation, tenure or promotion should have their teaching evaluated on an annual basis. The recommendation regarding frequency should be viewed as a minimum. Student evaluations of teaching should be conducted serially and information gathered should be collected from various course levels. Student evaluation of teaching is mandatory.

2. Instruments for the summative student evaluation of teaching should be applied consistently to the teaching of all teachers within an academic unit.

3. Colleges should review their practices with respect to providing summative information from teaching evaluations to students and are encouraged to make the information available, on a restricted basis, to all admitted and registered students at the University of Saskatchewan. This can be done by making the information available through department or college offices, in a department or college resource room, via electronic means or through a local branch of the University Libraries.

4. Teachers who gather formative information from students should advise the students how their views and comments will be taken into account or what the effect will be on future instruction.

5. The optimum time period for administering a student evaluation during the term is after the deadline when students are permitted to withdraw from the course and prior to the last two weeks of the term. It is recommended that evaluation instruments be administered at the beginning of a class. Students should be given enough time to complete the questionnaire and give feedback.

6. Proper instruction and training should be given to individuals charged with administering student questionnaires. Students should be advised of the importance of teaching evaluations and be provided with information on how the information will be used. Teachers should not administer summative student evaluation instruments.

7. Colleges are responsible for approving and validating teaching instruments used in departments and colleges.

8. Departments and colleges are encouraged to develop a norm base of evaluation outcomes for comparative purposes and to assist with the interpretation of evaluation results over time. The interpretation of student ratings and weightings of the criteria should be used to determine teaching effectiveness. Results from a series of a particular course, gathered over a reasonable period of time, should be analyzed and compared to generated averages and percentile rankings

in a particular unit. Comparative norms are developed through a year-by-year accumulation of evaluation results for teachers across all evaluated courses. Comparisons are only made when sufficient data have accumulated across years to ensure confidentiality.

Recommendations for Next Steps

Once University Council has approved this Framework, a number of issues related to teaching evaluations and attention to teaching-related activities should be addressed. It is therefore recommended that:

1. The Instructional Development Committee of Council work with individuals knowledgeable about statistics and evaluation measurement to develop a proposal for financial resources and technical expertise to assist departments and colleges with the validation process for teaching instruments. The proposal will be forwarded to the appropriate Council committees and the Provost's Committee on Integrated Planning for consideration.

2. The Instructional Development Committee of Council consult with stakeholder groups to determine the level of interest in and viability of developing a common questionnaire for student evaluation of teaching. Financial, technological and human resources associated with a University-wide questionnaire are critical pieces of the debate. The costs of developing a common questionnaire and the identification of unit(s) on campus that would be responsible for testing an instrument for reliability and validity and devising a strategy for making outcomes available should be explored. The development and implementation of a common questionnaire for the University of Saskatchewan has been raised during the consultation process and cannot be ignored.

3. The issue of web-based evaluation should form part of the discussion on the development and testing of a common questionnaire. Student groups have expressed a strong interest in collecting evaluation data and disseminating results of evaluations on a website; this is a common practice at many other universities. The IDCC has concluded that it should be a goal of this University to work towards this accomplishment. All stakeholders have an interest in reducing paper and being ecologically sensitive so it is logical that the electronic collection and dissemination of evaluation data should be explored.

4. If consensus is reached on the common questionnaire and resources can be obtained for a pilot project, the development and testing of a common questionnaire should occur in 2004-2005 or as soon as possible thereafter. Further consultation with departments, colleges, student groups, unions representing instructional staff and other parties will be required to determine the interest in adopting a University-wide questionnaire, whether departments and colleges will be given a choice to adopt the common questionnaire or continue to use their own instruments and other such transition issues.

5. Department heads and deans should promote participation in instructional development programs. Follow-up processes and systems should be formalized to support teachers who want to improve their teaching. The Gwenna Moss Teaching and Learning Centre offers workshops to all teachers at the University of Saskatchewan and teachers should be encouraged to take

advantage of these learning opportunities. Similarly, the Educational Support and Development unit in the College of Medicine offers a wide array of programs including workshops on teaching and learning and clinical teaching that are of particular interest to faculty in the Health Science colleges. The Gwenna Moss Teaching and Learning Centre will also design and develop programs to meet the needs of specific departments or colleges. Attendance at teaching-related conferences and symposia such as the events sponsored by the Society for Teaching and Learning in Higher Education (STLHE) should be encouraged.

6. Colleges should establish teaching and learning committees to identify instructional development issues and to devise strategies to support teachers. For example, coaching and mentoring programs, discussion and focus groups, opportunities to videotape lectures and other such approaches can serve to enhance teaching effectiveness.

7. The College of Graduate Studies and Research work with departments and colleges to design evaluation systems for graduate courses and graduate student supervision that provide useful feedback to the teacher while at the same time protecting student confidentiality and anonymity, wherever possible. The evaluation of graduate courses and the supervision of graduate students make it difficult to protect confidentiality and anonymity of students due to the small number of students. In the *Principles of Evaluation of Teaching at the University of Saskatchewan*, University Council recommended that courses with fewer than eight students should use a different instrument for evaluation than those used for larger classes and outcomes from questions with fewer than five responses should be suppressed in order to protect anonymity.

8. Departments and colleges continue to explore best practices for the evaluation of teaching for clinical teaching and team-taught courses consistent with the *Principles of Evaluation Teaching at the University of Saskatchewan.*

9. Participation in teaching events and excellence in teaching should be rewarded and recognized by departments and colleges. College teaching awards should be developed and open to all instructional staff. The nomination information should be well publicized and students should be knowledgeable on how to nominate their best teachers. Deans and department heads should regularly nominate superior teachers for the Master Teacher Award. The University of Saskatchewan should nominate its excellent teachers for external awards such as the 3M Teaching Fellowship. For faculty, excellence in teaching should be rewarded through special increases.

The Framework document does not address evaluation of clinical courses or courses taught by a team of teachers. Nor does it address evaluation processes for other individuals involved in instructional activities such as laboratory demonstrators and teaching assistants. A search for best practices in the evaluation of clinical teaching and team-taught courses at other institutions was inconclusive; thus the document does not include a model or framework for these areas at this time. Further work is required in the evaluation processes for teaching and the University should continue its effort on adopting best practices.

It is important to create a culture on our campus that reflects accountability to our students and a commitment to teaching effectiveness. The University of Saskatchewan should strive to become

known for its teaching excellence and its attention to the evaluation and improvement of teaching.

References

Arreola, A.R. (2000). *Developing a comprehensive faculty evaluation system: A handbook for college faculty and administrators on designing and operating a comprehensive faculty evaluation system*. Bolton, MA: Anker Publishing Company, Inc.

Galbraith, P. (1997). *Student evaluation of instruction: Research implications and potential application.* Calgary, Alberta: University of Calgary.

Lawall, M.L. (1998). *Students rating teaching: How student feedback can inform your teaching*. Winnipeg, Manitoba: University of Manitoba.

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

Strobino, J. (1997). "Building a better "better mousetrap:" Praise for the "start-stop-continue" model of instructional evaluation." University of Nevada, Reno: *The Teaching Professor,* January, p. 6.

University of Saskatchewan. (1998). *A Framework for Planning at the University of Saskatchewan*. Saskatoon, Saskatchewan: University of Saskatchewan.

University of Saskatchewan (2002). *University of Saskatchewan Council - Principles of evaluation of teaching at the University of Saskatchewan.* Saskatoon, Saskatchewan: University of Saskatchewan.

University of Saskatchewan (2002). *University of Saskatchewan standards for promotion and tenure.* Saskatoon, Saskatchewan: University of Saskatchewan.

Van Note Chism, N. (1999). *Peer review of teaching: A sourcebook.* Bolton, MA: Anker Publishing Company, Inc.

Appendix A - Formative Evaluation for Use by Teachers

Stop, Start and Continue Technique

The following approach to formative evaluation is a technique used to gather information to improve the classroom experience for students and to improve instructional techniques (Garner and Emery, 1994 as cited by Strobino). Start, Stop and Continue (SSC) is recommended for use mid-way through the course or whenever it may be appropriate for a teacher to gather information on a new instructional technique or approach. Students are advised to include

feedback on articles, books, assignments, procedures, policies, case presentations, discussion formats, class activities and teaching techniques. Typically, students make three columns on a piece of paper labeling them start, stop and continue (see attached sheet).

- In the Start column, students list any instructional policies, practices or behaviors they would like the teacher to start using;
- In the Stop column, students comment on any instructional policies, practices or behaviors they would like stopped; and
- In the Continue column those that they would like to see continued.

The SSC technique should be administered during class time. Teachers are encouraged to discuss the outcomes the next time the class meets and provide some information to the students about what suggestions may be incorporated into the course and why some may not be or why some changes may take longer to incorporate or adapt. Jane Strabino (1997) from Pennsylvania's Marywood College has found that as she implemented more student suggestions from the evaluations, student responses to the Stop and Start items decreased, while Continue comments increased. When she did not respond adequately to student feedback, a typical response in the Start column might challenge me to "take heed to the evaluation results" (Stabino, 1997, p.6). This is a good mid-course test that should provide some formative feedback to the teacher and may gauge some issues that could emerge in summative evaluations.

START	STOP	CONTINUE

Appendix B - Sample Summative Evaluation by Students

from the Department of Psychology

*** NOTE: Although the practice in the Department of Psychology is to permit unsigned evaluations, it is recommended that evaluations by students be signed and data aggregated or summarized when used for summative purposes.**

1. The questions (in the form of statements) are grouped under various headings. It will be helpful in completing the form if you read over these headings before you begin.
2. One response should be checked for each question.
3. Suggested time for completion is approximately 20 minutes.
4. Respond to each statement by selecting one of the following:

A. Strongly Agree
B. Agree
C. Somewhat Agree
D. Somewhat Disagree

E. Disagree
F. Strongly Disagree

5. If a statement is not applicable, mark "G" on the answer sheet for that item.

ON THE ANSWER SHEET PROVIDED, PLEASE FILL IN THE BLANKS FOR CLASS AND SECTION, INSTRUCTOR, AND DATE. ALSO WRITE "PSYCHOLOGY MAJOR" OR "OTHER" AT THE TOP OF THE ANSWER SHEET.

YOU DO NOT HAVE TO SIGN YOUR NAME ON THE ANSWER SHEET OR ON THIS QUESTIONNAIRE.

STUDENT PERCEPTION OF PROFESSOR KNOWLEDGE

1. The instructor is familiar with the concepts and information included in the course.
2. The instructor understands the material sufficiently to discuss the subject matter fluently and in a sophisticated way.
3. The instructor is familiar with related materials other than those assigned in the course.
4. The instructor demonstrates an understanding of how this course relates to other courses in the program.
5. The instructor demonstrates awareness of current developments relating to the course.

COURSE ORGANIZATION

6. The subject matter of this course is organized in a coherent way.
7. The content of this course is what I expected from the description I was given in the course outline.
8. The intellectual level of this course is what I expected from the description I was given in the course outline.
9. It is clear what the objectives of the course were.
10. The amount of time allocated to each topic is appropriate.
11. The materials for this course adequately cover the content of the course.
12. The materials assigned are relevant to the content of the course.
13. The instructor selected materials that are helpful in understanding the content of the course.
14. The materials are edited in a helpful way.
15.The materials are well organized.

COURSE PRESENTATION

The instructor...

16. ..... has a good voice projection.
17. ..... uses clear sentence construction.
18. ..... employs an appropriate vocabulary.
19. ..... varies inflection and expression.
20. The instructor shows sensitivity to the general level of comprehension of the class.

21. The instructor is aware when she/he has not been understood.
22. The instructor is able to adapt to the needs of individual students without losing the interest and attention of the remainder.
23. The instructor maintains the focus of the class, and knows when to terminate discussion.
24. The instructor is enthusiastic and stimulating.

25. The instructor uses techniques which challenge thought and stimulate discussion.

STUDENT-PROFESSOR INTERACTION
26. The instructor has the respect of the class.
27. The instructor provides a positive example of professional responsibility.
28. The instructor helps students deal with problems encountered in the course.
29. The instructor gives positive reinforcement.
30. The instructor makes specific suggestions for improvement.
31. The instructor encourages students to see her/him outside class.
32. I feel comfortable approaching the instructor with my questions outside class.
33. The instructor shows a readiness to modify his/her teaching techniques in light of student feedback.
34. The instructor uses a variety of techniques to present material.
35. The instructor uses innovative teaching methods which are effective.

COURSE GRADING CRITERIA
36. I find the feedback received from the instructor with my exams and assignments to be helpful.
37. The instructor's expectations of how much we are required to learn is reasonable.
38. Assignments are given in a way that makes it clear what is expected of me.
39. The types of exams I have to prepare for facilitate my learning.
40. The types of assignments we are given facilitate my learning.
41. Grading criteria are always stated clearly as the assignments are assigned.
42. I think the instructor's grading of exams is fair.
43. I think the instructor's grading of assignments is fair.
44. I think the content of exams fairly tests what is emphasized as important concepts to be learned in the course.
45. I understand, based on feedback from the instructor, how my grades are determined.

OTHER COMMENTS - It would be helpful if you could give some thought to the following questions:

46. What <u>ONE</u> thing <u>COULD HAVE BEEN DONE</u> differently in this course which would have made the course better?
47. What <u>ONE</u> thing <u>WAS DONE</u> in the course that most facilitated your learning?
48. What other comments do you have about this course and your instructor?

Appendix C - Sample Summative Evaluation by Students

from the College of Law (Seminar Course)

Please note:

1. The questions (in the form of statements) are grouped under 6 headings. It will be helpful in completing the form if you read over these headings before you begin.
2. Complete section (x) (questions 51-56) only for a seminar involving the writing of a major paper.
3. One response should be checked for each question.
4. Suggested time for completion is approximately 15 minutes.
5. Respond to each statement by selecting one of the following:

a. Strongly Agree
b. Agree
c. Somewhat Agree
d. Somewhat Disagree
e. Disagree
f. Strongly Disagree

ON THE ANSWER SHEET PROVIDED PLEASE FILL IN THE BLANKS FOR CLASS AND SECTION, INSTRUCTOR, AND DATE.

(i) Knowledge of the Subject

1. The instructor is familiar with the concepts and information included in the course.
2. The instructor understands the material sufficiently to discuss the subject matter fluently and in a sophisticated way.
3. The instructor is familiar with related materials other than those assigned in the course.
4. The instructor demonstrates an understanding of how this course relates to other courses in the law program.
5. The instructor demonstrates awareness of current developments in the law relating to the course.

(ii)Organization of the course and reasonable adherence to course description in the context of curricular design.

6. The subject matter of this course is organized in a coherent way.
7. The content of this course was what I expected from the description I was given.
8. The intellectual level of this course was what I expected from the description I was given.
9. It was clear what the objectives of the course were.
10. The amount of time allocated to each topic was appropriate.
11. The total amount of material covered was reasonable.

(iii) Selection, editing and currency of course materials.

12. The materials for this course adequately cover the content of the course.
13. The materials assigned were relevant to the content of the course.

14. The materials assigned were at a level of difficulty appropriate for the students in the course.
15. The instructor selected materials that were helpful in understanding the content of the course.
16. The materials were edited in a helpful way.
17. The materials were well organized.
18. The materials were up-to-date.
19. The syllabus or other course guide was a helpful aid to using the materials.
20. Useful materials other than those assigned were identified.

(iv) Clarity of communication.
The instructor . . .
21. . . . had a good voice projection.
22. . . . spoke too quickly.
23. . . . spoke too slowly.
24. . . . enunciated clearly.
25. . . . used clear sentence construction.
26. . . . employed an appropriate vocabulary.
27. . . . varied inflection and expression.

(v)Awareness of and responsiveness to students' level of comprehension.
28. The instructor showed sensitivity to the general level of comprehension of the class.
29. The instructor was aware when she/he had not been understood.
30. The instructor was able to adapt to the needs of individual students without losing the interest and attention of the remainder.
31. The instructor maintained the focus of the class, and knew when to terminate discussion.

(vi) Rapport with students in the class/seminar room.
32. The instructor had the respect of the class.
33. The instructor seemed to enjoy teaching.
34. The instructor did not intimidate or patronize students.
35. The expression of viewpoints in conflict with the instructor's own was encouraged.
36. The instructor was sensitive to and respectful of the dignity of individual students.
37. The instructor showed sensitivity in handling matters of: the economically disadvantaged; minority cultures; Aboriginal peoples; women; the disabled; gay or lesbian persons.

(vii) Ability to foster in students skills and understanding relevant to the legal profession, and to develop in them an awareness of and sensitivity to the ethical standards and social demands of the profession.
38. The instructor provides a positive example of professional responsibility.
39. The instructor contributes to my understanding of the ethical standards of the legal profession.
40. The instructor presents a balance, appropriate to the course, of practical, policy and theoretical aspects of the law.
41. The instructor made use of opportunities to raise and address issues of race, culture, gender, sexual orientation, disability and economic status.

(viii) Extent to which student interest is aroused, curiosity quickened and independent, critical

thought engendered.
42. The instructor is enthusiastic and stimulating.
43. The instructor encourages students to take responsibility for their own learning.
44. The instructor uses techniques which challenge thought and stimulate discussion.
45. The instructor directs students to useful additional material.
46. I want to learn more in this subject area.

(ix) Provision of feedback on student progress.
47. The instructor provided sufficient direction and feedback on student performance.
48. The instructor helps students deal with problems encountered in the course.
49. The instructor gives positive reinforcement.
50. The instructor makes specific suggestions for improvement.

(x)Supervision of research paper. [Complete this section only if applicable]
51. The instructor was helpful in the selection of an appropriate topic.
52. The instructor directed me to the best available sources of information in my subject area.
53. The instructor made him/herself available for regular consultations on the subject.
54. The instructor's insights helped me to focus and clarify my reasoning.
55. The instructor's guidance improved my paper.
56. I had the freedom to deal with my subject in a way that was comfortable for me.

(xi) Availability to students outside the classroom for guidance and assistance in their studies.
57. The instructor encouraged students to see her/him outside class.
58. The instructor was available to discuss questions in his/her office.
59. I felt comfortable approaching the instructor with my questions outside class.
60. The discussions outside class helped me to understand the material covered in class.
61. The discussion outside class increased my interest in and curiosity about the legal issues discussed.

(xii) Innovation in teaching techniques.
62. The instructor has a concern for the quality of her/his teaching.
63. The instructor showed a readiness to modify his/her teaching techniques in light of student feedback.
64. The instructor uses a variety of techniques to present material.
65. The instructor uses innovative teaching methods which are effective.

Appendix D - Issues Related to Reliability and Validity of Evaluations by Students

There has been considerable research on the validity and reliability of student evaluation of teaching. A well-designed teaching evaluation system should ensure that the information provides useful data on teaching and provides consistent and accurate results. Validity focuses on the whether one is measuring what one intended to measure and reliability focuses on whether one is measuring consistently and accurately (Van Note Chism, 1999, p. 14). Reliability and

validity are technically complicated issues. Individuals or units interested in ways of assessing reliability and validity are encouraged to research these issues or get advice from someone knowledgeable about psychometric testing and analysis. Studies by Murray (1983) and Cranton and Hillgarten (1981) demonstrate that a high correlation exists between evaluations by students and teaching behaviours observed by external observers (as cited by Galbraith, 1997, p. 6).

Scriven (1995) has indicated reasons why student ratings of instruction are potentially valid. These reasons are:

1. The positive and statistically significant correlation of student ratings with learning gains.

2. The unique position and qualifications of the students in rating their own increased knowledge and comprehension.

3. The unique position of the students in rating changed motivation (a) toward the subject taught; perhaps also (b) toward a career associated with that subject; and perhaps also (c) with respect to a changed general attitude toward further learning in the subject area, or more generally.

4. The unique position of the students in rating observable matters of fact relevant to competent teaching, such as the punctuality of the instructor and the legibility of writing on the board.

5. The unique position of the students in identifying the regular presence of teaching style indicators. Is the teacher enthusiastic; does he or she ask many questions, encourage questions from students, etc.?

6. Relatedly, students are in a good position to judge--although it is not quite a matter of simple observation-such matters as whether tests covered all the material of the course.

Scriven cautions that checks and balances must exist to ensure that teachers do not gather information used for summative purposes, that no special considerations or incentives be given to students that might influence their evaluations, and appropriate time must be given to administer evaluations. It is important to get an appropriate evaluation in place to gather information that is useful, reliable and valid for making important career-related decisions.

Appendix E - Methods for Assessing Reliability and Validity of Evaluations by Students

The methods for assessing reliability and validity of evaluations by students set out in Appendix E were contributed by Professor Dave Scott, an expert in applied statistics, psychometrics and measurement, in the Department of Psychology at the University of Saskatchewan.

Assessment of the reliability and validity of teaching evaluation questionnaires involves application of data analysis procedures broadly characterized as psychometric analyses. The

definitions (and typical assessment methods) of reliability and validity of a measurement procedure are discussed next.

Reliability refers to the extent to which a test or questionnaire measures its intended characteristic with a minimum of measurement error.

Typical methods for assessing reliability include:

(1) administering the same test twice to a group of individuals, and correlating the scores across the two administrations - this method is called test-retest reliability;

(2) administering two similar versions of the same test (i.e., two forms of the test consisting of questions which are different but judged to similarly assess the same characteristic) - this method is called parallel forms reliability; or

(3) analysing the scores for a single test or questionnaire administered once to determine the extent to which the individual test questions all consistently measure the same intended characteristic - this method uses a reliability analysis technique called "Coefficient Alpha" and is the method most commonly encountered when teaching evaluation questionnaires are assessed to determine reliability.

Reliability is quantitatively assessed and theoretically ranges from 0.00 (indicating that the test or questionnaire scores only reflect random errors) to 1.00 (indicating that the test or questionnaire scores are perfectly accurate and involve no random errors). Typical reliability values range from 0.95 and up for standardized tests of ability such as the Wechsler Adult Intelligence Scale - Third Revision. Reliability values between 0.80 and 0.90 are often seen for standardized tests of attitudes or sentiments such as the Strong Interest Inventory, a commonly used questionnaire designed to assess potential satisfaction with various occupations and careers.

For any test or questionnaire, including teaching evaluation questionnaires, it is obviously desirable that reliability be as high as possible. One well-known influence on reliability is the number of questions that are totalled to produce a score. Simply stated, the reliability of a test is a direct function of the number of questions on that test; longer tests are more reliable than shorter tests. While it is possible to have a teaching evaluation procedure involving only one or two questions (such as: "How do you rate this teacher?" and "How willing are you to register for another course taught by this teacher?"), it is inevitable that having only one or two questions would lead to lower reliability. This is why most teaching evaluation questionnaires have a number of questions assessing a common characteristic of the teacher such as "effectiveness of course organization" to ensure that the final, summed score assessing that characteristic has acceptably high reliability.

Reliability is a necessary characteristic of any test. A test can only be valid if it is reliable. This leads next to the concept of validity.

Validity refers to the extent to which a test or questionnaire measures what it is intended to measure. Unfortunately, this apparently simplistic definition leads to a complex set of issues and quantitative assessment procedures.

Traditionally, validity is conceptualized in three major ways; these are: (1) content validity; (2) criterion-related validity; and (3) construct validity.

Content validity refers to the extent to which the questions on a test or questionnaire adequately reflect the known content domain of the characteristic the test is intended to assess. As an example, tests such as the Canadian Tests of Basic Skills are developed by experts in educational curricula to ensure that the questions relate to what is known to be the knowledge and skill areas believed important in elementary and secondary schools. For teaching evaluation questionnaires, content validity should be judged by the extent to which the questionnaires survey opinions on what are known to be important characteristics and behaviours of teachers. Ultimately, decisions regarding the content validity of any measurement procedure are based on subjective evaluations comparing what is believed to be the content domain of the measured characteristic with the extent to which the test or questionnaire content taps into that domain.

Criterion-related validity refers to the extent to which a test or questionnaire adequately predicts performance in situations removed in time or space from the test setting. As a simple example, tests of driving skill are intended to predict the skill level of a person applying to receive a driver's license. As another example, tests such as the Graduate Record Exam are intended to predict potential academic performance of applicants for graduate study.

Applying this concept to the area of teaching evaluations, it could be predicted that teachers receiving very positive ratings from their students are more likely to receive external, positive recognition of their teaching prowess. Highly rated teachers could also be expected to have students who are more successful in their later endeavours as assessed by such things as academic rewards and scholarships.

It is often difficult, if not impossible, to use criterion-related validity concepts to assess the validity of teaching evaluation questionnaires. Most teachers do not receive external rewards recognizing teaching prowess, and even when such rewards are made, the decision is usually based on a consideration of what that teacher has done over a period of several years. Data which could be used to demonstrate a connection between how a teacher is evaluated and the extent to which that teacher receives external recognition of teaching excellence would be sparse, and time-consuming to obtain.

Assessing the validity of a teaching evaluation questionnaire by noting the accomplishments of students who have been taught by highly-rated teachers would also involve data which is sparse and time-consuming to obtain. Another, more serious problem exists however. Many teaching evaluation questionnaires promise confidentiality to the responding student, even if that student is required to sign the evaluation form. To obtain the necessary data to link future student performance to earlier evaluations of the teachers of that student would breach any promise of confidentiality.

For these and other reasons not discussed, the concept of criterion-related validity is not particularly helpful in determining the extent to which a teaching evaluation questionnaire is valid.

The concept of construct validity stems from the fact that any test or questionnaire consists of a collection of items that are designed to collectively assess some conceptualized trait. A conceptualized trait (e.g., verbal intelligence or depressive ideation or approachability of a teacher) is operationally defined as a constellation of behaviours or characteristics. Literally, the definition of the trait to be assessed is by "construction" of a collection of observable characteristics. Construct validity basically asks two questions. The questions are: (1) is the conceptualization of the measured trait "correct"?; and (2) if the answer to question (1) is "yes", does the test or questionnaire purporting to measure that trait do so in an empirically verifiable manner?

Many measurement experts would say that the concept of construct validity subsumes all other types of validity. Whatever position one takes on this issue, it is clear that the concept of construct validity must be central to any research designed to show that a teaching evaluation procedure is valid.

Evidence for construct validity is multi-faceted. Evidence for construct validity cannot be reduced to a single quantitative index as is possible for reliability. It is never "proven" that any test or questionnaire has construct validity. Instead, it is necessary to subjectively consider multiple sources of information from empirical studies. Hopefully, the results of those studies will provide converging evidence in support of a hypothesis that a test or questionnaire possesses construct validity.

Types of empirical studies that provide construct validity evidence for teaching evaluation questionnaires are as follows:

(1) Demonstration through a multivariate analysis technique called factor analysis that the questionnaire items do assess what has been shown in the research literature to be the salient characteristics of a teacher. Typically, a well-designed questionnaire has been put together with an intent to assess what is believed to be important and relevant teacher activities. Having a clear understanding the content domain of teacher characteristics hopefully leads to collections of items that accurately assess those characteristics. Evidence from factor analytic studies should be provided to demonstrate that this is so;

(2) If a collection of questionnaire items comprising a scale can be shown to be highly reliable through the use of Coefficient Alpha (e.g., a collection of items addressing issues regarding the effectiveness with which a teacher presents the course), it will also be the case that those items can be shown through factor analysis to collectively assess a common underlying construct or trait. Thus, evidence for questionnaire reliability also provides evidence for construct validity;

(3) Accumulation of teaching evaluation data for a group of teachers across several years permits other empirical studies of the construct validity of the questionnaire. As a specific example,

investigations of the questionnaire developed by the College of Law at the University of Saskatchewan showed the following:

(3.1) In those cases where a specific course had been taught by various Law faculty across several years, ratings for individual teachers across years were more consistent than ratings for all faculty teaching that specific course. In short, if a teacher was positively or negatively viewed by one group of students in a course, that teacher would be viewed similarly by other groups of students in subsequent years when the same course was taught.

(3.2) Ratings for Law faculty teaching various courses were more consistent across multiple offerings of a specific course, when compared to the ratings of those same faculty across different courses. In short, the Law teachers were seen by students to be better at teaching some courses than other courses, independent of the overall level of reported student satisfaction for those teachers; and

(3.3) Ratings for Law faculty across years, when changes did occur, were seen to be in the direction of consistently improving ratings. The conclusion here was the Law faculty were taking their teaching ratings into consideration and modifying some of their teaching behaviours as a result. These modifications were recognized by their students and resulted in improved ratings.

These are but three examples of empirical studies that were conducted to assess the construct validity of a specific teaching evaluation questionnaire currently in use at this university. In general, such studies must investigate the extent to which scores on the questionnaire reflect predicted results anticipated from use of the questionnaire over time.

Investigations conducted to show reliability and validity of a teaching evaluation questionnaire requires several years of data accumulation. The reason is that the unit of analysis is the teacher of a specific course. Evaluation of that teacher by 20 or 50 or even 350 students results in only one data point. Psychometric analyses to assess reliability and validity are data intensive. A rough "rule of thumb" accepted by most measurement experts is that the number of necessary data points to conduct statistically valid investigations is ideally ten times the number of questions. Thus, a questionnaire with 30 questions should be administered in 300 courses before the data is sufficient to permit psychometric analyses. In the College of Law, data were gathered for approximately six years from 1988 to 1994 before empirical investigations were carried out.

Any unit wishing to develop a student-based teaching evaluation questionnaire from the ground up needs to be aware of how lengthy a process will be necessary for that development to be demonstrably successful. That said, some units might decide to rely on the concept of "validity generalization" as they contemplate adopting a new questionnaire for evaluation of teaching. Validity generalization refers to the idea that a test or questionnaire shown to have reliability and validity in one setting can be presumed to also be reliable and valid if used at another setting relatively similar to the first.

If an existing questionnaire, shown to be reliable and valid, is modified or adapted for use in a different setting, only those components which have been modified need to be analyzed psychometrically. For example, if five new items are added to a questionnaire, or replace five

existing items, it should be sufficient to assess the merit of those items by as few as 50 teacher ratings. In units with many course offerings, such could be accomplished in one or two years.

This appendix was prepared in response to requests for more detailed information regarding how one assesses the reliability and validity of a teaching evaluation questionnaire. As mentioned earlier in the report, determining the reliability and validity of a questionnaire involves technically complicated analyses. Individual units contemplating such analyses are well advised to seek expert advice and assistance as soon as possible in the process of questionnaire development and implementation.