24

# Developing and Validating an Instrument for Student Ratings of Teaching

Gary Hunt, Lyn Baldwin, Ernest Tsui, & Les Matthews
Thompson Rivers University

*In May 2007, the Thompson Rivers University Faculty of Science established an ad hoc subcommittee to develop a new student ratings of teaching survey. The final survey, approved by the Faculty in February 2011, includes statements categorized in the dimensions of teaching shown in previous studies to be correlated with student achievement. The survey is learner-centred, discipline and pedagogically neutral, and includes only items that can be reasonably evaluated by students. The survey consists of 40 items including eight statements of student background information, 32 statements to rate on a six-point Likert scale, and four open-ended questions. We demonstrated that a faculty group with no formal training in survey design and informed by the literature, can, in collaboration with faculty, develop a survey established as having a high degree of inter-rater reliability.*

## Introduction

At Thompson Rivers University (TRU), student ratings of teaching are obligatory for new faculty and are mandated components of promotion and tenure packages. In 2007, the Faculty of Science determined that our existing survey was not meeting our needs for a number of reasons: faculty expressed a wide range of concerns about specific items, it was not consistent with recommended practices in the current literature, and it had never been properly assessed for validity and reliability. The need for a new student ratings of teaching form can be met through one of several ways including modifying an existing form or the purchase of a commercial product. Following a review of several surveys from other institutions and commercial products, we concluded that the best solution to meet the needs within the TRU Faculty of Science was to develop our own survey. The survey resulting from this work is included in the Appendix.

While it is recognized that student ratings of instruction are controversial and best used in combination with other forms of evidence (e.g., peer-observation and dossiers), they remain widely used in evaluating teaching and provide the best option for gathering quantifiable and comparable data (Abrami, 2001). In our Faculty of Science, student ratings of teaching remain controversial among some members yet play important roles in formative and summative evaluations. Our intent was to develop a ratings survey that supported the mission of our University and met with faculty approval.

# Methodology

Two primary goals underlined our approach to developing a new student ratings survey. First, we wanted our process to be collaborative, gathering input and approval through focus groups and presentations to our Science Faculty Council. Second, we wanted to employ the recommended practices as outlined in Berk (2006) and Gravestock and Gregor-Greenleaf (2008). Below, we outline the steps we completed from 2007 to 2011.

1. Considering TRU's mission statement, we identified teaching dimensions to be included in the survey.
2. We developed a schematic of our approach based on best practices, summarized the teaching dimensions planned for the survey, and presented this approach to Faculty Council for approval.
3. We compiled a list of potential statements for each teaching dimension drawn from published examples (Adams et al., 2008; Gravestock & Gregor-Greenleaf, 2008). Once compiled, each statement was rewritten, if necessary, to be student-centered and evaluated against the criteria outlined by Berk (2006). The criteria range from grammatical guidelines to rules regarding the relevance or applicability of each statement to potential student respondents.
4. We received ethics approval from the TRU Human Ethics Review Board.
5. We held focus groups with faculty and students. Both students and faculty in each focus group were asked individually to: 1) rank the importance of all statements in each dimension; and 2) respond to a series of questions regarding the applicability and neutrality of all statements.
6. We modified the statements based on focus group feedback.
7. We conducted the field test with classes of students. We used email solicitation and presentations within our Faculty Council to solicit faculty who would volunteer their classes to complete the draft survey online.
8. We completed statistical analyses of field test data. We completed the quantitative analysis suggested by Berk (2006) including the following: 1) determine the mean and standard deviation of each item in the survey; 2) assess inter-item correlations between statements in each dimension; and 3) asses item-scale correlations. The TRU Faculty of Science includes a diverse group of faculty teaching a wide variety of lecture, lab, clinical, and field courses. Thus, we also included a "Not Applicable" (NA) response for each statement in the draft survey. As part of our analysis, we identified statements that elicited a large number of NA responses.
9. Based on the results of the field test (see Results), we made the final selection of statements.
10. We prepared a report (Baldwin, Matthews, Tsui, & Hunt, 2011) and requested approval from Science Faculty Council.
11. We incorporated the final changes suggested by Faculty Council and released the survey for use.

# Results

## Selection of teaching dimensions and modification of statements

We referred to TRU's mission statement to guide our process:

> Thompson Rivers University is a comprehensive, learner-centred, environmentally responsible institution that serves its regional, national, and international learners and their communities through high quality and flexible education, training, research and scholarship. (Thompson Rivers University, 2007)

Thus, we selected six teaching dimensions that had been previously shown to be highly correlated with student achievement including 1) preparation and organization; 2) clarity and understandableness; 3) perceived outcome or impact; 4) stimulation of interest in content; 5) encouragement and openness; and 6) availability and

helpfulness (Abrami, d'Apollonia, & Rosenfeld, 2007; Abrami, Rosenfeld, & Dedic, 2007; Feldman, 1989, 2007).

Although learner-centred teaching has many components (Barr & Tagg, 1995), we felt that a critical aspect of learner-centred teaching was linking the influence of teaching activities to student learning. Given that learner-centred teaching requires that teachers evaluate the role that teaching activities play in student learning, the committee decided that it would focus the evaluation on those dimensions of teaching that previous research has found to be most strongly linked to student learning. By actively tailoring the evaluation to aspects (or dimensions) of teaching that have been found to support student learning, we believed that we would be constructing a student evaluation of teaching form that would best support the goals of TRU's Faculty of Science.

We also recognized that students cannot be asked to respond to questions for which they are not qualified to answer. Students are most qualified to report on their experience within the class and are not qualified to report on the experience of other students. Thus, we included only statements that directly reflect a student's experience within a course. To emphasize this distinction, all statements in the evaluation form were written from the student's perspective. For example, compare the two statements below:

> Instructor centred: The instructor was well organized for class.
>
> Student centred: I think the instructor was well prepared for class.

Likewise, students do not have the expertise to comment on scope and currency of the curriculum or an instructor's depth of knowledge, which are best evaluated by faculty peers (Berk, 2006). Because the Faculty of Science includes diverse departments and faculty using a wide variety of teaching strategies, it was important that individual statements be pedagogically neutral. That is, as much as possible, each statement would be equally applicable to all students within all classes. Thus, to guide our selection of statements in each teaching dimension we followed the principles of selecting measurable factors that relate to student achievement and are independent of both discipline and pedagogy. In addition, we used only factors that students are qualified to report upon and that faculty can control and improve upon through targeted development activities.

## Focus groups and field test

In total, 17 faculty members and 26 students participated in the focus groups. At the time we completed the focus groups, the Faculty of Science consisted of approximately 70 faculty members and 500 students. In the field test, 14 faculty members participated in volunteering 16 courses, including four lab sections, one online course and 10 lecture courses. The courses also ranged across the curriculum from 100-400 level courses. By program, the proportion of participating students ranged from approximately 21 to 66% (Table 1). The departments of Computing, and Math and Statistics were not in the Faculty of Science at the time of this study.

Table 1

*Number of students participating in the field test and total enrolment by program*

| Program of Students | Number of Students | Enrolment |
|---|---|---|
| Animal Health Technology | 44 | 126 |
| Biological Sciences | 186 | 281 |
| Physical Sciences | 50 | 246 |
| Natural Resource Science | 44 | 97 |
| Respiratory Therapy | 30 | 145 |
| Unreported | 12 | N/A |
| Total | 366 | 895 |

### Quantitative analysis of field test results

Overall, the global mean score provided for all faculty on all items was 3.36 (out of a possible 4.0). Mean scores for individual items ranged from 2.95 to 3.64. The standard deviation of each item indicates whether or not each item solicited a range of responses from students. The standard deviations of items evaluated in the field test ranged from 0.5 to 0.87, which Berk (2006) suggests is an adequate range of variation.

Items within each dimension are meant to evaluate the same teaching dimension. Correlations between mean scores for each statement evaluate how well the statements correlate with one another. Berk (2006) recommends that all correlations should be positive and should be as high as possible. Inter-item correlations resulting from our field test ranged from 0.30 to 0.73. Overall, teachers who score high in one category typically score high in the other dimensions, and item-scale correlations evaluate how well the score for each item correlates with the global score for all items (minus the specific item being evaluated). The item-scale correlations ranged from 0.52 to 0.79. We also evaluated the number of NA responses each item solicited in the field test. Based on these three items, inter-item correlations, item-scale correlations and the number of NA responses, we identified two statements that should be deleted from the survey. We also moved two statements into a more appropriate dimension (Table 2).

While the range of mean scores is higher than the 2.0 mean that Berk (2006) recommends, we believed that the relatively high scores may have resulted from the self-selection process whereby faculty who typically have high student evaluation scores were the most likely to volunteer their classes for the field test. To evaluate the long-term trends in scores, the Faculty of Science is currently attempting to collect baseline data. At this time, the collective agreement language provides no mechanism for the collation of faculty scores.

## Discussion

Feedback from student ratings of teaching surveys is one source of information that contributes to evaluating teaching effectiveness. Given the complexity of teaching, it is vital that student feedback be complemented by evaluations from peer and expert observers, as well as self-evaluations (Adams et al., 2008; Murray, 2005). This is necessary to ensure that the whole picture of an instructor's teaching is evaluated.

We decided not to use "overall" or "global" statements such as, "Overall, I would rate this instructor's teaching performance as…." These are quite common items in student questionnaires, but since they do not reflect the multidimensionality of teaching, the information they provide may not be very meaningful at best, and can be misleading at worst (Murray, 2005). Global items are recommended by some experts (Abrami, 2001; Arreola, 2007) because they are correlated moderately with student learning and they provide a single-value summary of teaching (Berk, 2006). Because global items provide no value for formative feedback and are less reliable than subscale and total scale scores, we chose not to include them in this survey. Sub-scale means are

Table 2

*List of questions that were deleted or moved to another teaching dimension based on the results of the field test*

| Dimension | Full text of Question | Action |
|---|---|---|
| Preparation and Organization | I was satisfied with the time it took for the instructor to return graded material. | Moved to Dimension 6 based on low inter-item correlation coefficients |
| Clarity and Understanding | The work I completed in this course (e.g., assignments, homework or class activities) increased my learning. | Deleted due to high NA responses |
| Perceived Outcome or Impact | My problem solving skills improved as a result of this course | Deleted due to high NA responses |

often averaged across student evaluation tools to provide a global mean for individual faculty. However, global means estimated across all statements obscure the relative importance of each dimension (i.e., Preparation and Organization explains 30-35% of student achievement, while Availability and Helpfulness explains less than 10% of student achievement [Feldman, 1989, 2007]). While global means may be weighted by the amount each dimension explains of overall student achievement, any single value produced by a student evaluation masks the details of variation among the dimensions. The committee was concerned that reviews of faculty member's teaching (especially summative reviews for promotion and tenure) would rely too heavily upon a global mean, if produced.

Research shows that the information gained from student ratings surveys has a limited impact on teaching improvement if not accompanied by appropriate professional development activities (Cohen, 1981; Feldman, 1989, 2007; Marsh, 2007; Murray, 2005; Kulik, 2001; Wachtel, 1998). The Centre for Teaching and Learning at TRU offers professional development opportunities to assist faculty in addressing individual concerns about item ratings. In considering the design of our new ratings form, it was precisely this formative value that we wanted to emphasize. In other words, we would like the instrument to be used as a tool to help instructors improve their teaching skills, in addition to its current use in summative or personnel decision processes. Furthermore, we needed to ensure the teaching practices addressed in the form align with the educational goals and objectives of TRU Science (e.g., learner-centredness and student learning). Our decision to make the new ratings form based primarily on learners' experiences addresses this. We believe our collaborative approach led to the successful adoption of the survey by our Faculty and that our process can serve as a model for other faculties.

## References


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

Abrami, P.C., d'Apollonia, S., & Rosenfeld, S. (2007). The dimensionality of student rations of instruction: An update on what we know, do not know, and need to do. In R. P. Perry & J. C. Smart (Eds.), *The Scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 385-445). Dordrecht, The Netherlands: Springer.

Abrami, P.C., Rosenfeld, S., & Dedic, H. (2007). Commentary: The dimensionality of student ratings of instruction: What we know, and what we do not. In R.P. Perry & J.C. Smart (Eds.), *The Scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 385-446). Dordrecht, The Netherlands: Springer.

Adams, V., Bleicher, R., Buchanan M., Nuhfer E., Elliot J., Furmanski, M. Renny C., Smith, P., & Wood, G. (2008). *Final report of the task force to create a new student ratings of teaching for California State University Channel Islands*. Retrieved from http://www.csuci.edu/academics/faculty/facultyaffairs/evaluation.htm

Arreola, R. (2007). *Developing a comprehensive faculty evaluation system: A handbook for college faculty and administrators on designing and operating a comprehensive faculty evaluation system.* Anker Publishing Co., Bolton, MA.

Baldwin, L., Matthews, L., Tsui, E., & Hunt, G. (2011). *Developing and evaluating an instrument for student evaluation of teaching.* Retrieved from http://www.tru.ca/ctl/resources.html

Barr, R. & Tagg, J. (1995). From teaching to learning: A new paradigm for undergraduate education. *Change, The Magazine of Higher Learning.* Nov/Dec., p.13-25.

Berk, R. (2006). *Thirteen strategies to measure college teaching*. Sterling, VA: Stylus.

Cohen, P.A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research*, *51*, 281-309.

Feldman, K.A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education*, *30*, 583-645.

Feldman, K. (2007). Identifying exemplary teachers and teaching: evidence from student ratings. Pages 93-143 *in* R. P. Perry and J. C. Smart, editors. The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective. Springer Netherlands, Dordrecht.

Gravestock, P. & Gregor-Greenleaf, E. (2008). *Student course evaluations: research, models and trends.* Higher Education Quality Council of Ontario. Retrieved from http://tinyurl.com/97fhcst

Kulik, J.A. (2001). Student ratings: validity, utility, and controversy. *New Directions for Institutional Research*, 9–25.

Marsh, H. (2007). Students' evaluation of university teaching,: dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective.* (pp. 319-383) Springer Verlag, Dordrecht.

Murray, H.G. (2005). Student evaluation of teaching: has it made a difference? Paper presented at the Annual Meeting of the Society for Teaching and Learning, Charlottetown, PEI.

Thompson Rivers University. (2007). Retrieved from http://www.tru.ca/__shared/assets/2007-2012_strategic_plan8326.pdf

Wachtel, H.K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education, 23*, 191–212.

## Biography

Gary Hunt is Coordinator, Teaching and Learning Support at Thompson Rivers University. His interests in higher education include student ratings of teaching, learning outcomes and assessment, and learning-centred instruction.

# Appendix

## Thompson Rivers University Faculty of Science Student Ratings of Teaching Survey

THOMPSON RIVERS UNIVERSITY

**Student Feedback on Course Learning Experience**

**Course Number & Section:_____Instructor Name:_____**
**Date:_____(day/month/year)**

**Introduction**
You are a critical source of information about the effectiveness of the instruction that you have received. Your thoughtful responses are appreciated and will be used to identify aspects of your instruction that are meeting your learning needs and those that need to be improved.

Please complete the survey alone, not in consultation with your classmates.

Student feedback will remain confidential and responses will be returned to the instructor only after all grades have been submitted to the registrar.

This information will be used by individual faculty to improve their teaching. In addition, information from this survey will be made available to department Chairs and Deans for the purpose of assessing instructors.

This questionnaire contains three sections. In the first section, we would like you to tell us a little about yourself.

The second section contains statements about your learning experience with your instructor. Please read each statement carefully and the rate the extent to which you agree with the statement as a reflection of your experience in the class. Consider each statement separately and assess it based on your actual experience.

Finally, the third section asks you to comment more generally about your experience within the course. Please answer these questions in the most constructive and objective way possible.

**I. Background Information**

1. Of all classes and other sessions scheduled for this course (e.g., labs, tutorials, etc.), I attended approximately:

| 90% or more | 70-89% | 50-69% | 20-49% | less than 20% |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

2. My anticipated grade in this course is:

| A | B | C | D | F |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

The next three statements concern your involvement with the course. You will rate the instructor's teaching in the remaining statements. Please respond using the following scale.

**1=Strongly Agree—**this statement definitely reflects my experience in all cases
**2=Moderately Agree—**this statement reflects my experience most of the time
**3=Slightly Agree—**this statement reflects my experience in some cases but not the majority
**4=Slightly Disagree—**this statement differs somewhat from my experience

**5=Moderately Disagree—**this statement in general does not reflect my experience
**6=Strongly Disagree—**this statement definitely does not reflect my experience in any way

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| **Strongly Agree** | **Moderately Agree** | **Slightly Agree** | **Slightly Disagree** | **Moderately Disagree** | **Strongly Disagree** |

3. I asked the instructor for additional guidance or feedback when I needed it.
4. I came to class prepared (e.g., reviewed posted notes, read from the course text or completed other activities as directed by the instructor) even if it was not going to be graded.
5. I think that the instructor's main role is to explain all the course content, not to make students think about it.

**II. Ratings of Teaching in this course**

6. I think the instructor was well prepared for class.
7. I think the class sessions were well organized.
8. I clearly understood the relevance of the assignments to the course objectives.
9. I think the evaluation (all graded material) clearly reflected the course content.
10. I think the course content was well organized.
11. I clearly understood what I was expected to learn in this course.
12. The time I spent in class helped my understanding of difficult course content.
13. Examples and illustrations provided in this course aided my understanding.
14. I think the instructor communicated the course material clearly.
15. I think the instructor delivered the course material at a pace I could follow.
16. I clearly understood how my work would be evaluated in this course.
17. I learned skills in this course that I will be able to use in other courses.
18. I learned ways of reasoning that I could apply to other subjects.
19. I think the instructor made the course content relevant to my overall education
20. The instructor helped me understand the relevance of the material to the real world.
21. I felt the instructor presented the course material in a way that challenged me to think.
22. I think the instructor was enthusiastic about the course content.
23. I felt comfortable participating in class activities.
24. My experience in the class increased my interest in the course content.
25. I was engaged in learning the course content during class time
26. My interactions with the instructor encouraged me to learn.
27. I think the instructor was approachable.
28. The class atmosphere supported my learning.
29. I was treated with respect in this class.
30. I felt encouraged to ask questions in class.
31. I think that the instructor was receptive to suggestions from students.
32. I was satisfied with the time it took for the instructor to return graded material.
33. The instructor provided me with all the information I needed to seek help.
34. I felt welcome to seek help from the instructor.
35. I think the instructor made a genuine effort to be available outside of class.
36. I think the instructor cared about my learning.

37. The feedback I received on work that I completed was helpful to my learning.

**III. Additional Background Information**

**1.** My program of study is: _____

**2.** My year in **this program** of study is: _____

3. My reasons for taking the course are **(check all that are applicable)**:
   - ○ Interest
   - ○ Program requirement
   - ○ Program elective
   - ○ Reputation of the instructor
   - ○ Reputation of the course
   - ○ Course fit in my timetable

**IV. Knowing what you know now about the course, if it were possible to turn back time and you could experience this course again….**

1. What changes would you make in your own approach in order to improve your learning?

2. What aspects of the course would you advise your instructor to retain?

3. What suggestions would you provide to your instructor for revisions that would produce a better learning experience for you?

4. Do you have any other comments about your learning experience in this class?