Student Evaluation of Teaching

Response Rates

April 15, 2010

Table of Contents

Student Evaluation of Teaching Response Rates (Summary Document) 1

Executive Summary 3

Summary of the Literature 4

Studies Performed at UBC 4

UBC's Response Rates 7

Appendix A. Summary of the Literature for Online Student Evaluations of Teaching (Rawn) 8

Summary 9

Typical Response Rates for Online Student Evaluations of Teaching 9

Methods to Increase Response Rates for Online SEoTs 12

Appendix B. Online Administration of the University Module Items in the Faculty of Arts and Faculty of Science (Hakstian) 20

Response Rates: Before and After the Change to Online Administration 21

Long Term Stability of UMI Mean Scores 25

A Brief Look at Changes in Mean UMI Scores over Time 29

Appendix C. Effects on Average UMI Ratings of Online vs. Paper Administration: The 2008-09 Results for the Faculties of Arts and Science (Hakstian) 32

Findings in the Faculty of Arts 34

Findings in the Faculty of Science 40

A Brief Look at Item 6 Stability 44

Appendix D. Report to on the New University-Module Items and their Online Administration at the University of British Columbia (Hakstian) 48

(original pagination retained)

Student Evaluations of Teaching Response Rates

Executive Summary

In May 2007, the UBC Senate passed a policy on Student Evaluation of Teaching. The Policy stipulates that, with limited exceptions, Student Evaluations of Teaching should be administered in every course section (or learning experience) at UBC each time the section is offered. A committee chaired by Dr. Anna Kindler, Vice Provost and Associate Vice President Academic Affairs, provides strategic oversight of the implementation of the Policy.

In alignment with the policy, the end of term surveys contain 6 common items across the institution; these items are referred to as the University Module Items (UMIs). The UMI results are stored in a centrally managed, secure database and are available to authorized University and Faculty administrators. Individual faculty members may opt to share their UMI results with students and other members of the UBC community through a password protected (CWL-enabled) website.

In order to support the implementation of the policy, UBC implemented an online system to conduct end of term surveys. Students are sent emails when the surveys are available, and invited to login and complete the surveys within a defined time period (typically the last two weeks of the term). The survey system and all data are housed on secure servers managed by UBC IT. All users authenticate using CWL.

From the outset of the policy's implementation, the committee took an empirical approach, evaluating the data collected. Dr. Ralph Hakstian, a psychometric expert from the Department of Psychology, is contracted for this purpose. The research reported below was completed by Dr. Hakstian, or under his supervision, by graduate students supporting him in this endeavour. Highlights from the research documents are noted below. The full reports are attached to this document. This research shows that although response rates are lower, there is no meaningful difference in ratings, and suggests that online evaluations reduce some forms of bias.

Summary of the Literature (Rawn, 2008)

- Response rates for online student evaluations of teaching typically range from 30 to 60%.
- In general, published studies show that response rates for online evaluations are 20 to 30% lower than those for paper based evaluations.
- In some studies, online response rates are shown to increase over time.
- If a representative sample of students complete online evaluations, then confidence intervals of 80% are sufficient (see McGill's acceptable response rates below in Table 1).

Table 1. Recommended Response Rates

Class size	McGill: Acceptable Response Rate (%)	Nulty (2008): Recommended Response Rates with 80% Confidence Interval (10% Sampling Error)
5-11	minimum 5 responses	at least 75%
12-30	at least 40%	74-48%
31-100	at least 35%	47-21%
101-200	at least 30%	20- 12%
201-1000	at least 25%	11-3%

(Adapted from Rawn, p. 11).

- Despite reduced response rates, research consistently indicates that there are no meaningful differences in ratings of instructors.
- Studies indicate that students responding online are more likely to provide qualitative comments.
- An increase in response rates can be achieved by: 1) careful timing of email reminders; and 2) instructors indicating the value of the feedback to their students.

Studies Performed at UBC

1. Hakstian (March 2008):

- Overall there are no significant differences between response rates for online and paper-based administrations as shown in Table 2 (see next page).

Table 2. Response Rates: Online vs Paper Administrations.

Admin. Unit (No. of Classes)	Mean Response Rate Online (%) Fall 2007	Mean Response Rate Paper (%) 2006-07
Psychology (41 Online; 99 Paper)	63.30	63.31
Land and Food Systems (52 Online; 32 Paper)[a]	60.98	79.24
Law (104 Online; 82 Paper)	65.18	65.91
Science (359 Online; 624 Paper)	66.52	66.40
Weighted Mean Response Rates (556 Online; 837 Paper)	65.51	66.48

[a] Some of the discrepancy between the two response rates in this case is due to the fact that students who were not registered for the courses (or were registered in cross-listed sections) were able to complete paper evaluations, but only those actually registered for the courses (and were LFS students) completed the online forms.

(Table 5 on p. 27 in Hakstian 2008)

- On a within Faculty basis, only the Faculty of Land and Food Systems showed a significant difference, but paper-based evaluations showed a higher than 100% response rate in some cases.

2. Hakstian (March 2010):

- The Faculty of Arts shows higher drop in response rates than Science. Students evaluating courses in the Faculty of Arts complete almost twice the online evaluations; students are asked to evaluate both instructors and TAs. Results are shown in Table 3 below.

Table 3. Response Rates: Online vs Paper Administrations for Arts & Science.

Faculty	Administration Format	No. of Sections	Students Responding (%)
ARTS	Pencil/Paper 2007-08, Terms 1 & 2	1898	75.97
	Online 2008-09 Terms 1 & 2	2691	59.70
SCIENCE	Pencil/Paper 2006-07, Terms 1 & 2	783	66.25
	Online 2008-09 Terms 1 & 2	1342	62.90

(Table 1 on p. 22 in Hakstian 2010)

- Response rates greater than 100% on paper-based evaluations were adjusted to 100%. The fact that there were more responses than registrations introduces a bias in favour of paper administrations; there is no way of knowing how much.

- Correlations between response rates and UMI ratings are low; online evaluations reduced those correlations uniformly and significantly.
- Negative correlations between class size and UMI ratings in paper-based administrations are reduced significantly with online administration.
- In order to make sense of any differences between paper-based and online evaluations, the year to year variability was examined.
- As shown in Table 4, the variability between administrations is lower in online vs online comparisons than paper-based vs online.

Table 4. Changes in UMI 6 Over Time & Administration.

Faculty of Arts - UMI6 Means

	1st Administration	2nd Administration	Change
Cohort 1: (*n* = 317) Paper, 2007 – Online, 2008	4.298	4.137	.161
Cohort 2: (*n* = 398) Online, 2008 – Online, 2009	4.096	4.134	–.038

Faculty of Science - UMI6 Means

	1st Administration	2nd Administration	Change
Cohort 1: (*n* = 124) Paper, 2006 – Online, 2008	4.088	4.022	.066
Cohort 2: (*n* = 196) Online, 2008 – Online, 2009	4.068	4.068	.000

(Please note in these analyses the same course/instructor results are compared across administrations)

(Table 5 on p. 31 in Hakstian 2010)

3. Hakstian (November 2009)

- The shift from paper to online in matched course/instructor comparisons across administrations has had a relatively small effect on UMI means (ranging from -.085 to -.177) in the Faculties of Arts and Science.

UBC's response rates

- Compared with other institutions, UBC's current online response rates rank relatively high as shown in Table 5.

Table 5. Response Rates for Online Administration

FACULTY	2007W	2008S	2008W	2009S	2009W[1]
Applied Science					
Architecture			35%		
Engineering			58%	42%	57%
Landscape Arch.			56%	57%	65%
Nursing		35%	69%	61%	41%
Arts	60%[2]	44%	56%	42%	57%
Dentistry	66%		70%		77%
Education[3]	63%	48%	57%	64%	61%
College of Health					
Disciplines	50%		54%		54%
Law	56%		45%		43%
Land & Food Systems	58%	48%	37%	42%	60%
Medicine[4]					53%
Occupational Therapy	74%				68%
OLT – Distance Ed.[5]	47%	46%	55%		62%
Pharmacy	53%		45%		47%
Physical Therapy	68%	48%	55%		
Rehab Grad Program	84%		92%		79%
Science	61%	58%	62%	50%	62%
School of Environ					
Health			60%		26%[6]
School of Population					
Health			65%	56%	66%

1 Term 1 and Distance Education courses ending in December 2009.
2 In 2007W, Psychology, Music and the Masters of Creative Writing had online evaluations.
3 Includes Masters of Educational Technology courses.
4 Includes courses evaluated by Faculty of Science. In previous sessions, those courses were included in the Science calculations.
5 Includes several courses for which there is no online evaluation presence (all other distance courses are represented in Faculty data).
6 Small number of courses evaluated; all courses with low n's.

Appendix A.

Summary of the Literature

For Online Student Evaluations of Teaching

Catherine D. Rawn
Department of Psychology
September 1, 2008

Summary of the Literature on Response Rates for Online Student Evaluations of Teaching

Summary

The purpose of this report is to compile typical response rates for online student evaluations of teaching (SEoTs) and to illuminate research-based strategies to improve those rates. In published data, response rates for online SEoTs typically range from 30% to 60%. A number of strategies may improve these rates, particularly when used in tandem (Nulty, 2008; Porter, 2004; Umbach, 2004). Based on the literature, the following strategies are recommended: (1) contacting students at least three times about the survey via email from the institution, including a deadline and a personal plea for help in evaluating teaching, and (2) strongly encouraging individual faculty members to remind students to complete the evaluations, including an explanation of the value of students' feedback. A lottery incentive may be a viable route to improving response rates, particularly if many relatively small prizes are offered. However, extrinsic incentives may undermine any positive effect of communicating to students the value and uses of their feedback, and therefore should be considered with caution.

Typical Response Rates for Online Student Evaluations of Teaching

Overall response rates for online SEoTs are often lower than response rates for paper-based SEoTs (e.g., Avery, Bryant, Mathios, Kang, & Bell, 2006; Dommeyer, Baum, & Hanna, 2002; Hardy, 2003; Johnson, 2003; Laubsch, 2006; Layne, DeCristoforo, & McGinty, 1999). In these studies, response rates for online SEoTs are generally 20-30% lower than for paper SEoTs, which range from a 60% to over 90% response rate. However, not all studies report a difference in response rates. Leung and Kember (2005) report no differences between online and paper response rates, which all hovered around 50% (except in Engineering, where the online response rate was significantly higher, at 60%). A study of the 2003 National Survey on Student Engagement (NSSE) in the United States revealed a similar average response rate (around 40%) per school and range (between 5% and 80%) regardless of paper or online administration

(Porter & Umbach, 2006). A study among computer science students revealed a benefit to online over paper administration: students were much more likely to complete online SEoTs (82%) in 2002 and 2003 compared to earlier paper versions of the same instrument (67%; Liegle & McDonald, 2005). Another aspect of response rates to consider is the frequency of written qualitative comments. Studies consistently show that students responding online are more likely to provide qualitative comments about their course and instructor than are those responding on paper (e.g., Hardy, 2003; Johnson, 2003; Layne et al., 1999). Thus, although overall response rates may be lower for online than paper based SEoTs, qualitative feedback can be enhanced in the former method.

Do response rates change over time?

Studies of response rates tend to report pilot data often within a single department at one institution, which may not reflect the long-term response rates of online SEoTs. However, some published reports offer data beyond a pilot project. In both cases, response rates for online SEoTs increased over time. An extensive study at Cornell University's economics-based Public Policy Department conducted across four academic years (1999-2001) examined differences in online versus paper-based SEoTs in the same courses taught repeatedly by experienced faculty (Avery et al., 2006). Twenty sections were evaluated using paper measures: the mean response rate was 77.75% of students (response rate range: 50.2-100%; class size mean = 98, range: 22-325). Nine sections were evaluated using online measures: the mean response rate was 49.83% of students (response rate range: 33.3-77.8%; class size mean = 120, range: 36-151). Over time, as more courses used online SEoTs, response rates improved. By 2003 the online mean response rate was 72% with at least a 50% response rate in all courses. Johnson (2003) also reported an increasing trend over time as a pilot program testing online SEoTs was extended across the Brigham Young University campus. An early pilot test in 1999 had a 50% response rate for online SEoTs and 71% for paper. However, later campus-wide response rates for online SEoTs increased to approximately 60% in the three subsequent years. Ballantyne (2003) also reported a low response rate initially (30% in a 1999 Engineering Department pilot), that increased to 50-70% in the following three years after implementing some promotion strategies (see the next section for more on strategies).

What is an acceptable response rate?

Nulty (2008) draws from statistical sampling theory and offers some guidance about what constitutes acceptable response rates for SEoTs (see table below). The key factors he highlights are class size, degree of sampling error deemed acceptable, and the degree to which the respondents represent the class as a whole. (Sample representativeness is an important issue

that falls outside the purview of this report; see Kreiter & Lakshman, 2005 for a selective sampling strategy; see Nulty, 2008 for a discussion.) According to Nulty, if liberal sampling conditions are deemed satisfactory (e.g., 80% confidence level, 10% sampling error), classes with 20 to 200 students need response rates of no more than 60%, provided those responses are representative of the full class. In contrast, very high response rates are required to satisfy statistically conservative sampling conditions (e.g., 95% confidence level, 3% sampling error). In practice, response rates for liberal conditions are reasonable. For example, as noted below, McGill University (see *Policy*) advocates rates closer to the liberal conditions recommended by Nulty.

Class size	McGill: Acceptable Response rate (%)	Nulty (2008): Recommended Response Rates under *Liberal* conditions	Nulty (2008): Recommended Response Rates under *Conservative* conditions
5-11	minimum 5 responses	at least 75%	100%
12-30	at least 40%	74-48%	99-96%
31-100	at least 35%	47-21%	95- 87%
101-200	at least 30%	20- 12%	86-77%
201-1000	at least 25%	11-3%	76-41%

Additional information when considering response rates.

Despite the generally relatively lower response rates for online SEoTs, studies consistently report no meaningful differences in instructor's scores between online and paper versions (e.g., Avery et al., 2006; Dommeyer, Baum, Hanna, & Chapman, 2004; Hardy, 2003; Leung & Kember, 2005).

The vast majority of literature on this topic collected data in the late 1990s and early 2000s. Some authors note that limited access to computers may have contributed to the low online response rates (e.g., Johnson, 2003). Access to computers has increased greatly since then. Statistics Canada reports that Vancouver had one of the highest home internet use rates in the country between 2005 and 2007: 78% of homes had some form of internet access (see *Canadian)*. It is likely that the limited internet access that may have reduced some of the low

response rates in early studies is no longer an issue today, given the improvement in computer access among students.

Methods to Increase Response Rates for Online SEoTs

Increasing response rates for online SEoTs is a common concern (Sorenson & Reiner, 2003). Numerous studies offer insight into strategies to improve response rates. No studies were found that tested the effect of timing in the term; all studies that noted timing were conducted in the final weeks of the term. What follows is a synthesis of literature on the four major means of increasing response rates found in the literature: multiple contacts, wording of contacts, faculty promotion, and external incentives. Studies specifically investigating SEoTs are presented whenever possible; research on other types of surveys is included when pertinent.

Contacting Participants Multiple Times

One of the clearest and most cost-effective ways to improve response rates for online surveys is to send multiple e-mails. A meta-analysis of strategies across web-based surveys in general indicated that number of contacts was the strongest predictor of response rates (r = .44), stronger than 14 other factors (Cook, Heath, & Thompson, 2000). Of the 68 surveys included in analyses, those with neither pre- nor follow-up contacts with participants yielded an average 30% response rate. Three contacts (e.g., one pre-notification and two follow-ups) yielded the highest average response rate across studies: over 50%.

Most responses—up to 80% of the final number—occur in the first five days after releasing a web-based survey (Crawford, Couper, & Lamais, 2001). To maximize the number of responses in a student survey, one study suggests that two follow-up emails, spaced 2 days and 4 days from the initial notification, is a more effective strategy than sending a single reminder email after 5 days (Crawford et al.). The two message follow-up method resulted in 80% of responses collected by day five, whereas the single message follow-up method resulted in 60% of responses collected in that time period. Ultimately, the two message follow-up method offered a small advantage over the single message follow-up method; final response rates were 26% and 21%, respectively. (Note that the low overall response rates in this study were likely due to the timing of the survey; it was live from 29 November 1999 to 11 January 2000.)

Drawing from research on paper and general web surveys, Porter (2004) recommends four contacts with potential respondents (see also Umbach, 2004). First, he recommends sending a pre-notification email so that potential respondents know to watch for the survey in their

inboxes. After sending the opening link to the survey, with a cover message, he recommends at least two follow-up contacts to remind participants to complete the survey. Considering the results of Crawford et al. (2001), these follow-up contacts should occur two and four days after the survey.

Wording of Contacts: Personalization and Special Requests for Participation

A second way to increase response rates to web surveys is to personalize the contact emails sent to respondents. In Cook et al.'s (2000) meta-analysis, sending personalized contacts was a strong predictor of response rates (r = .41), second only to number of contacts. However, a survey of high school students found that personalization had little effect on response rates (Porter & Whitcomb, 2003). Instead, response rates were improved by 8 percent when students were given a survey shutdown deadline *and* when they were told they were part of a specially selected group of informants selected randomly to participate. Other research on SEoTs shows that students believe their opinions do not matter to instructors and administrators (Spencer & Schmelkin, 2002). Adding a special note of invitation personally asking for help (Porter, 2004) may help students believe their responses will matter, thereby increasing response rates.

Faculty Support and Encouragement

Response rates improve when individual instructors encourage students to complete online SEoTs (Ballantyne, 2003; Norris & Conn, 2005). In two reports, reduced response rates for online SEoTs after pilot testing have been attributed to a concurrent reduction in faculty encouragement. Goodman and Campbell (1999) report a drop from 52% to 27% in their online SEoT response rates from semester 1 to semester 2 in 1998 at an Australian Computer Science department. In their report, the authors surmise that a relative dearth of promotional material by faculty and administration in the second semester may explain the drop. In another study, faculty volunteered for their classes to take part in a pilot test of online rather than paper SEoTs (Anderson, Brown, & Spaeth, 2006). The overall response rate in this pilot test was over 80%. The following year the whole campus adopted online SEoTs and the overall response rate was cut in half. Notably, individual faculty had higher response rates than average if their department head was involved in the administration of the online SEoT pilot project, if the individuals were active on teaching and learning committees, and if they participated in teaching development activities themselves.

Results of these cases suggest that online SEoT response rates may depend importantly on promotion by individual faculty members. Further research substantiates this observation. After an initial disappointing response rate of 30% in 1999, faculty in the Engineering

Department at Murdoch University (Australia) actively promoted the survey, explained how they planned to use the students' feedback, and held a prize draw (Ballantyne, 2003). After one semester, the average response rate rose to 54%, and after two semesters, it rose further to 72%. After two years the average response rate fell to 50%, perhaps indicative of lowered enthusiasm and promotion over time. In another study, instructors who announced online SEoTs in their syllabi, by email, by a link on their course web pages, and who also sent students a reminder email were rated with much greater frequency than were instructors who did not use these strategies (Norris & Conn, 2005). These strategies increased response rates in face-to-face courses from an average of 34% to 67%, and in web-based courses from 42% to 74%. Importantly, these strategies had a large impact on courses that were receiving extremely few evaluations; the minimum response rate in the strategy group was 41%, compared to zero in the no strategy group.

The strategies recommended by Norris and Conn (2005) dovetail with those mentioned earlier regarding sending announcement and follow-up emails. I know of no study that has empirically tested the differential impact of emails about online SEoTs sent from individual instructors versus institutions. It is possible that Norris and Conn's drastic increase in response rates (by 30%) is related to students perceiving that individual faculty members value their responses. As mentioned earlier, students tend to feel like their opinions on teaching are not valued highly (Spencer & Schmelkin, 2002). When people do not feel a survey is important, they tend to not respond to it (Porter, 2004). Faculty promotion of SEoTs may indicate to students that their opinions do, in fact, matter. Students in focus groups have mentioned that such faculty encouragement is one key strategy for improving response rates for online SEoTs (Johnson, 2003). The impact of showing students their feedback is valued can be extreme. In one department where SEoTs are regularly used to effect reflection and change, and where such changes are communicated regularly to students via reports made by individual instructors, the average response rate is over 95% (Tucker, Jones, & Straker, 2008).

To synthesize results of these studies, a most effective approach may be to send pre-notification and reminder emails from the institution, as well as to encourage individual faculty members to take responsibility for promoting SEoTs in their courses (Nulty, 2008). They can promote SEoTs relatively easily by announcing them in class, on syllabi, and on course webpages, as well as by explaining how they intend to use the results. This joint-responsibility approach is underway at McGill University (see *Teaching*).

Possible effect of faculty encouragement on response rates in other courses.

One issue that arises when discussing response rates is whether students will, once logged into a SEoT website, rate all their courses or simply one or two. In a large pilot study at Brigham Young University, the overall response rate was 62% (Johnson, 2003). Of the students who rated at least one course (n=1892), only 34% of them (n=638) rated *all* of their courses. However, students who were enrolled in more than one course involved in the pilot project (i.e., which had special announcements and reminders) were more likely to rate all their courses than those who were enrolled in only one pilot project course. Forty percent of students enrolled in two pilot courses rated all of their courses, and 74% of students enrolled in three pilot courses rated all of their courses. This result suggests that a teamwork approach in which all faculty members promote completion of online SEoTs may lead to the highest response rates across the institution.

Use of Incentives

Research on the effect of incentives (e.g., lotteries, small grade increases) on response rates to SEoTs and other online questionnaires reveals mixed results (Umbach, 2004). One study has compared the effect of three different incentives on response rates to SEoTs (Dommeyer et al., 2004). Researchers recruited 16 instructors from the business school at California State University, Northridge. Each instructor was teaching two sections of a course; one was randomly assigned to provide a paper SEoT, and the other an online version of the same questionnaire. Courses with online SEoTs were assigned to one of four conditions: no incentive, a one quarter percent grade increase, feedback about grades before the final exam if 2/3 of the class completed the evaluation, or an in-class demonstration of how to complete the questionnaire. The average response rate for paper SEoTs was 75%. The average response rate for online SEoTs with no incentives was 28%. The four courses in which students received a negligible grade increase had an average 87.5% response rate. The two courses in which students received a demonstration or feedback averaged response rates of 53% and 51%, respectively. Based on this study, adding a very small grade increase may greatly improve response rates to online SEoTs. This result is consistent with Johnson (2003), who reports an average response rate of 87% among instructors who offered points for online SEoT completion. However, others have questioned the ethics of using grade incentives in this way (Ballantyne, 2003).

Lotteries for prizes have worked to improve response rates at some institutions (e.g., Ballantyne, 2003). However, most research on the effect of lotteries has examined response rates to surveys other than SEoTs, and results are mixed. In a large survey of American high

school students, researchers offered prizes ranging from $50 to $200 in cash (Porter & Whitcomb, 2003). None of these prizes yielded more than a 2.3% boost to response rates relative to the control group. A survey among first year students at a Belgian university experienced an insignificant 3% improvement in response rate resulting from a lottery for ten €25 gift certificates, relative to the no-incentive control group (Heerwegh, 2006). However, the lottery also decreased the break-off rate (i.e., failure to complete the survey) by 2.5%, resulting in an overall benefit of nearly 6% more complete surveys in the lottery condition relative to the control condition. In contrast, a large online survey of real estate agents experienced almost twice the number of complete responses in a lottery condition (23.4%) relative to the control condition (12.9%; Bosnjak & Tuten, 2003).

If a lottery is to be used, one study has tested the effect of multiple smaller prizes (ten €25 gift certificates), a few larger prizes (five €50 gift certificates), or one large prize (a DVD player) on response rates to an online survey in a student population (Deutskens, de Ruyter, Wetzels, & Oosterveld, 2004). Smaller prizes with a higher probability of winning yielded a significantly greater response rate (35.5%) than did few larger prizes (26%) and a single large prize (22.5%). In summary, lotteries and other incentives may improve response rates to online SEoTs; however, given the discrepancies in available data, the degree of expected gain is questionable.

Additional considerations regarding use of incentives.

Research on motivation reliably finds that extrinsic rewards can undermine intrinsic attraction to a variety of activities (Deci, Koestner, & Ryan, 1999). Evaluating teaching can be viewed as an important task. When it is, response rates can soar (Tucker et al., 2008). Implementation of incentives for completing evaluations of teaching risks devaluing the process as well as potentially setting expectations for future surveys from the institution (Porter, 2004). Thus, incentives should be attempted with caution.

References

Anderson, J., Brown, G., & Spaeth, S. (2006). Online student evaluations and response rates reconsidered. [Electronic Version.] *Innovate: Journal of Online Education, 2*. Retrieved 21 August 2008 from http://www.innovateonline.info/?view=article&id=301.

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

Ballantyne, C. (2003). Online evaluations of teaching: An examination of current practice and considerations for the future. *New Directions for Teaching and Learning, 96*, 103-112.

Bosnjak, M., & Tuten, T. L. (2003). Prepaid and promised incentives in web surveys: An experiment. *Social Science Computer Review, 21,* 208-217.

Canadian internet use survey. (2008, June 12). *The Daily Statistics Canada*. Retrieved 29 August 2008 from http://www.statcan.ca/Daily/English/080612/d080612b.htm.

Cook, C., Heath, F., & Thompson, R. L. (2000). A meta-analysis of response rates in web- or internet-based surveys. *Educational and Psychological Measurement, 60*, 821-836.

Crawford, S. D., Couper, M. P., & Lamais, M. J. (2001). Web surveys: Perceptions of burden. *Social Science Computer Review, 19*, 146-162.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125*, 627-668.

Deutskens, E., de Ruyter, K., Wetzels, M., & Oosterveld, P. (2004). Response rate and response quality of internet-based surveys: An experimental study. *Marketing Letters, 15*, 21-36.

Dommeyer, C. J., Baum, P., & Hanna, R. W. (2002). College students' attitudes toward methods of collecting teaching evaluations: In class versus on-line. *Journal of Education for Business, 78*, 11-15.

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

Goodman, A., & Campbell, M. (1999). Developing appropriate administrative support for online teaching with an online course evaluation system. Retrieved 27 August 2008 from http://www.deakin.edu.au/~agoodman/isimade99.html.

Hardy, N. (2003). Online ratings: Fact and fiction. *New Directions for Teaching and Learning, 96,* 31-38.

Heerwegh, D. (2006). An investigation of the effect of lotteries on web survey response rates. *Field Methods, 18*, 205-220.

Johnson, T. D. (2003). Online student ratings: Will students respond? *New Directions for Teaching and Learning, 96*, 49-61.

Kreiter, C. D., & Lakshman, V. (2005). Investigating the use of sampling for maximising the efficiency of student-generated faculty teaching evaluations. *Medical Education, 39*, 171-175.

Laubsch, P. (2006). Online and in-person evaluations: A literature review and exploratory comparison. [Electronic version]. *Journal of Online Learning and Teaching, 2*. Retrieved 27 August 2008 from http://jolt.merlot.org/Vol2_No2_Laubsch.htm.

Layne, B. H., DeCristoforo, J. R., McGinty, D. (1999). Electronic versus traditional student ratings of instruction. *Research in Higher Education, 40,* 221-232.

Leung, D. Y. P., & Kember, D. (2005). Comparability of data gathered from evaluation questionnaires on paper and through the internet. *Research in Higher Education, 46,* 571-591.

Norris, J., & Conn, C. (2005). Investigating strategies for increasing student response rates to online-delivered course evaluations. *The Quarterly Review of Distance Education, 6*, 13-29.

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment and Evaluation in Higher Education, 33*, 301-314.

Policy on official end-of-term course evaluations. (23 January 2008). *McGill University*. Retrieved 28 August 2008 from http://www.mcgill.ca/tls/courseevaluations/policy/.

Porter, S. R. (2004). Raising response rates: What works? *New Directions for Institutional Research, 121*, 5-21.

Porter, S. R., & Umbach, P. D. (2006). Student survey response rates across institutions: Why do they vary? *Research in Higher Education, 47*, 229-247

Porter, S. R., & Whitcomb, M. E. (2003). The impact of lottery incentives on student survey response rates. *Research in Higher Education, 44*, 389-407.

Teaching and Learning Services, McGill University. (2008). *Strategies to increase online response rate*. Retrieved 30 August 2008 from http://www.mcgill.ca/tls/courseevaluations/strategies/

Tucker, B., Jones, S., & Straker, L. (2008). Online student evaluation improves Course Experience Questionnaire results in a physiotherapy program. *Higher Education Research and Development, 27,* 281-296.

Umbach, P. D. (2004). Web surveys: Best practices. *New Directions for Institutional Research, 121*, 23-38.

Appendix B.
Online Administration of the University Module Items

In the Faculty of Arts and Faculty of Science

Ralph Hakstian
Department of Psychology
March 13, 2010

Online Administration of the University Module Items (UMIs) in the Faculty of Arts and Faculty of Science

Response Rates: Before and After the Change to Online Administration

On the basis of one full year of student responses to the UBC University Module Items (UMIs) administered via both pencil-and-paper and online formats, we can now compare response rates (*defined as the percentage of students in a class that responded to the student-evaluation inventory containing the UMIs*) between the two administration formats. First, for the Faculty of Arts, we have the student responses to: (a) the full 2007-08 academic year *pencil-and-paper* (referred to in the sequel as simply "*paper*") administration of the UMIs (included with the Arts inventory) and (b) the full 2008-09 academic year online administration of the Arts inventory and the UMIs. Second, for the Faculty of Science, we have the student responses to: (a) the full 2006-07 academic year paper administration of the Science inventory and (b) the full 2008-09 academic year online administration of the UMIs.

Potential Bias in the Paper Results

In our examination of the paper teaching-evaluation results from previous years, we have discovered a number of sections with a greater-than-100% reported response rate. This, of course, signifies an impossible occurrence. In all of our present analyses, we have set the student response rate for such classes to 100%, and this is reflected in all the results reported in the following tables. This adjustment, however, obviously cannot correct all the bias in the results. For these sections for which the response rate has been set to 100%, this adjustment will almost certainly still result in an overestimate of the actual response rate. Further, we have no way of knowing how many of the reported response rates that are less than 100% are accurate or, perhaps, similarly inflated. For this reason, we must interpret the following results with caution.

Recorded Student Response Rates for Arts and Science

In Table 1 below, the recorded student response rates for the two faculties appear (with the noted adjustment to the paper results).

Table 1. Student Response Rates for Paper and for Online Administration

Faculty	Administration Format	No. of Sections	Students Responding (%)
ARTS	Paper 2007-08, Terms 1 & 2	1898	75.97
	Online 2008-09 Terms 1 & 2	2691	59.70
SCIENCE	Paper 2006-07, Terms 1 & 2	783	66.25
	Online 2008-09 Terms 1 & 2	1342	62.90

It seems clear from Table 1 that there has been some reduction in response rate in going from the paper administration format to the online. Given the inflation noted above, however, in the paper results, we cannot be sure of just how much reduction has actually occurred. On the face of it, we seem to have about a 16% drop in the Faculty of Arts, but only about a 3% drop in the Faculty of Science. Given the uncertainty surrounding our paper data, we are reluctant to attempt much of an interpretation of these apparent reductions in response rates.

It might be of interest to know that, compared with the student response rates achieved at many U.S. universities with online student evaluation of teaching, our current rates of 60+% rank near the top. Many of these universities are seeing response rates in the 30% - 40% range, although there is evidence that, with some creative administration methods, this need not be the case, and much higher response rates are possible. We (the Student Evaluation of Teaching—SEoT—Implementation Committee) consider the 60+% response rate at UBC reasonable, but are reviewing and considering methods by which this percentage can be raised.

Relationships between Student Response Rate and UMI Scores and Class Size

In an effort to determine whether or not student response rate changes might affect UMI scores obtained by individual instructors, we correlated the response rate for each instructor/course unit with the mean UMI scores obtained for the unit, along with the size of the class. We remind the reader that, as with all of our previous analyses of teaching-evaluation data, the *instructor/course unit* is the unit of analysis. Thus, by "score" here, we are referring to the *mean* score—over the responses of all the students in the class—obtained by

an instructor on a particular UMI. Thus, if there were to be a large correlation between student response rate and a particular UMI score, this would indicate that classes in which the response rate was high tended to be those in which the highest UMI mean scores were obtained.

Similarly, we were interested in determining whether student response rate was related to class size. In the absence of any empirical results, we might, for example, think that smaller classes tend to have a higher response rate from the students when courses are being evaluated.

Finally, we were interested in determining whether these relationships (whatever they turned out to be) were constant over administration modalities. In other words, was the relationship between student response rate and, say, UMI mean scores different when comparing paper results with those obtained online?

These questions were answered by the results appearing below in Table 2.

Table 2. Correlations between Student Response Rate, UMIs and Class Size

Faculty	Administration Format (sections)	Mean Scores						Class Size
		UMI1	UMI2	UMI3	UMI4	UMI5	UMI6	
ARTS	Paper 2007-08, Terms 1 & 2 (1898)	.182	.188	.183	.155	.231	.189	–.439
	Online 2008-09 Terms 1 & 2 (2691)	.066	.092	.109	.095	.115	.092	-.173
SCIENCE	Paper 2006-07, Terms 1 & 2 (783)	--- UMIs not used ---						-.402
	Online 2008-09 Terms 1 & 2 (1342)	.086	.050	.060	.099	.120	.062	-.155

Note: All differences between correlations from one year to the other are statistically significant (for UMIs 3 and 4, $p < .05$; for all other comparisons, $p < .01$).

The results in Table 2 are interesting. First, we see, from the results for the Faculty of Arts, that the correlations between student response rate and mean scores on the UMIs are low. However, the change from paper to online administration reduced these correlations uniformly and significantly. What little tendency there was, with the paper format, for classes having the largest student response rates to also be those in which the instructor was more-highly rated (correlations averaging around .19) has been almost eliminated with the online administration (correlations averaging around .095 for Arts and .08 for Science). As noted, the reduction in this relationship (in the Arts data) is statistically significant.

What this means, with the online administration format, is that there is almost no relationship at all between the student-response rate for a class and the mean UMI scores. In general, this is good news. The change to the online format has almost eliminated an interpretively-extraneous variable from the UMI mean scores. At the same time that we are exploring ways to increase student response rates, it is comforting to know that, at the very least, differences in these rates are having next to no effect on the mean UMI scores that instructors receive from their classes.

The second interesting result in Table 2 is the correlation between student response rate and class size under the two administration formats. It can be seen that, formerly, with the paper format, there was a fairly substantial correlation between percentage of students in the class responding to the student evaluation inventories and the class size: –.439 for Arts and –.402 for Science (the two correlations not significantly different). The fact that these correlations are negative indicates, of course, that the largest classes tend to have the lowest student response rates, and the smallest classes, the highest response rates. With the change to online administration, this relationship is greatly reduced. Now, instead of correlations in the –.42 range, we have correlations in the –.165 range: –.173 for Arts and –.155 (again these two not significantly different from each other) with the online format. This reduction in the strength of relationship of, on average, .255 is statistically significant and large. Thus, although these reduced correlations do not reflect a complete absence of relationship between student response rate and class size, the magnitude of the effect of this extraneous variable (for interpretive purposes) has, via the change to online administration, been substantially reduced to the point of being almost irrelevant. As with the reduction in the relationships between student response rates and mean UMI scores, this result with class size is a good outcome. It means that the results received by instructors of large classes are based on nearly the same *percentage* of student respondents as for those instructors of smaller classes.

Summary of Results Involving Student Response Rates

To date, we have evidence of a reduction in student response rates following the change from paper inventory administration to online administration. We cannot be certain of the actual magnitude of the reduction; the Faculty of Science results suggest that it is very small. The SEoT Implementation Committee will continue to explore ways to increase student response rates.

Other outcomes associated with the shift from paper to online administration are positive. First, it is clear that the effects of individual differences between classes in student response rate and mean UMI scores received by instructors have been significantly reduced. This means that, although we will continue to do whatever we can to raise these response rates, instructors need not feel that their obtained UMI results are the result of low response rates and that these results are, therefore, an unfair indication of their teaching performance. Second, the negative effects of large class size on student response rates appear to have been substantially reduced with the change to online administration. This means that instructors of very large classes will no longer be faced with substantially lower rates at which their students respond to the UMI inventory. There still remains a low relationship between these two variables, but it is now almost irrelevant.

Long-Term Stability of UMI Mean Scores Resulting from Online Administration

In order to assess the long-term (defined here as one-year) stability of the UMI mean scores obtained through online administration, we assembled a sample of instructor/course units that were identical in the Fall Term, 2008 and Fall Term, 2009. That is, we isolated 398 sections in the Faculty of Arts (185 taught by women and 213 by men), and 196 sections in the Faculty of Science (47 taught by women and 149 by men) in which the same course was taught by the same instructor in the two fall terms, separated by one year. The purpose was to examine the extent to which the UMI-score rank-ordering remained constant from one year to the next. We would, for example, have some serious concerns about our UMI measurements if there were little or no relationship from year to year between score ranks achieved by the same instructors. On the other hand, we must be mindful of the typical fluctuations that occur over time with virtually all performance measures.

To refresh the reader's memory of the content of the six UMIs, this is given on the next page.

Table 3. University Module Items.

UMI 1:	The instructor made it clear what students were expected to learn.
UMI 2:	The instructor communicated the subject matter effectively.
UMI 3:	The instructor helped inspire interest in learning the subject matter.
UMI 4:	Overall, evaluation of student learning (through exams, essays, presentations, etc.) was fair.
UMI 5:	The instructor showed concern for student learning.
UMI 6:	Overall, the instructor was an effective teacher.

The response scale is:

(1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree

One-Year Correlations by Faculty

Below in Table 4 are the correlations for each UMI, along with some summary means, for each faculty separately.

Table 4. One year Correlations (2008-09 Term 1 to 2009-10 Term1)

	Faculty		
UMI	**Arts (*n*=398 Sections)**	**Science (*n*=196 Sections)**	**Wt'd Average (*n*=594 Sections)**
UMI1	.56	.62	.58
UMI2	.64	.67	.65
UMI3	.70	.68	.69
UMI4	.54	.52	.53
UMI5	.68	.69	.69
UMI6	.65	.65	.65
Mean	.63	.64	.64

There are several conclusions we can draw from the results in Table 4. First, there is considerable similarity between the results obtained in Arts and those in Science (none of the Arts–Science correlational differences is statistically significant). Thus, the weighted averages in the third column give us a stable representation of the correlations for each UMI. Second, there are some interesting differences among the UMIs themselves. Most noticeably, UMI4 shows the lowest one-year correlation of the UMIs—a value that is significantly lower than the one-year correlations for all other UMIs except that for UMI1, which is also on the low side. It

is important to note, however, that beginning in Term 1, 2009-10, a *N/A*, or "Not Applicable" option was added to UMI4 only. The fact that this option was absent in the 2008-09 administration, but present in the 2009-10 edition may help explain the lower long-term correlation for this item.

Although not manifesting a significantly different one-year correlation from those of the remaining UMIs (except for UMIs 1 and 4), UMIs 3 and 5 have the highest correlations, at around .69. The all-important UMI6 shows a .65 correlation over one year. Finally, the average magnitude of these long-term correlations (close to .65) is very close to what behavioral scientists have found over the years for many variables.

We might be tempted to search for the reasons for the differences in one-year correlations noted above among the UMIs. Why, for example, should student ratings of inspiring interest in students (UMI3) or showing concern for students (UMI5) be more consistent from year to year than the ratings of setting clear expectations (UMI1) or, particularly, providing fair evaluations of student performance (UMI5)? We have seen that the *average ratings* for UMI4 are somewhat lower than those for UMIs 3 and 5, but this is a different phenomenon than the question of stability over time.

The biggest issue in connection with the results in Table 4, however, is just what to make of the numbers. What is represented by a one-year correlation between measurements on the same variable? We have found nothing in the literature that contained any stability results for measures like our UMIs—essentially measurements provided by two different sets of other people one year apart. It is true that many longitudinal studies of variable stability can be found in the behavioral literature. However, time intervals have varied widely in these studies, and many have been conducted with either young children or aging (and, in some cases, infirm) adults. Nonetheless, it is possible to get some idea of the stability of personality traits like those that make up the Big Five model (the model used almost exclusively today in the study of personality traits). A meta-analysis from 2000 suggested that one-year correlations for these variables could be expected to fall around .70, very close to the values we saw in Table 4 for the UMIs.

Still, these meta-analytic results are with reference to what are assumed to be stable personality constructs that should, according to theory, remain constant, and should manifest a high similarity in the rank-ordering of individuals over a one-year time interval. In addition, the

correlations reported in the studies compiled in the meta-analysis are for *scales* consisting of many items (say, from 8 or 10 up to 30 or 40), and it is scores on the scales that are being correlated.

These characteristics of the variables correlated in the studies noted above differ from what we have with our UMIs. First, the published stability studies have, as noted, been conducted with *scale* scores. The UMIs are, on the other hand, *single items*. It is a psychometric truism that the more items in a scale, the higher the reliability of the scale, including the retest variety considered in the studies to which we have alluded. Thus, any retest correlation— particularly with a long time interval like one year—involving a single personality or ability item would be considerably lower than those correlations obtained with scale scores. Second, the stabilities reported in the literature are for scores obtained by either (a) self-report or (b) performance on a maximum-performance measure (like an ability test). The unit of analysis is the individual responder. With the UMIs, of course, the instructor's score arises from the aggregation of ratings by *others*. Finally, although we might expect considerable similarity of mean UMI scores from year to year, we are not dealing with a behavioral phenomenon assumed to be a constant except for the fact of random measurement error. Instead, we are dealing with variables (the UMIs) the scores on which, although certainly containing measurement error, might also be expected to contain what measurement specialists refer to as *function fluctuation*, or an actual change in the underlying behavior, not in merely a measurement of it. Indeed, we would hope for such behavioral change in the cases of instructors receiving low UMI mean scores.

Further, the "error" component of UMI scores differs from that of standard self-report or maximum-performance measures (like ability tests). Where does this "error" come from in the present context? First, the classes performing the assessments at the two time points are composed of entirely different students. These students will have differing preconceived views of good teaching, and each class as a whole may differ to some extent in this respect. Thus, at Time 1, one particular aggregate of n_1 students may view Instructor X's teaching performance (when averaged) as, let's say, deserving a rating of 4.0. At Time 2, another aggregate of n_2 students may view Instructor X's ***identical*** *teaching performance* as deserving a rating of 3.94. Thus, two identical phenomena may easily be rated differently. These rating fluctuations will be random across classes, and this fact will account for changes in the rank-ordering of instructors and lower one-year correlations.

Another cause of random error is the combination of external factors that can conspire to either improve or decrease students' perceptions of a course and its instructor. The need for a

number of room changes, breakdowns in A-V equipment, extremely good or poor work on the part of the course TA(s), the existence in the class of a single student course booster or, more frequently, a single detractor or malcontent that influences the whole class to view it in ways they would not have without this single student (or several students), news that their instructor has received a prestigious award, along with many other extraneous factors, can cause different aggregated ratings for two identical teaching performances.

Along with function fluctuation—or true planned change in teaching by the instructor—we have other, unintended, instructor changes during the term due to such factors as illness, personal matters, and so on. In such cases, although the instructor is attempting to perform as s/he has in the past with respect to a course, the perception of that instructor's actual performance is diminished. On the other hand, news the instructor receives about a positive promotion or tenure decision could trigger a spurt of enthusiasm and happiness during a term that might result in slightly higher student ratings. These factors are transitory, of course, and thus add to the random error component of the UMI mean scores. In the present particular time period, we had the additional factor of the H1N1 flu phenomenon which may have affected the environments of both instructors and students, and, perhaps, lowered the consistency of UMI scores over this one-year interval.

A Brief Look at Changes in Mean UMI Scores over Time

In an earlier report, we saw some decline in mean UMI scores in the transition from paper to online administration, and we thought that it might be informative to examine and compare, in addition, a one-year change occurring with online administration alone. For this analysis, we took two cohorts of instructors from both the Faculties of Arts and Science, each cohort having taught the same course on successive occasions, either one or two years apart. In the Faculty of Arts, Cohort 1 is a group of 317 instructors that taught the same course in both the Fall term, 2007 (and had their teaching evaluated via the paper forms) and in the Fall term, 2008 (when they had their teaching evaluated by the online system). Arts Cohort 2 is a group of 398 instructors (many of whom would undoubtedly also be in Cohort 1) that taught the same course in both Fall, 2008 and Fall, 2009 and had their teaching evaluated on both occasions by the online format.

In the Faculty of Science, Cohort 1 is a group of 124 instructors who taught the same course in both the Fall term, 2006 (and had their teaching evaluated by the Science paper form, in which the summative item was almost identical to UMI6, and is treated as identical in Table 6 on the next page) and in the Fall term, 2008 (when their teaching was evaluated via the online system). Science Cohort 2 is a group of 196 instructors (many of whom would undoubtedly also be in Cohort 1) that taught the same course in both Fall, 2008 and Fall, 2009 and had their teaching evaluated on both occasions by the online survey.

The reason for using these particular groups of instructors in these analyses is to control as much as possible for factors extraneous to the intended analysis, such as comparing different instructors and courses on the two occasions which would, of course, introduce additional error into the data. Results for these analyses appear in Table 5 (Faculty of Arts) and 6 (Faculty of Science).

Table 5. Changes in UMI Over Time & Administration for the Faculty of Arts

Faculty of Arts - UMI6 Means

	1st Administration	2nd Administration	Change 2nd-1st
Cohort 1: (*n* = 317) Paper, 2007 – Online, 2008			
UMI1	4.137	4.125	-.012
UMI2	4.251	4.118	-.133
UMI3	4.128	4.096	-.032
UMI4	4.157	4.038	-.119
UMI5	4.239	4.219	-.020
UMI6	4.298	4.137	.161
Cohort 2: (*n* = 398) Online, 2008 – Online, 2009			
UMI1	4.091	4.122	.031
UMI2	4.092	4.130	.038
UMI3	4.064	4.094	.030
UMI4	3.989	4.051	.062
UMI5	4.193	4.206	.013
UMI6	4.096	4.134	.038

Table 6. Changes in UMI Over Time & Administration for the Faculty of Science

Faculty of Science - UMI6 Means

	1st Administration	2nd Administration	Change 2nd-1st
Cohort 1: (*n* = 124) Paper, 2006 – Online, 2008			
UMI6	4.088	4.022	-.066
(Other UMIs not used in 2006 administration)			
Cohort 2: (*n* = 196) Online, 2008 – Online, 2009			
UMI1	4.088	4.101	.013
UMI2	4.020	4.027	.007
UMI3	3.969	3.965	-.004
UMI4	3.903	3.884	-.019
UMI5	4.140	4.112	-.028
UMI6	4.068	4.068	.000

Please note: In these analyses the same course/instructor results are compared across administrations.

The results in Table 5 and 6 are far from dispositive, but do provide some evidence of what we might expect in the year-to-year fluctuation of UMI mean scores. First, the previously-noted decline in UMI scores in the transition from paper to online administration is seen again in these Fall term comparisons, although it is a smaller decline than we found earlier when we analyzed data gathered from a full, two-term academic year. Second, some of the decline in the Faculty of Arts results (Table 5) can be seen to be made back up in the 2008–2009 online change. The parallel 2008–2009 online change for the Faculty of Science (Table 6) is completely negligible. Interestingly, the earlier paper-to-online decline with respect to UMI6 was considerably smaller in Science (.066) than in Arts (.161) for these Fall term courses.

We will continue to monitor year-to-year mean UMI scores and examine the shifts that occur over time. This will add to our overall understanding of the performance of the UMIs at UBC.

Appendix C.

Effects on Average UMI Ratings of Online vs. Paper Administration:

The 2008-09 Results for the Faculties of Arts and Science

Ralph Hakstian
Department of Psychology
27 November 2009

Effects on Average UMI Ratings of Online vs. Paper Administration: The 2008-09 Results for the Faculties of Arts and Science

The objective of the present study was to determine whether or not there was a reliable difference in mean ratings on the six University Module Items (UMIs) between those arrived at through paper-and-pencil administration of the items to students in the classroom and those arrived at through online administration of the items to students to be filled in on their own time. In the Faculty of Arts, the UMIs had been administered in paper-and-pencil form in the 2007-08 academic year, but via online means in the 2008-09 year, thus providing a basis for comparison of the two administration modes. In the Faculty of Science, however, the last year of paper-and-pencil administration was 2006-07, and this inventory's items were different from the UMIs and had a different response scale than found with the UMIs. Thus, in this faculty, our comparison involved the single item that was similar in the two inventories—the paper-administered Science inventory in 2006-07 and online-administered UMIs in 2008-09.

It is important to realize that the shift from the paper-and-pencil-administered earlier inventories to the online-administered UMIs represents a change which may result in mean differences. Possible changes in scale means (calculated over the student ratings in a class) should not concern instructors, as the new norms represented by the UMIs administered online will apply to all instructors equally and within several years will define UBC teaching effectiveness levels (as these are perceived by students). Nonetheless, it is of interest to know whether ratings are affected by the transition from paper-and-pencil to online administration. One benefit of this knowledge will be to understand with greater insight (and possibly some adjustments) comparisons of teaching effectiveness for individual instructors as measured by the new UMIs administered online with their measured effectiveness via earlier inventories and paper-and-pencil administration.

Within both faculties, we isolated all the instructor/course combinations that were given in 2008-09 (both terms) and evaluated using the online SEoT inventory of UMIs **and** that had been given in an earlier academic year and evaluated with the pencil-and-paper (abbreviated simply to "paper" in the sequel) SEoT inventory. This (paired-comparison) design using dependent samples contributed to greater precision in the analyses since it led to a reduction in the sampling error normally associated with independent-samples comparisons. Since a number of characteristics differed between the two faculties, we report the results for each separately. After presentation of numerical results, we will attempt to synthesize the findings and identify some general conclusions that can be drawn.

Findings in the Faculty of Arts

Method

In the Faculty of Arts, we were able to isolate 707 instructor/course units that were given in 2008-09 (both terms) and evaluated using the online SEoT UMI inventory and that had also been given in 2007-08 (both terms) and evaluated with the paper version of the UMI inventory (in which the wording of the items was very slightly different). We first examined the data across all administrative units within the faculty, calculating the means and standard deviations (*SD*s) of the six UMIs and their average, over all 707 instructor/course units for both academic years. This analysis enabled us to see whether, first, there were any shifts in mean ratings—upward or downward—and, second, whether any such shifts were constant across all six UMIs or instead varied according to the item content.

Next, we examined shifts in mean ratings, as a result of the changeover to online administration, at the administrative-unit level for 22 administrative units within the Faculty of Arts. To prevent the reporting of these results from becoming too piecemeal and voluminous, we examined the differences between the two academic years on two measures: (a) a summary measure—the *average of the six UMIs* (justifiable because all six UMIs correlate substantially and positively)—and (b) the one summative item, UMI 6.

To refresh the reader's memory of the content of the six UMIs, they are given on the next page.

Table 1. University Module Items.

UMI 1:	The instructor made it clear what students were expected to learn.
UMI 2:	The instructor communicated the subject matter effectively.
UMI 3:	The instructor helped inspire interest in learning the subject matter.
UMI 4:	Overall, evaluation of student learning (through exams, essays, presentations, etc.) was fair.
UMI 5:	The instructor showed concern for student learning.
UMI 6:	Overall, the instructor was an effective teacher.

The response scale is:

(1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree

Results for the Faculty as a Whole

In Table 1, the faculty-wide means and SDs on the six UMIs are presented for the (a) 2007-08 (paper) and 2008-09 (online) administrations based on the sample of identical instructor/course combinations for the two years.

Table 2. Academic Year 2007-08 to 2008-09 (Terms 1 & 2)

UMI	Mean		Difference in Means[a]	Standard Deviation		Difference in SDs[a]
	2007-08 Paper	2008-09 Online		2007-08	2008-09	
1	4.19	4.16	-.03	.36	.43	.07
2	4.30	4.15	-.15	.40	.47	.07
3	4.17	4.11	-.06	.46	.50	.04
4	4.19	4.07	-.12	.39	.43	.04
5	4.28	4.25	-.03	.38	.42	.04
6	4.34	4.16	-.18	.41	.48	.07
Averages	4.245	4.150	**-.095**	.398	.456	**.058**

Note: These results are based on 704-707 sections (identical instructor/course combinations) given in both 2007-08 and 2008-09 (both terms).

[a] The differences are given as the 2008-09 value minus the 2007-08 value. Thus, a positive difference means that the 2008-09 value is larger than the 2007-08 value; a negative difference means the 2007-08 value is larger.

It will be noted, from Table 1, that there is virtually no change in item means for UMIs 1 and 5, and a negligible change for UMI 3, arising from the transition from paper to online administration. Using a conservative inferential criterion (to protect against excessive Type I

error), we find that the item means differ significantly between the two academic years (and administration types) for UMIs 1, 2, 3, 4, 6, and the overall average, but not for UMI 5. It is important, however, to distinguish between statistical and *practical* significance, and for most of the UMIs (with the possible exceptions of UMIs 2 and 6), it could be argued that the differences in Table 1 are of little practical significance. With a very large *n* and powerful design like this, small differences often emerge as statistically significant. We will return to a discussion of the differences in *variability* over the two academic years in a later section.

Results Involving Means for Individual Administrative Units

In Table 2 below, we present the results by administrative unit of the decrease or increase in average ratings on two variables (a) the Average of the Six UMIs and (b) UMI 6.

Table 2. Differences in Ratings on UMI Average and UMI6

Administrative Unit	Number of Paired Sections	Difference in Means[b] from 2007-08 to 2008-09 for:	
		Average of 6 UMIs	UMI 6
Art History, Visual Art & Theory	37	+.003	–.089
Anthropology	14	+.060	–.020
ARTSsm[a]	6	.000	–.115
Asian Studies	96	–.072	–.140
Central, Eastern and Northern European Studies	56	–.072	–.177
Classical, Near Eastern and Religious Studies	28	–.133	–.181
Economics	62	–.154	–.271
English	93	–.138	–.245
French, Hispanic and Italian Studies	66	–.169	–.296
Geography	40	–.163	–.217
History	29	–.134	–.230
Linguistics	4	–.125	–.253
Music	7	–.264	–.321
Philosophy	20	–.026	–.102
Political Science	34	+.003	–.031
Library, Archival and Information Studies	20	+.103	+.045
Sociology	27	–.089	–.114
Social Work	25	–.094	–.091
Theatre & Film	29	–.116	–.140
Women's and Gender Studies	13	–.158	–.253
Weighted Average (by # of sections in unit)	706	**-.095**	**-.177**

[a] Administered by the Dean of Arts Office.

[b] The mean differences are given as the 2008-09 value minus the 2007-08 value. Thus, a positive difference means that the 2008-09 value is larger than the 2007-08 value; a negative difference means the 2007-08 value is larger.

Average of the Six UMIs.

We see from Table 2 that, over 706 sections of the same instructor/course combination, the change in the *average of the six UMIs* from paper (2007-08, both terms) to online (2008-09, both terms) is a decrease of **.095** scale points on the 1.0–5.0 scale. (Notice, however, that in four administrative units, this average rating actually increased.) This average decrease would have to be considered small.

On the basis of our previously-discussed conservative inferential criterion (allowing a familywise Type I error rate of .10 over all comparisons), only 5 of the 20 departments demonstrated a statistically significant decline in the six-UMI average measure between 2007-08 and 2008-09, these departments being Economics, English, French, Hispanic and Italian Studies, Geography, and History.

UMI 6.

With this single UMI, the one that deals with overall teaching performance, we see a greater decrease in the course means than with the UMI Average measure, an average decrease of **.177** scale points over the 20 administrative units in Table 2. To some, an average decrease of this magnitude—from a mean of 4.341 in 2007-08 to one of 4.164 in 2008-09 (also presented in Table 1)—may be seen as having practical significance.

Looking at the administrative units individually, we see that 7 of the 20 manifested a statistically significant overall decline in UMI 6 between 2007-08 and 2008-09, these departments being Asian Studies, Central, Eastern and Northern European Studies, Economics, English, French, Hispanic and Italian Studies, Geography, and History.

Results Involving Item Variabilities for the Faculty as a Whole

Although changes in average item mean levels may be of the greatest interest, changes in the spread, or variability, accompanying the distributions of item mean levels are also worthy of consideration. The item standard deviations (SDs) for the faculty as a whole are given in Table 1. From this table, we can see that there was an increase in the item variabilities when going from the 2007-08 (paper) results to the 2008-09 (online) results, and that this trend held for all six UMIs. On average this increase was .0575 (.398 to .456) scale points. This represents approximately a 14.5% average increase in variability of item means across the 707 instructor/course combinations examined here.

With the exception of Item 4, these differences in item standard deviations are all highly significant—again using our conservative hypothesis-testing strategy—and the change with Item

4 comes very close to statistical significance. There is a clear trend in the new UMIs and their online administration for *individual differences* in perceived teaching effectiveness to have increased somewhat (where the "individual" is the instructor/course combination). That is, there is now a somewhat greater spread of class means on these items from one instructor/course unit to another.

From a psychometric perspective, this is a welcome outcome. One of the goals (perhaps the main goal) in the design of measurement instruments is for these to demonstrate as much variability as possible among the subjects scored. For one thing, greater variability enhances our ability to discriminate among subjects (in this case instructor/course combinations). This increased discriminability in turn results in higher reliability of the scores, in the sense that they are more stable and dependable. A very high mean on a scale (as we have with an overall mean well over 4.0 on a scale running from 1 to 5) produces *ceiling effects*, and reducing the mean and increasing the variability enables the distribution of scores to spread out into a slightly more symmetric form. One consequence of this is that better discrimination is possible at the upper end of the distribution. This improved discrimination in the upper, right tail could be important, for example, when using the results to decide on recipients of teaching awards.

We know that large differences in teaching performance exist, and having our scales represent this underlying reality a little better (as the greater variability permits) represents an improvement in measurement.

Some Observations from the Results for the Faculty of Arts

(a).Item Mean Differences from Paper (2007-08) to Online (2008-09) Administration

(i.) These differences are very small—completely negligible for 3 of the 6 items—with the possible exceptions of UMIs 6, 2, and perhaps 4, for which the decrease in item means might be considered non-negligible. Whether these mean decreases amount to more than the usual year-to-year variation will have to await the gathering of further data, such as that which will be obtained in the 2009-10 administrations.

(ii.) These small differences in item means will have little or no effect on the interpretations we have attached to mean scores in the various ranges of the scale. Thus, mean scores in the 3.0–3.5 range, for example, will continue to be seen as low student evaluations of teaching, whereas values in the 4.50+ range will continue to index very high student ratings.

We note that the norms and the standards against which instructors will be compared, using the UMIs with online administration, have been calculated on results with the new item wording and administration format. Given the item (and Average UMI) correlations from 2007-08 to 2008-09, we know that relative standings have generally not changed greatly. The changes made to the scales and administration format can be expected to affect all instructors relatively equally. The average item changes accompanying the shift from paper to online administration should, as noted earlier, be seen as simply part of the recalibrating of the teaching-evaluation process at the University. Within a short period—maybe two or three years—the new metric (which is almost identical to the old one) will be part of the UBC SEoT culture.

(b) Possible Reasons for the UMI Item Mean Differences from Paper (2007-08 Academic Year) to Online (2008-09 Academic Year) Administration

(i.) The wording of the items was changed slightly. In addition, the response scale for all items was changed from "Very Poor"… "Excellent" to "Strongly Disagree"… "Strongly Agree." Consider UMI 4 as an example (and one for which the mean rating decreased from 4.19 to 4.07, one of the larger mean-rating shifts). In the earlier, paper-administered, form, the item had read:

The fairness of the instructor's assessment of learning (exams, essays, tests, etc.),

and was responded to via the response scale:

(1) Very Poor (2) Poor (3) Adequate (4) Good (5) Very Good.

In the revised form, and using online administration, UMI 4 now reads:

Overall evaluation of student learning (through exams, essays, presentations, etc.) was fair

and is responded to via the scale:

(1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree.

Although to the eye the change is minor, it is possible that the revised wording and change in the response scale elicit slightly lower ratings.

(ii.) The response rates declined. In Arts there was a 10–15% decline in response rates, from 2007-08 to 2008-09.[1] This decline might account, to some extent, for slight shifts in mean ratings. In any case, possible response-rate-induced shifts can be considered a part of the recalibration process, and instructors will not be compared, of course, with what the numbers used to mean when the inventory of slightly different items was administered in paper format, but instead against the new norms that take into account the change in response rates.

(iii.) There are other possible explanations. It may be (although we have no empirical evidence at present supporting this) that those students who are more comfortable with the Internet are slightly more represented in the online administration sample, and that these students have slightly different views of teaching effectiveness than those of less computer-literate students. Other differences than comfort with the Internet, however, may characterize the new pool of respondents, and these pool differences might be seen as contributing in part (although likely a small part) to the small changes in item means, which, as noted earlier, define the recalibration process. Changes in the setting, from completing the inventory in the classroom to doing so online, may contribute, with students perhaps having more time to complete the items than in the classroom. Perhaps distractions that may occur while taking the online inventory (as opposed to the quieter classroom setting) contribute. We are planning further research studies to better understand the underlying causes of paper-to-online rating response shifts.

Findings in the Faculty of Science

Method

In the Faculty of Science, we compared paired instructor/course units that were given in 2008-09 (both terms) and evaluated using the online SEoT UMI inventory and that had also been given in 2006-07 (both terms) and evaluated with the Science paper inventory. Thus, the content—as well as the administration format—of the inventories was different on the two occasions. The total number of paired instructor/course sections in this sample is 268. With some departments, the number of such paired sections was very small (in some cases, one), and these departments were not included in the department-level results reported here, although they were included in the analysis of the faculty as a whole. We have included departmental results in the analyses for all those in which there were at least four instructor/course combinations that matched between the two academic years.

Since the items differ in wording and accompanying response scale and, to some extent, also content between these two inventories, we have considered only the one item that is relatively similar between the two—the summary items, Item 6 on the 2006 Faculty of Science paper inventory and UMI 6 on the current online inventory:

2006, Paper: Item 6: *The instructor taught effectively*.

Response Scale: (1) Strongly Disagree (2) Mildly Disagree (3) Neutral (4) Mildly Agree (5) Strongly Agree

2008, Online: Item 6: *Overall the instructor was an effective teacher*.

Response Scale: (1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree

[1] We note that an online-inventory response rate of around 60% (approximately that found in the Faculty of Arts), although lower than that previously experienced at UBC with the paper form, is *very high* in comparison with response rates to online surveys conducted at most other universities, where rates of 30–50% are commonplace.

Results: Differences in Means

Table 3 contains the decrease or increase by department in average ratings on inventory Item 6 in both cases, along with the two-year-lag correlations for Item 6 and the standard deviations on each occasion.

Table 3. ***Results by Faculty and Individual Department on Item 6***

Department (# of Paired Sections)	r_{xy}^{a}	Difference in Means on Item 6 between 2006-07 and 2008-09[b]	SD 2006-07	SD 2008-09	Difference in SDs[b]
Faculty (268)	**.613**	**-.085**	**.532**	**.460**	**-.072**
Biology (69)	.62	–.11	.469	.453	–.016
Chemistry (39)	.55	–.16	.479	.414	–.065
Computer Science (28)	.35	–.04	.408	.504	+.096
Earth/Ocean Sciences (52)	.43	–.16	.411	.385	–.026
Mathematics (22)	.81	+.10	.656	.587	–.069
Microbiology (19)	.60	–.11	.429	.354	–.075
Physics (28)	.63	+.03	.678	.492	–.186
Statistics (4)	.94	–.02	.782	.520	–.262

n = 268 sections in Faculty row (261 total in the departments tabled).

[a] These are the correlation coefficients between instructor/course mean scores on Item 6 from the 2006-07 paper form and those (Item 6) on the 2008-09 online inventory.

[b] The differences are given as the 2008-09 value minus the 2006-07 value. Thus, a positive difference means that the 2008-09 value is larger than the 2006-07 value; a negative difference means the 2006-07 value is larger.

On the basis of our previously-discussed conservative inferential criterion (allowing an overall Type I error rate of .10 over the mean comparisons above), the mean decrease between the two years is statistically significant for the faculty as a whole (–.085) and for Earth/Ocean Sciences (–.16), but not for any other department (and in Mathematics and Physics, there was a small, but nonsignificant, *increase* in Item 6 mean scores).

Differences in Item Variabilities

The item standard deviations (SDs) are given in Table 3 for the faculty as a whole and the eight departments considered over the two academic-year administrations. From this table, we can see that there was generally a decrease in the Item 6 variability from the 2006-08 (paper) results to the 2008-09 (online) results, and that this trend held for all but one department (Computer Science). It should, however, be noted that none of the separate-department SD differences was statistically significant, undoubtedly in most cases because of very small sample sizes. If we consider the faculty as a whole, though, we see a statistically significant decline in variability of Item 6 scores (SDs of .532 for 2006-07 and .460 for 2008-09) of approximately 13.5% (in the SD metric) when going from the paper to online format. We note here that this result is just the opposite of what we saw in the Faculty of Arts results.

Some Observations Regarding Differences in Means and SDs from the Faculty of Science Results

(a) Item Mean Differences on Item 6 from Paper (Science Inventory; 2006-07 Academic Year) to Online (UMI 6; 2008-09 Academic Year) Administration

(i.) These differences are, for the most part (and certainly for the faculty average of –.085), small and may be little different from what we might see from year to year with the same administration format.

(ii.) As was found with the Faculty of Arts results, these differences can be expected to have little or no effect on the interpretations we have attached in the past to mean scores in the various ranges of the scale.

(iii.) The mean difference (decrease) on Item 6 for the Faculty of Science as a whole of .085 is very close to that found for the Faculty of Arts (.095), when the Arts results were in terms of the average of the six UMIs. When the UMI 6 means were compared in the Faculty of Arts data, however, the difference was somewhat larger, –.177.

(b) Possible Reasons for the Item Mean Differences from Science Item 6 via Paper (Science Inventory; 2006-07 Academic Year) to UMI 6 Online (2008-09 Academic Year) Administration

(i.) The wording of the item and the response scale was changed to a slight degree. As noted above and shown again, the items read:

2006, Paper: Item 6: *The instructor taught effectively*.

Response Scale: (1) Strongly Disagree (2) Mildly Disagree (3) Neutral (4) Mildly Agree (5) Strongly Agree

2008, Online: Item 6: *Overall the instructor was an effective teacher*.

Response Scale: (1) Strongly Disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly Agree

(ii.) In Science, there was no appreciable decline in the student response-rates from paper-based administration to the online form, and for this reason, unlike with the Faculty of Arts results, simple response-rate differences cannot explain the differences in mean scores. However, there may be differences between the students who came to class and completed the paper form used in the past and those willing to go online to complete the inventory. It is possible that those students who are more comfortable with the Internet and, for this reason, slightly more represented in the online administration sample, have slightly different views of teaching effectiveness than those of less computer-literate students. To some extent at least, any possible shifts that have arisen from a slightly-different population of student-respondents can be considered a part of the recalibration process, and instructors will not be compared with what the numbers used to mean (with a different respondent population), but instead against the new norms that take into account the change.

(iii.) Other differences than comfort with the Internet, however—as noted earlier in connection with the Faculty of Arts results—may characterize the new pool of respondents, and these pool differences might be seen as contributing in part (although likely a small part) to the small changes in item means, which, as noted earlier, define the recalibration process. Changes in the setting, from completing the inventory in the classroom to doing so online—convenience, time available, etc.—possible distractions, and other unknown factors may all contribute.

(c) Differences in the Variability of Class Item 6 Means from Paper (Science Inventory; 2006-07 Academic Year) to Online (UMI 6; 2008-09 Academic Year) Administration

We noted above a faculty-wide decrease of approximately 14% in the variability (quantified by the SDs) in the Item 6 instructor/course means in going from the paper to the online

version. In attempting to understand this decrease in variability, two points are worth considering. First, the phenomenon noted above in (b)-(iii.) may be playing a role here. If, in fact, the students who are willing to respond online represent a subset of the larger population of Science students, it may be that their responses tend to be somewhat more uniform than we would see from the larger population of students. We have, however, no explanation for why the same factor would not account for a similar decrease in variability in the Faculty of Arts results, where, as we have seen, there was actually an increase in the SDs.

Second, it is worth noting that the UMI 6 variability in the 2008-09 online administration (SD = .458, *n* = 275) is very close to the UMI 6 variability in the 2008-09 online administration in the Faculty of Arts (SD = .483, *n* = 707). In fact, these two standard deviations are not significantly different. If we expand this discussion by adding in the UMI 6 item means, we see that these too were similar between the two faculties (Table 4 below).

Table 4. Results by Faculty on UMI 6 in 2008-09 Online Administration

Faculty	Number of Sections	Mean	SD
Science	268	4.10	.458
Arts	707	4.16	.483

In fact, as with the two standard deviation estimates, the two faculty means of 4.10 and 4.16 are not significantly different. The .06 scale-point difference should be regarded as nothing more than sampling error and of absolutely no consequence. It must be stressed, however, that the present results and conclusions involving faculty mean and SD estimates are restricted to only those courses/sections that were taught by the same instructor over a one- or two-year interval (this particular sampling done to enable a more precise examination of mean/SD changes from the paper to online format). A corresponding analyses of *all* courses taught in 2008-09 in both faculties might yield different results.

A Brief Look at Item 6 Stability

In addition to information about Item 6 means and standard deviations, we see some correlation coefficients in Table 3. As noted there, these are the correlation coefficients between instructor/course mean scores on the 2006-07 paper form and those on the 2008-09 online inventory. As such, these values could be seen as lower-bound estimates of the stability (a form of psychometric reliability) of Item 6 mean scores over time.

Normally, stability is assessed for measures that tap constructs that *should* be stable over time, such as ability, personality traits, interests, etc. However, in the present case with the UMIs, we might expect some *function fluctuation* (that is, actual change on the construct itself) because of instructors' perceptions of their results and efforts to improve them. To the extent that this occurs, assessed stability of construct measures will decrease from usual levels. Further, there are other factors here that are not generally present in assessments of measurement stability: (a) a wording change, (b) a change in administration format, and (c) a longer-than-usual time interval between the two administrations. All of these factors can be expected to reduce measurement stability estimates, and for this reason we have characterized the present correlational results as *lower-bound estimates*.

All of the correlations in Table 3 are either statistically significant or very close to being so (missing significance because of extremely small sample sizes in addition to a conservative decision rule). If we focus on the one value based on a large sample, that for the faculty as a whole, we see a stability coefficient of **.613** *for the two-year time interval*. This should be regarded as exceptionally high, particularly given the factors noted in the just-preceding paragraph . By way of providing some perspective on this, we note that measures of most stable personality traits produce long-term (two-year) stability coefficients not substantially larger than this, and these latter coefficients are for entire scales of (perhaps 30) items administered exactly the same way on the two occasions. Thus, all evidence at present suggests that our assessments of the construct measured by UMI 6 are stable.

From the Faculty of Arts data, we calculated corresponding (one-year) stability estimates for UMI 6 (administered on both occasions). With these data, some of the same extraneous factors as with the Faculty of Science data were present: (a) a (very) slight wording change and a response-scale change and (b) the change in administration format. For UMI 6 in the Faculty of Arts as a whole, the 2007-08 to 2008-09 stability coefficient was **.584**, a value very close to the corresponding stability estimate for the Faculty of Science.

We conclude this section by noting that after the Term 1 and 2, 2009-10 UMI administrations, we will have the necessary data to calculate better long-term stability estimates for all the UMIs and their average. We will correlate UMI means from the 2008-09 instructor/course combinations with those from the corresponding 2009-10 instructor/course combinations for all instructor/course combinations that are identical over the two academic years. This measurement design will remove all three of the contaminating factors noted above in that the items will have exactly the same wording and response anchor points, along with

administration format, and the design will have the more-common one-year time lag—which is still a long one and which will yield authentic *long-term* stability estimates.

Summary and Conclusions from the Analyses

Although we have focused on a number of statistical phenomena in the two faculty-specific analyses, the central question has been whether the change from paper to online administration has caused instructor means to either increase or decrease. On the basis of results obtained from use, by the Faculty of Arts and Faculty of Science, of the online UMIs in the 2008-09 academic year—and based on a total of nearly 1,000 separate instructor/course combinations—we believe that our assessment of the magnitude of change is sound. In Table 5 below, we summarize the relevant findings:

Table 5. Summary of Means for Both Faculties

Faculty	# of Instructor/ Course Sections	Nature of Mean	Earlier Mean (Paper)	2008-09 Mean (Online)	Difference[a]	.95 Confidence Interval for $\mu_1 - \mu_2$
Arts	707	Avg. of 6 UMIs	4.245	4.150	-.095	(-.070, -.122)
		UMI 6 (both academic yrs)	4.340	4.163	-.177	(-.147, -.207)
Science	268	Item 6 (2006-07) UMI 6 (2008-09)	4.184	4.099	-.085	(-.033, -.137)

[a] 2008-09 mean minus the earlier mean.

We thus see that, on average, decreases of .095 for the average of all six UMIs and of .177 for UMI 6 in the Faculty of Arts have accompanied the change from paper to online administration in this examination involving one year of paper-inventory results and one of online results. In the Faculty of Science, we have found a mean decrease from the previous summative Item 6 to the current UMI 6 of .085 when comparing the paper-administered inventory with the online version. Our conclusion is that the shift from paper to online inventory administration has had a relatively small effect on the magnitude of class mean scores obtained by instructors and that these new mean scores obtained from online inventory administration will be interpreted contextually (particularly as high, average, or low) in much the same way as previously.

Even if there had been a substantial difference in the class means resulting from the transition from paper to online administration, however, we note that the norms and the standards against which instructors will be compared in the future, using the UMIs with online

administration, will be calculated on results with the new item wording and administration format. On the basis of one year's use of the online results, a large normative base now exists for comparative purposes, and this base will be expanded by each subsequent year's results.

It is anticipated that relative standings generally will not change greatly. The scale/administration changes can be expected to affect all instructors relatively equally. As noted earlier, the analyses accompanying the transition from paper to online administration should be seen as simply a part of the recalibrating of the student teaching-evaluation process to be used in the future at UBC. We believe that, within a short period—maybe two or three years—the new metric (which is almost identical to the old one) accompanying the UMIs will be part of the UBC SEoT culture.

! LJLJŜyŘȺÈ 5¢ REPORT TO THE UNIVERSITY OF BRITISH COLUMBIA VICE PRESIDENT ACADEMIC AND PROVOST

ON THE

NEW UNIVERSITY-MODULE ITEMS AND THEIR ONLINE ADMINISTRATION AT THE UNIVERSITY OF BRITISH COLUMBIA

A. RALPH HAKSTIAN
DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF BRITISH COLUMBIA

WITH ASSISTANCE FROM

CATHERINE D. RAWN, MA
AND
CARRIE CUTTLER, MA
PHD STUDENTS IN
THE DEPARTMENT OF PSYCHOLOGY, UBC

13 MARCH 2008

ACKNOWLEDGEMENTS

Catherine Rawn wrote all of Section 2 of this report. Carrie Cuttler did all of the data analyses as well as some editing. Both should be regarded as co-authors of this report. Russell Ball provided much-needed editing assistance.

Work like this is not done in a vacuum. The comments along the way and the writings of Gary Poole, Director of UBC's Centre for Teaching and Academic Growth (TAG), have helped to inform this report.

TABLE OF CONTENTS

Contents **Page**

EXECUTIVE SUMMARY iii

SECTION 1 – THE PURPOSE AND SCOPE OF, AND BACKGROUND TO, THE PRESENT REPORT 1
Purpose and Scope 1
Background 1
Summary 2

SECTION 2 – A BRIEF HISTORY OF AND SOME ISSUES CONNECTED WITH STUDENT EVALUATION OF TEACHING 3
Do Students Discriminate Meaningfully? 3
Are Online or Paper-Based Measures Better? 3
Main Uses of Student Evaluations and Perceptions of Each Use 4
Potential Sources of Bias in Student Evaluations of Teaching 5
Use of Student Evaluations at Noteworthy Universities 6
Current Use of Student Evaluations at UBC 7
Summary of How Student Evaluations of Teaching Are Generally Seen in the Context of Comprehensive Evaluation of Teaching at UBC and Elsewhere 8

SECTION 3 – DESCRIPTION AND ASSESSMENT OF UMIS AND THEIR ADMINISTRATION 9
The Six University-Module Items and Some Psychometric Results with Them 9
Inspection of Item Means 9
Inspection of Item Content 14
Reliability 15
Validity 22
Brief Summary of What We Know about the Six UMIs 27
Online Administration of the UMIs 27
Effects of Administration Format on Response Rates 27
Effects of Administration Format on Scores 28
Brief Summary of Results Concerning Administration Format 31
A Very Brief Look at Some Correlates of Student Ratings of Teaching 31
Some Comments on the Posting of Teaching-Evaluation Results on a University Website 33
Rationale for Posting Student-Evaluation Results 34
Features of the Proposed Posting Initiative 34

SECTION 4 – SUMMARY AND RECOMMENDATIONS 36
Summary 36
A General Recommendation 36
More Specific Recommendations – I. Short-Term 36
More Specific Recommendations – II. Longer-Term 38

REFERENCES 42

EXECUTIVE SUMMARY

BACKGROUND

An examination was undertaken, on behalf of the Vice President Academic and Provost, of the performance of the new *University-Module Items* (*UMI*s) administered at the University in the Fall term, 2007. The scope of the investigation included evaluation of psychometric characteristics of the UMIs and of the similarity of item performance from paper to online administration of student-evaluation inventories. Another aspect of the University initiative—that of the posting of the student-evaluation results to a password-protected website for access by students and faculty—is considered. Recommendations for future student-evaluation activities at the University are provided.

The implementation of the UMIs followed several Senate recommendations for a uniform student-evaluation process, with the most recent recommendation also urging the online publication of teaching-evaluation results. The body most closely responsible for the implementation of these recommendations is the Student Evaluation of Teaching (SEoT) Committee.

Student evaluations of teaching at colleges and universities have been ubiquitous for the past 40 years or more. Research on student evaluations has shown them to be useful for the purposes of teaching improvement and not merely a reflection of instructor popularity. The research evidence on paper- vs. online-administered student evaluations has revealed that both produce comparable results, and, with increasing familiarity with computers and the Internet by students, online administration is increasing. Research has also revealed certain extraneous factors that can affect student evaluations. At UBC, student evaluations have, for some time, been used across campus for both formative and summative purposes. Although administration of the inventories has for the most part been in paper format, some administrative units have experimented with online presentation and some, in addition, have, in the past, published student-evaluation results. At present, only the Faculty of Arts publishes these results on a website.

ANALYSES PERFORMED

The six UMIs were examined for a number of performance characteristics. Inspection of item content and scores revealed that the UMIs measured instructional themes and produced score levels and distributions very similar to those seen in the past with existing items. The items were seen as tapping into central aspects of effective teaching and learning and were sufficiently comprehensive to provide adequate assessment of the important facets of instruction.

Further, the UMIs were found to be of comparable reliability to existing items, in terms of both stability over time and internal consistency. The question of inter-rater reliability of mean item results (the unit used for formative and summative purposes) was addressed, and this form of reliability was seen as fully adequate as long as the means were based on at least 10-15 student raters.

Validity of the UMIs was assessed by correlating scores on them with those on existing items that were designed to measure the same aspects of teaching. Results indicated that the UMIs provide valid assessment of what have been identified as central aspects of teaching.

The question of the effects of changing from paper administration of the student evaluation inventories to online administration was addressed by an examination of the student response rates under the two formats, as well as that of the general level of results obtained under each. Results of analyses of student response rates in four administrative units for which online data were available indicated very little, if any, reduction in response rates for the online presentation format. Similarly, comparisons of mean item ratings showed no systematic differences favouring one format over the other. There is no evidence from the present evaluation, therefore, to suggest that online administration of the UMIs (along with other faculty- or department-specific items) will cause a decline in response rates or any change in the general level of ratings awarded by students.

Several factors that can be expected to affect student evaluations were investigated. The relationship between class grades and mean ratings, that between class grades and student response rates, and that between mean ratings and response rates were all examined. Very low correlations were found in analyses across nine administrative units, with those between class grades (actually awarded) and mean UMI scores the highest—averaging .27. Although low, this correlation raises the question of whether grading leniency affects student ratings, but this relationship can be explained in several ways, with grading leniency only one possibility. This topic deserves further research, but the correlations found with extraneous factors do not undermine the use of the UMIs (or any items) for reliable and valid student evaluation of teaching.

DISCUSSION OF POSTING OF RESULTS

The subject of the posting of student evaluation results on a University website is discussed. Although no data have been collected in this regard and no local empirical findings inform this discussion, prior experience with posting evaluation results is available and suggests some ways in which the goals of Senate—in providing systematic, reliable, and valid course-planning information—can be realized.

RECOMMENDATIONS

A number of recommendations are made, based on the findings of the present investigation. It is generally recommended that each administrative unit consider adding items to the UMIs that reflect the specifics of teaching in that unit. More specifically, a number of recommendations are presented that concern tasks to be performed in the short term—in time for the Fall, 2008 administration of student evaluations. In addition, a list of recommendations is provided containing tasks that can be completed over a longer period of time. The intention of these recommendations is to aid the University in systematically developing, over the next few years, fully effective procedures by which student evaluation of teaching is made more precise, more useful, more widely-accepted and optimally-used, and more clearly tied to pedagogical upgrading opportunities available on campus than it currently is.

One summary observation that deserves mention is the generally high regard that UBC students have for the quality of their teaching. In the various analyses performed, mean scores falling between “Good” and “Excellent” were routinely found, indicating that students have reported that teaching is generally at a high level at UBC.

SECTION 1 – THE PURPOSE AND SCOPE OF, AND BACKGROUND TO, THE PRESENT REPORT

Purpose and Scope

The primary purpose of the investigation performed and reported here was to examine in some detail the newly-introduced *university-module items* (or *UMI*s in the sequel) developed to provide a uniform component or basis across campus for evaluation of teaching by students. The psychometric characteristics of these six items, as well as some effects of their presentation online—a departure from the traditional paper-format method of administration—have been examined, with the results and recommendations that follow largely concerned with these matters.

I have, however, understood my mandate with respect to the present evaluation to be somewhat broader than a strict examination of the items and administration format. On the basis of approximately 30 years of close involvement in the teaching-evaluation activities in the Department of Psychology, I have acquired experience not only in the writing and scoring of items and presentation of results to instructors, but also in the thornier aspects of student evaluation of teaching, such as the issues surrounding its use as well as the development of a departmental system to implement the online posting of student-evaluation results. Thus, although we do not at present have solid empirical data on the basis of which to address the topic of posting results, I do have experience with and some observations about it, and I will discuss it, albeit somewhat briefly, and make some recommendations about implementation.

The scope of this evaluation does not, however, extend to matters concerning larger principles of the proper use of teaching evaluation results or University policies in this regard. Thus, matters such as, for example, how tenure and promotion committees will use student-evaluation data and how the various administrative units should best combine the various available forms of teaching-evaluation information fall outside the terms of the present investigation.

Background

Over a number of years, the UBC Senate has recommended certain initiatives concerning the evaluation of teaching, these arising in 1978, 1991, 1996, 1999, 2000, and 2006. In April, 2005, a committee was established under the leadership of Vice-Provost Anna Kindler: the initial Student Evaluation of Teaching (or SEoT) Committee. The Committee drafted a preliminary report that recommended a revision of the process and approach used in the student evaluation of teaching. The report was brought to the UBC Senate in May, 2006 by the Senate Teaching and Learning Committee, which proposed a number of recommendations. The Senate approved those recommendations in-principle. A new joint SEoT Committee was then established, co-chaired by Dr. Joy Johnson, the Chair of the Senate Teaching and Learning Committee, and by Dr. Kindler, to focus on implementation of the recommendations.

A pivotal recommendation of the committee was that any UBC student evaluation of teaching be capable of serving multiple constituencies: the central administration; the faculties; the instructional units; the individual instructors; and students. Following from this principle was the recommendation of a modular approach to student evaluation of teaching. That is, there will be four “modules” within the evaluation, reflecting items of interest to different constituents. Faculties, departments, and individual instructors would be free to design questions at the

appropriate level of the questionnaire to meet their own needs for evaluation. Thus, the four modules would be University items, faculty items, department items, and instructor-written items. All levels of the questionnaire would have the fundamental Senate objective to improve teaching and learning, but there might be very different perspectives on what to ask, and how to use the information across different modules. Much of the work involved in developing the teaching evaluation process was in fact already done, in that all departments already had modules in use. Sometimes these were at the faculty level, such as in Science, and sometimes at the departmental level, as in Psychology. The committee was very clear that where good practices already existed, nothing would be mandated to change those practices. The remaining challenge was to develop a module which could be used at the university level; a small set of questions from which information could be aggregated across all courses (with a few exceptions mentioned in the Senate policy). The recommendations of the committee reflect the value expressed by students and central administration in having some questions which are comparable and can be aggregated across the entire university.

The SEoT Committee then consulted with other universities (e.g., Purdue, Stanford) who had implemented online evaluation systems similar to that recommended at UBC. Consideration was given to developing an appropriate number of inventory items that would tap the most important aspects of effective teaching. Since these university-wide items would in many cases supplement items developed by individual administrative units, their number had to be kept to a minimum to prevent an overly-long evaluation experience for students. Yet the committee also wanted the items to be as comprehensive as possible, covering the multidimensional nature of teaching. Hence they are quite general, with the understanding that a drilling down to more specific and diagnostic questions would be possible in a lower-level module more appropriate to the range of concerns and contexts. It is important to note that students played a significant role in the item-development process described here.

The result of the item-generation work overseen by the SEoT Committee was a set of six items that captured important facets of effective teaching and were consequently recommended to the Provost for implementation. The items were further edited in response to additional feedback received from faculties, departments, and individuals after the initial list was released to the units. In addition, further implementation concerns were discussed; two of these were (a) the online presentation of the six items and (b) the posting each term on a university website the results of the assessment for individual instructors.

Summary

It should be reiterated here that the concern in this report is mostly with the six UMIs that were arrived at from the work described above. The objective of the University has not been that these items be used solely, but rather that they be used as part of a larger student evaluation system in faculties and departments to ensure that more specific aspects of pedagogy in those units are reflected in the evaluations. This overall system would ideally include items designed for specific administrative units and those written by individual instructors for their own, primarily formative-evaluation, use. The latter augmentation of the UMIs makes sense, in that such a combination of items—if not yielding so lengthy an inventory as to induce evaluation-fatigue—would produce some cross-university uniformity at the same time as more finely-grained faculty- , department-, and instructor-specific information. The primary concern in this report, however, is with the performance of these UMIs and their presentation format.

SECTION 2 – A BRIEF HISTORY OF AND SOME ISSUES CONNECTED WITH STUDENT EVALUATIONS OF TEACHING

Students in higher education have evaluated teaching in North America for almost a century (Doyle, 1983, as cited in d'Apollonia & Abrami, 1997). During that time a large body of research has explored the psychometric properties and uses of student evaluations of teaching. Historically, there has been debate regarding the reliability and validity of student evaluations of teaching, primarily during the 1970s and 1980s (Greenwald, 1997). The current consensus among scholars in this area is that student evaluations of teaching are generally reliable and valid indicators of students' perceptions of their learning experience (Harrison, Moore, & Ryan, 1996; Johnson & Ryan, 2000; Marsh, 2007). This statement is particularly true when the instruments have been developed with the aid of psychometricians and have undergone reliability and validity analyses (Marsh & Roche, 1997).

Do Students Discriminate Meaningfully?

Teaching is a multifaceted endeavour and research shows that students are sensitive to this reality. Students discriminate among different facets of teaching that they can observe, such as lesson organization versus charisma (Aleamoni, 1976, 1999). Ratings of conceptually distinct facets such as these tend to be unrelated to each other in student evaluations (d'Apollonia & Abrami, 1997; Marsh, 1984). Moreover, students do *not* tend to use characteristics like warmth, friendliness, and use of humour as the primary source of their ratings of teaching (Aleamoni, 1999). Instead, students use variables such as teaching methods, perceived fairness, and respect for students (Moore, 2006; Remedios & Lieberman, 2008), and organization, motivation, stimulating interest, treating students courteously, and answering students' questions (Feldman, 1989; Tang, 1997) as the basis of their evaluations[1]. In sum, student evaluations of teaching do not merely reflect instructor popularity (Aleamoni, 1999).

Are Online or Paper-Based Measures Better?

Traditionally, student evaluations of teaching have been conducted using paper-and-pencil measures. There is increasing interest in online versions of these scales. Only two percent of American institutions reported using online student evaluations in 2000 (Hmieleski, 2000, as cited in Hoffman, 2003). However, ten percent of a large sample of American institutions (N = 256) reported using a university-wide online version in 2003, with even higher rates reported for online courses (56 per cent; Hoffman, 2003). Online and paper-based measures produce comparable evaluations of teaching in terms of the ratings given (Hardy, 2003; Heath, Lawyer, & Rasmussen, 2007; Layne, DeCristoforo, & McGinty, 1999). There are a variety of practical and research-supported benefits of online over paper-based evaluations, including lower costs, more class time for teaching rather than completing forms, greater accessibility for students, longer and more thoughtful student comments, more accurate data collection and reporting, and reduced processing time, among others (Ballantyne, 2003, see also Sorenson & Johnson, 2003). Students tend to prefer online evaluations yet are more suspicious about their anonymity with an online versus paper method of evaluating teaching (Layne et al., 1999). This finding suggests that the process of ensuring anonymity may need to be transparent to students in order to gain full acceptance of the online system. One major concern regarding online versions is the possibility of a reduction in response

[1] Predictors differ depending on those which are measured in the particular study.

rate relative to paper versions administered in class, yet extant research does not support this conclusion. Response rates have been found to be similar across the two methods (Hardy, 2003; Heath et al., 2007; Kelly, Ponton, & Rovai, 2007; Johnson, 2003), with online versions tending to glean more and much lengthier open-ended comments than do paper versions (Heath et al., 2007; Johnson, 2003). Based on this research, online student evaluations of teaching are a viable alternative to paper-and-pencil measures.

Main Uses of Student Evaluations and Perceptions of Each Use

Student evaluations of teaching are primarily used for two purposes: (1) *formative evaluation* intended to improve the quality of teaching, and (2) *summative evaluation* intended to inform personnel decisions such as tenure and promotion (Murray, 1987). Formative evaluation tends to be valued by faculty (Murray, 1982, as cited in Murray 1987; Schmelkin, Spencer, & Gellman, 1997). Moreover, students value the opportunity to provide meaningful feedback that will result in teaching and course improvement (Chen & Hoshower, 2003; Giesey, Chen, & Hoshower, 2004; Spencer & Schmelkin, 2002), yet doubt it is taken seriously by administration (Spencer & Schmelkin). Indeed, when student evaluations of teaching are used for formative purposes, positive changes can result in increased student learning (Barnett & Matthews, 1997), particularly when instructors are supported through the interpretation process (Aleamoni, 1999; Cohen, 1980; Murray, 1997).

A relatively more controversial use of student evaluations is for summative purposes (Ryan, Anderson, & Birchler, 1980). When reliable and valid measures are used and interpreted in appropriate ways, personnel decisions can be well-informed by considering students' perspectives in the evaluation of teaching (Marsh, 1984, 2007; McKeachie, 1997) in concert with other forms of data as presented in a complete teaching portfolio. Anecdotal evidence tends to indicate resistance toward student evaluation of teaching (see Schmelkin et al., 1997), which may have spurred the view that there are commonly held myths regarding these ratings (e.g., Aleamoni, 1987, 1999). However, little research has empirically documented faculty's perceptions of student evaluations for the summative purposes. Existing data suggest that faculty's view of summative teaching evaluations ranges from neutral (Barnett & Matthews, 1997) to generally useful for informing personnel decisions despite some disagreements about particular items (Schmelkin et al., 1997). From students' perspectives, summative use tends to be viewed as less important than formative evaluation (Chen & Hoshower, 2003; Giesey et al., 2004). The use of student evaluations of teaching for personnel decisions does not appear to enjoy the same positive attitudes by students and faculty as do formative evaluations. However, this use of evaluations may be perceived more favourably by faculty than anecdotal evidence tends to portray.

A third use for student evaluations of teaching is to inform students regarding their course selections (Richardson, 2005). The popularity of websites such as *ratemyprofessors.com*, which has received nearly six million postings across over 6000 schools since 1999 (Felton, Koper, Mitchell, & Stinson, 2008), suggests that there is a demand for obtaining peer input regarding course selection. Indeed, research shows that students are interested in viewing their peers' ratings of teaching to inform course selection, but faculty have expressed concerns about this use of evaluations, citing concern for privacy among others (Howell & Symbaluk, 2001). These attitudes are consistent with the potential risks for each group: students stand to gain information whereas faculty stand to risk negative publicity. Perhaps of import, these data (Howell & Symbaluk) come primarily from 2-year institutions in the United States; further research is needed to determine how faculty feel at 4-year

research-focused institutions. Nonetheless, student evaluations of teaching are being published for student view (Llewellyn, 2003). In a survey conducted in 2002, 12 per cent of American colleges and universities reported sharing evaluations with their students online, and another three percent were implementing this process in 2003 (Hoffman, 2003). Notably, the Faculty of Arts at UBC has had (with the consent of the participating faculty members) a searchable online database of student evaluations of teaching since 2006, geared toward students as partners in the learning process. Posting student evaluations of teaching online for student use is a relatively new way that these ratings are being used at UBC and across the continent. (We note here that evaluation data in paper form have been posted at UBC, off and on, for many years.)

Potential Sources of Bias in Student Evaluations of Teaching

Many variables that may influence student evaluations of teaching—besides instructional quality—have been investigated to various extents (Pounder, 2007). Overall, the literature on these potential biases has few definitive answers at this point. For example, there are few studies examining the impact of *native English-speaking* faculty versus faculty whose primary language is not English. One large study (N = 2,039) has found that native English-speakers tend to be rated as more clear in lecture delivery and more favourably overall, relative to non-native English-speakers (Ogier, 2005). However, the dearth of research in this area precludes recommendations based on a single study.

Research suggests that *course content* influences student evaluations (Cashin, 1990; Ogier, 2005; Santhanam & Hicks, 2002), supporting the view that student evaluations of teaching should be referenced to local faculty norms—performance standards that exist for courses in specific content areas. Moreover, the effect of *class size* tends to demonstrate a quadratic function, such that student evaluations of very small and very large classes tend to be higher than moderate sized classes, although the evidence is mixed (Aleamoni, 1999; Pounder, 2007). Much research has investigated the role of *instructor gender* on student evaluations of teaching. These results are inconclusive (Pounder, 2007). Some research reports females faring better than males (e.g., Feldman, 1993; Kierstead, D`Agostino, & Dill, 1988); some reports females faring worse than males (e.g., when providing negative feedback to students, Sinclair & Kunda, 2000); and still other research emphasizes the role of gender stereotypes by discipline (e.g., Langbein, 1994).

There is a concern that courses involving light workloads or resulting in high grades will be evaluated more favourably by students than courses that involve heavy workloads or result in relatively lower grades. Workload tends to have a small positive relationship with student evaluations of teaching (Heckert, Latier, Ringwald-Burton, & Drazen, 2006; Marsh, 1987; Marsh & Roche, 2000; cf. Dee, 2007 for zero relationship, and Greenwald & Gillmore, 1997b for a negative relationship). More recent research parsed out the effects of workload that was perceived by students as beneficial to their learning, and that which was not (Heckert et al., 2006; Marsh, 2001). Results of these studies indicate that both teaching effectiveness and student evaluations of teaching can be improved by increasing good (beneficial) workload and decreasing bad workload.

Anecdotal evidence suggests that faculty tend to believe that grading leniency leads to higher student evaluations and sometimes act in ways to capitalize on this perceived relationship (e.g., see Greenwald & Gillmore, 1997a; Marsh & Roche, 2000). Greenwald and Gillmore (1997a, 1997b) have noted a negative correlation between perceived workload and expected grades and have argued that the validity of student evaluations can be improved by removing the effects of grading leniency. There is a small positive relationship between grades (actual and expected) and student

evaluations of teaching. However, many studies, including meta-analyses and those with very large datasets, demonstrate correlations in the .20 range (varying from .10 to .30), which indicates a small effect (e.g., Marsh, 1980, 1983; Marsh & Roche, 2000; Stumpf & Freedman, 1979). Researchers have argued that at least a portion of this relationship is due to a valid relationship between teaching effectiveness and higher grades, concluding that grading leniency is not a substantial concern when interpreting student evaluation of teaching (Marsh & Roche, 2000).

In summary, a variety of variables have been examined with respect to their influence on student evaluations of teaching. The effects of course content and class size—two sources of data that are readily available at UBC under the current system—appear to impact course evaluations in ways outlined above. The role of instructor gender and fluency with the English language are unclear given extant research. Lastly, the roles of workload and grade leniency do not appear to contribute substantially to student evaluations of teaching.

Use of Student Evaluations at Noteworthy Universities

Student evaluations of teaching are ubiquitous across North America and increasingly in other parts of the world (Kwan, 1999). All of Canada's G13 universities (leading research-intensive universities in Canada) ask their students to evaluate teaching ("Summary," 2007)[2]. Two of those schools (Calgary and McMaster) have at least one common question that is administered across the entire school, and two others report similar questions appearing across campus (Montreal and Dalhousie). Four other institutions have a standard item bank from which departments can draw to construct their own questionnaires (McGill, Queen's, Alberta, and Laval).

Top Canadian schools tend, for the most part, to use paper-based evaluation methods ("Summary," 2007). Of 11 schools in the G13 group (excluding Waterloo and UBC), two use online rating forms exclusively (Calgary and McGill), although Calgary has experienced a drop in response rates and is considering reverting to paper format. Laval, Toronto, Dalhousie, and Ottawa have both online and paper options available, at least for some classes (e.g., distance education classes are evaluated online). The proportion of use of online questionnaires at U.S. schools was estimated in a 2002 study of 256 American colleges and universities. At that time, 10% used online evaluations, and 90% used a paper version (Hoffman, 2003). Twenty-two percent of the sample used the internet to disperse results to faculty, and 12 percent used the internet to inform students of their peers' ratings of instructors.

Unfortunately, the Hoffman (2003) study is now badly out-of-date, and there does not appear to be anything as thorough to give us current rates of online use (although we can reasonably assume present, 2008, levels to have risen considerably above those found in 2002). Articles concerned with this topic tend to cite Hoffman and note that the frequency of use of online student evaluations is increasing. One interesting online source of information, however, was located at: http://onset.byu.edu. This website, although not providing anything like a comprehensive or exhaustive account of schools that perform online student evaluations, is dedicated to this topic and does provide an informal listing of a number of institutions that use the online format in at least some departments. These institutions include seven in Canada and a number outside of North America.

[2] Data from the University of Waterloo were not included in the table compiled by McGill; however, other sources confirm that they do in fact ask students to evaluate teaching.

A study examining teaching evaluation procedures in schools of pharmacy also reveals the universality of student evaluations of teaching. A questionnaire was sent to all 79 members and affiliates of the American Association of Colleges of Pharmacy, including American, Canadian, and some non-North American schools (Barnett & Matthews, 1998). All 72 of the schools who responded to the survey (91.1%) regularly ask students to evaluate classroom teaching, which represents a 60% increase from the previous tally in 1976. Thirty-nine percent of schools used a university-wide instrument. Individual items on each school's instrument were content-analyzed to reveal common topics of evaluation. At least one third of schools included specific items addressing several key concepts: overall teaching ability (56% of schools), clarity of explanations (51%), accessibility to students (51%), clarity of objectives (49%), assignment of grades (44%), ability to stimulate thinking (42%), encouragement of class participation (39%), preparedness for class (37%), and organization (34%). The overlap between these items and the UBC UMIs is noteworthy. At least five of the six UMIs map directly on to concepts tapped by many of the schools of pharmacy in the United States, Canada, and abroad.

Combined, these data show that student feedback is a standard feature of teaching evaluation systems in higher education. Moreover, there is some consistency in the types of items asked of students, both within disciplines across schools (Barnett & Matthews, 1998), and across disciplines within schools in Canada (Hardy, 2003; "Summary," 2007).

Current Use of Student Evaluations of Teaching at UBC

Students evaluate teaching all across the UBC campus. These evaluations are used for both formative and summative processes. In terms of summative evaluation, the current requirements for reappointment, tenure, and promotion include demonstration of effective teaching rather than the popularity of the instructor ("Agreement on Conditions of Appointment for Faculty," 2006-10, Asrt. 4.02). The Agreement provides that methods of teaching evaluation may vary and can include student opinion, providing that where student opinion is sought it must be done through formal procedures. The University's *Guide to Promotion and Tenure Procedures at UBC* (2007/08) states that "the evaluation of teaching should include both peer and student evaluations" ("Guide," 2007/08, Section 2.5.2), suggesting that these should play a role in evaluation of teaching at UBC.

Student evaluations are also used by faculty in formative ways to improve their teaching. The UBC Centre for Teaching and Academic Growth (TAG) advocates reflecting on student evaluations with the aim of improvement ("Appendix C"). Indeed, TAG has recently introduced a series of workshops for faculty members entitled *Wisdom through Reflective Practice: Improving Teaching and Learning by Translating Feedback into Practice.* This series is designed to aid faculty in interpreting their teaching evaluations from multiple sources (e.g., students, peers, and self-assessments) with the goal of improving their teaching. The fact that this series of eight workshops is being offered strongly indicates that faculty are interested in using their evaluations to better their teaching practice.

The Faculty of Arts[3] publishes online, with the consent of participating faculty, the summary statistics for each item of its student evaluations of teaching ("Arts Course Evaluations"). The popularity of the *ratemyprofessors.com* website indicates that students are interested in rating and reading ratings of instructors, and anecdotal evidence suggests that probably the majority of

[3] The Department of Psychology processes its own evaluations and is therefore not represented on this website.

students at UBC consult this latter website. The quality of the data found there, however, is very low, and this fact makes the Faculty of Arts offerings in this regard a substantial improvement.

Summary of How Student Evaluations of Teaching are Generally Seen in the Context of Comprehensive Evaluation of Teaching at UBC and Elsewhere

Student evaluations of teaching are consistently promoted as one part of the set of methods to evaluate the many facets of teaching (Johnson & Ryan, 2000; Pounder, 2007). The University's *Guide to Promotion and Tenure Procedures at UBC* ("Guide," 2007/08, Section 2.5.2) and the UBC Centre for Teaching and Academic Growth ("Appendix C") advocate the evaluation of teaching from multiple perspectives, including, but not limited to, student and peer assessments. Many other universities require at least both peer and student evaluations of teaching for tenure and promotion, with some requiring comprehensive teaching dossiers that demonstrate reflection on these evaluations among other demonstrations of teaching effectiveness (e.g., McGill University, see "Regulations"; University of Guelph, see "Teaching Dossier"; University of Toronto, see "Provostial Guidelines"). Overall, student evaluations of teaching are viewed as one important component of a more complete system of teaching evaluation.

SECTION 3 – DESCRIPTION AND ASSESSMENT OF THE UMIS AND THEIR ADMINISTRATION

THE SIX UNIVERSITY-MODULE ITEMS AND SOME PSYCHOMETRIC RESULTS WITH THEM

For reference purposes, the six UMIs are given below. These items ask for students to rate their instructor on a 5-point scale with the following anchor points: (1) Very Poor; (2) Poor; (3) Adequate; (4) Good; and (5) Excellent.

UMI 1. The *clarity* of the instructor's expectations of learning.

UMI 2. The instructor's ability to *communicate* the course content effectively.

UMI 3. The instructor's ability to *inspire* interest in the subject.

UMI 4. The *fairness* of the instructor's assessment of learning (exams, essays, tests, etc.).

UMI 5. The instructor's *concern for* students' learning.

UMI 6. The *overall quality* of the instructor's teaching.

Some empirical findings from the first use of the six UMIs follow. These items were used for the first time at the end of Term 1 (September – December), 2007. It is important to understand that the analyses reported in what follows have all used the class as the unit of analysis. The data points, therefore, are class means, and all means, variances, and correlation coefficients that have been calculated have been based on class item and aggregate means as the basic units of analyses.

Inspection of Item Means

In an analysis of 1,341 sections across eight faculties, the item means (n = 1,341 classes) were: UMI 1: 4.07; UMI 2: 4.10; UMI 3: 4.01; UMI 4: 4.08; UMI 5: 4.22; and UMI 6: 4.12. These means hover around the "Good" anchor point on the scales and speak highly of the general level of teaching at UBC. The psychometric question that could be raised is whether there are ceiling effects operating with these scales. Since the item standard deviations are in the .40 to .60 range, generally, the scales do have sufficient space at the top to allow for relatively symmetric distributions of class mean scores about the overall means reported above. As just one example, for UMI 6 (Overall Evaluation of Instructor's Teaching), the item standard deviation over the 1,341 classes (with mean of 4.12) is .568, allowing a 1.55 standard deviation spread at the top. With some of the other items, this spread is closer to 2.0 standard deviations. This latter fact suggests that the class item means have adequate variability to allow for satisfactory reliability and validity.

Inspection of the item means provided some information about another question that has been raised, namely whether slight changes to the wording of the scale anchor points would have any effect on the results. In the Psychology Department scales, for example several items ask for an evaluative response using the anchor points Very Poor, Poor, Fair, Good, and Very Good, and corresponding Faculty of Arts items use Neutral in place of Fair, with the other anchor words the same. With the UMIs, the corresponding anchors are: Very Poor, Poor, Adequate, Good, and Excellent. With other Psychology Department and Faculty of Arts items, however, students are asked to indicate degree of agreement with statements about teaching. For example, whereas UMI 4 requests a "Very Poor…Excellent" response to: "The *fairness* of the instructor's assessment of learning (exams, essays, tests, etc.)", corresponding items in the Psychology and Faculty of Arts

inventories ask for degree of agreement ("Strongly Disagree...Strongly Agree") to "Evaluation procedures were fair and reasonable to students." One indication as to possible different interpretations in the scales would be whether the item means differed between the two.

Table 1 below contains some parallel items that differ primarily (although slightly in other ways too) from the UMIs because of the response scale used with each. If such differences do have an effect on the resulting scores earned, we should be able to detect this from a comparison of the item means. Results are given in Table 1 for comparable results taken from the Department of Psychology, Faculty of Arts, Faculty of Pharmaceutical Sciences, and Faculty of Land and Food Systems. (We note here that all Psychology Department results throughout this report are independent of those for the Faculty of Arts—that is, are not included in the latter.) These results were chosen because: (a) items that closely corresponded to the UMIs were found on the inventories used by these units, (b) the items—both UMIs and unit-specific items—were administered at the same time to the same students, and (c) the same administration format (paper or online) was used with both the UMIs and unit-specific items. These criteria were seen as necessary to prevent, as much as possible, extraneous factors from influencing the results. One casualty of imposition of the criteria was all the Faculty of Science data, since *only* the UMIs were administered in Fall, 2007 in Science.

Table 1

Comparison of Item Means: UMIs vs. Similar Items Used in Various Administrative Units (Fall, 2007 Administration) Sample Sizes (No. of Classes): Psychology: 41; Arts: 674; Pharmaceutical Sciences: 75; and Land & Food Systems: 53

Note: All UMIs Have 5-Point Scale: Very Poor, Poor, Adequate, Good, Excellent

UMI and Corresponding Item		***Mean***
1. UMI 1: (Rate) The *clarity* of the instructor's expectations of learning		
Psychology – Own Items—Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree		
UMI 1:		4.00
Psych. Item 28	The course requirements were clearly outlined to students.	4.11
Arts – Own Items—Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree		
UMI1:		4.12
Arts Item 6	The course requirements were clearly outlined to students.	4.20
Pharm. Sc's– Own Items—Strongly Disagree, Disagree, Undecided, Agree, Strongly Agree		
UMI1:		4.25
Pharm. Item 1	The instructor provided clear learning objectives.	4.26
Mean UMI 1 Score over Above Administrative Units:		*4.13*
Mean Score on Corresponding Items over Above Units:		*4.20*

(Table continues.)

Table 1 (Continued)

UMI and Corresponding Item		*Mean*
2. UMI 2: (Rate) The instructor's ability to *communicate* the course content effectively		
Psychology – Own Items—Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree		
UMI 2:		4.16
Psych. Item 13:	The instructor spoke with effective pacing, pitch, clarity and volume.	4.18
Arts – Own Items—Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree		
UMI 2:		4.25
Arts Item 10:	The instructor communicated at an appropriate level for the class.	4.27
Arts Item 11:	The instructor was articulate and communicated effectively	4.27
Pharm. Sc's– Own Items—Strongly Disagree, Disagree, Undecided, Agree, Strongly Agree		
UMI 2:		4.22
Pharm. Item 2:	The instructor provided the material in an organized manner.	4.22
Pharm. Item 3:	The instructor taught at a level appropriate to students' abilities.	4.28
Pharm. Item 4:	The instructor provided the material in a clear and understandable fashion.	4.18
Mean UMI 2 Score over Above Units:		*4.24*
Mean Score on Corresponding Items over Above Units:		*4.26*
3. UMI 3: (Rate) The instructor's ability to *inspire* interest in the course		
Psychology – Own Items—Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree		
UMI 3:		4.01
Psych. Item 25:	As a result of the instructor's efforts, you became more interested in the subject.	3.74
Arts – Own Items—Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree		
UMI 3:		4.10
Arts Item 15:	As a result of the instructor's effort, you became more interested in the subject.	3.97
Land & Food Systems – Own Items—Strongly Disagree, Mildly Disagree, Neutral, Mildly Agree, Strongly Agree		
UMI 3:		3.87
L&F S. Item 8:	The course stimulated my interest in the subject.	3.88
Mean UMI 3 Score over Above Units:		*4.08*
Mean Score on Corresponding Items over Above Units:		*3.95*

(Table continues)

Table 1 (Continued)

UMI and Corresponding Item		*Mean*
4. UMI 4: (Rate) The *fairness* of the instructor's assessment of learning (exams, essays, tests, etc.)		
Psychology – Own Items—Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree		
UMI 4:		3.75
Psych. Item 2:	Evaluation procedures were fair and reasonable to students.	3.79
Arts – Own Items—Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree		
UMI 4:		4.13
Arts Item 2:	Evaluation procedures were fair and reasonable.	4.12
Pharm. Sc's– Own Items—Strongly Disagree, Disagree, Undecided, Agree, Strongly Agree		
UMI 4:		4.13
Pharm. Item 10:	The instructor's examination questions or other evaluations were consistent with the stated learning objectives.	4.16
Mean UMI 4 Score over Above Units:		*4.11*
Mean Score on Corresponding Items over Above Units:		*4.11*
5. UMI 5: (Rate) The instructor's *concern* for students' learning.		
Psychology – Own Items—Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree		
UMI 5:		3.98
Psych. Item 3:	The instructor was patient when students requested course-related assistance.	4.13
Arts – Own Items—Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree		
UMI 5:		4.24
Arts Item 7:	The instructor was helpful when students requested course-related assistance outside of class.	4.27
Arts Item 9:	The instructor was readily available to students either through regular office hours or by appointment.	4.28
Pharm. Sc's– Own Items—Strongly Disagree, Disagree, Undecided, Agree, Strongly Agree		
UMI 5:		4.27
Pharm. Item 7:	The instructor provided sufficient time for outside-of-class consultation.	4.05
Land & Food Systems – Own Items—Strongly Disagree, Mildly Disagree, Neutral, Mildly Agree, Strongly Agree		
UMI 5:		4.21
L&F S Item 10:	Out-of-class assistance was available.	4.10
Mean UMI 5 Score over Above Units:		*4.23*
Mean Score on Corresponding Items over Above Units:		*4.26*

(Table continues.)

Table 1 (Continued)

UMI and Corresponding Item		*Mean*
6. UMI 6: (Rate) The *overall quality* of the instructor's teaching		
Psychology – Own Item: Very Poor, Poor, Fair, Good, Very Good		
UMI 6:		4.08
Psych. Item 31	Considering everything, how would you rate your instructor?	4.19
Arts – Own Item: Very Poor, Poor, Neutral, Good, Very Good		
UMI 6:		4.28
Arts Item 20	Considering everything, how would you rate your instructor?	4.39
Mean UMI 6 Score over Above Units:		**4.27**
Mean Score on Corresponding Items over Above Units:		**4.38**
7. Overall Unweighted Means over Above Units		
UMIs		**4.18**
Corresponding Items		**4.19**

When the results in Table 1 are examined in their totality and in some detail, it is very hard to see any consistent effects arising from the wording of the scale anchor points. It is true that with some specific cases (such as with UMI 3, for example), there appears to be some difference in mean levels, but this may just as easily have arisen from a slightly different understanding of the item content. Had the other UMIs shown the same trends as UMI 3, we might have suspected the scale-point wording to have had some effect, but we really do not see this, to any extent, with the other UMIs. This lack of an effect is particularly noticeable when the means for each UMI are compared with those for the corresponding items, where small differences wash out—a phenomenon seen most clearly in the overall unweighted means of the means given at the very bottom of the table. In some cases, these UMI vs. corresponding item mean differences are statistically significant, but this is generally a case of statistical, but not practical, significance resulting from very large sample sizes (such as the 674 classes in the Faculty of Arts sample). When we calibrate the mean differences in the metric of standardized effect sizes, most of the apparent effects fall below about .10, with a few rising to .20. On the basis of these results, we consider the matter of scale-point wording to be of little or no importance. The five anchor-point words used with the UMIs appear to us to be perfectly fine and more-or-less interchangeable with other possibilities.

It could be argued that an examination of variances, as well as means, is warranted. Although item variances are of little descriptive value, and, further, there was no *a priori* reason to expect them to vary between the two item forms compared, we nonetheless compared all variances for the items in the Department of Psychology and the Faculty of Arts. We should note that these are statistically very powerful comparisons, benefiting as they do from very high correlations between corresponding items (these correlations appear in Table 4). For the Department of Psychology items, none of the item variances differed between the UMI and corresponding Psychology item.

With the Faculty of Arts items, two of the seven item comparisons (UMI 2 was compared with two Arts items) yielded statistically significant differences at the .01 level between the compared variances. This result was expected given the very high correlations between item pairs (in the .80s) and enormous *n* (674). Still, the practical (as opposed to statistical) significance of these variance differences was very low, with the largest discrepancy between standard deviations 1.18. That is, in the case of the largest variance difference with Arts items, the standard deviations were in the ratio of 1.18:1. For the other significant difference, the SDs were in the ratio of 1.09:1. These variability differences are of no consequence.

Inspection of Item Content

An important question here is whether the item content represented by these six items is (a) ***central*** to teaching, as this is understood by university teachers and (b) sufficiently ***comprehensive*** to cover the important aspects of teaching. Certainly, with a limitation to only six UMIs, one of which is an overall, omnibus item, the task of ensuring that these would be sufficiently inclusive and diverse was a challenging one. The logistics in force here have required a small number of university-wide items.

(a) Centrality. An inspection of the item content suggests that these six items do cover important aspects of effective teaching. The themes represented by these items have been found on most teaching evaluation inventories I have surveyed. (Earlier, on Page 7, we noted how closely these items mapped onto mainline items found in a number of schools of pharmacy. It is interesting to note in passing that schools of pharmacy in North America have led the way in conscientious evaluation of teaching, a characteristic found also in the UBC Faculty of Pharmaceutical Sciences.) Further, the inclusion of one omnibus evaluation item can be seen as, in a sense, making up for the fact that, because of the small number of items, some specific aspects of teaching could not be covered in these university-wide items. This set of items can be seen as reasonable, including as it does questions involving the effectiveness of presentation, a clear setting out of learning objectives, fairness in student evaluation, inculcating of interest and excitement, and respect for students, all themes that, I think, would be seen as central to the teaching enterprise by almost anyone.

Thus, from the perspective of content validity (as well as face validity—or apparent content validity) it seems hard to fault this set of six items. They appear very straightforward, easily understandable, and concerned with what would be expected in a small set of items for use in evaluating a university course. The UMIs are not all stated in strictly behavioural terms, like most of the Psychology Department inventory items, instead soliciting an evaluation of the instructor's *ability* to *communicate* and *inspire* and his/her *level of concern* for students. (UMIs 1, 4, and 6 do not have this slippage between observable behaviour and the assessment of an ability or affective state.) Still, if (a) the items so worded correlate highly with corresponding behaviourally-worded items and (b) the item means are similar, there would seem to be no cause for concern about this particular choice of wording.

In the aforementioned sample of 41 Psychology classes, the correlations noted above were very high: (a) UMI 2 (communication) and the corresponding behaviourally-worded Psychology item: *r = .86*; (b) UMI 3 (inspiration) and corresponding behavioural item: *r = .94*; and (c) UMI 5 (concern for students) and corresponding behavioural item: *r = .85*. It should be kept in mind, in connection with these correlations, that not only did the correlated items differ in whether or not they were

behaviourally-worded, but they differed slightly in content. Example: UMI 5: (rate) "The instructor's concern for students' learning" and Psychology inventory Item 3: (strongly disagree to strongly agree) "The instructor was patient when students requested course-related assistance." These high correlations, when paired with the highly-similar means of these items (as seen several paragraphs back), suggest that the items are being responded to in extremely-similar ways and that the inferences called for are close enough to observable behaviour not to be seen as problematic.

(b) Comprehensiveness. Another perspective on the relevance of these items can be seen in their relationships to some teaching-evaluation *factors* derived years ago in the Psychology Department. The 32 items on the Department's Student Evaluation Inventory were factor-analyzed. Again, the class was the unit of analysis, and the data points were item means for the classes, approximately 500 in total. The 32 items had been fine-tuned over the years in the Department and were representative of all the important aspects of teaching. The domain of these aspects was well represented. Factor analysis revealed three broad factors of teaching effectiveness:

> *Factor 1, Instructor Competence* (involving the procedural aspects of teaching, such as preparedness, organization, clarity of presentation, use of effective examples and teaching methods, fairness of exams, etc.);
>
> *Factor 2, Respect for Students* (involving accessibility to students, establishment of rapport, respect for gender, race, ethnicity, etc.); and
>
> *Factor 3, Academic Standards and Motivation of Students* (involving setting of high standards of learning, getting students interested in the material, motivating students to do their best, etc.).

We might expect UMIs 1, 2, and 6 to correlate with Factor 1, *Instructor Competence*, since they tap into procedural aspects of teaching competence. The correlations, on a sample of 41 Psychology classes in Fall, 2007, were: UMI 1: **.90**; UMI 2: **.958**; and UMI 6: **.920**. Similarly, we might expect UMI 5 to correlate highly with Factor 2, *Respect for Students*; the obtained correlation in this same sample of 41 classes was **.877**. Finally, we might expect a substantial correlation between UMI 3 and Factor 3, *Academic Standards and Motivation of Students*; this correlation was **.871**.

The above correlations are very high, indicating that the UMIs measure the intended content. Further, these correlations indicate that the six UMIs effectively map onto the three major factors found via factor analysis and that the items were understood by students in the intended way. These correlations also provide some evidence of sufficient comprehensiveness for the six UMIs. More information about what these items measure will be provided later in a discussion of validity.

Reliability

As is well known, the notion of "reliability" of measurement can be understood in different ways. One conceptualization of the term—that of stability of measurement—requires, for its assessment, the administration of a measuring instrument on two occasions, separated by a reasonable time period. I believe that this meaning of reliability is the most intuitive and widely-understood. Another conceptualization of reliability is the extent to which the individual parts of an instrument "hang together" or are related to each other. With the present data we could inquire about the

extent to which the individual UMIs correlate with one another and, thus, how "internally consistent" their sum is. Assessment of this conceptualization of reliability requires only one administration of the instrument. Still another conceptualization of reliability (in a sense similar to internal consistency) concerns the extent to which two alternate measures of the same construct correlate. In the present case, we have data available that can provide information about two of the three forms of reliability, and can give us some insights about the third.

(a) Stability. This is sometimes referred to as "test-retest" reliability because it is assessed by correlating scores obtained for the measure on two different occasions. Since the Fall term, 2007 is the first time the UMIs have been administered, we do not strictly have the wherewithal to assess stability. We do, however, have data from two administrations, separated by one year, of measures that are very close in content. The logic here, then, is to use these correlations as lower bounds to estimates of the long-term stability of the UMIs and their aggregate.

To be more specific, we have results with items administered in Fall, 2006 from existing departmental inventories and with similarly-worded UMIs administered to the same course/ instructor combinations in Fall, 2007. A correlation between these two entities (Item X in 2006 and UMI Y in 2007), therefore, should shed some light on the likely test-retest stability of UMI Y. Since the items being correlated are not identical, however, we might expect a slightly lower correlation in this data-analytic setup than in a true test-retest design. Further, since the 2006 results are from paper administration, whereas those from 2007 are from online administration, we also have this difference threatening to lower slightly the correlations. What follow are some relevant correlation coefficients. We note here that the sample size with respect to the Psychology Department results for these correlations is *very* small, 18 classes. Thus, these correlations should be regarded with caution until corroborated by results based on much larger samples. The results presented below for the Faculty of Science data, however, are based on a sample of 124 classes, and as a result, can be seen as more solid. Results for these two administrative units appear below in Table 2.

Table 2

*Correlations between UMI Items Administered Online, 2007 and Corresponding Psychology Department and Faculty of Science Items Administered in Paper Format, 2006 (Also Included for Comparison are Correlations between Paper and Online Psychology Department Items) (*ns*: Psychology: 18 classes; Science: 124 classes)*

Items Compared			r_{xy}
1. UMI 1: (Rate)	The *clarity* of the instructor's expectations of learning.		
Psych. Item 28:	The course requirements were clearly outlined to students.	*(nonsignificant)*	**.33**
Psych. Item 28:	paper-format in 2006; online-format Item 28 in 2007.	*(nonsignificant)*	.24
2. UMI 2: (Rate)	The instructor's ability to *communicate* the course content effectively.		
Psych. Item 13:	The instructor spoke with effective pacing, pitch, clarity, and volume.		**.68**
Science Item 1:	Instructor presented material in a clear & understandable way.		**.70**
Psych. Item 13:	paper-format in 2006; online-format Item 13 in 2007.		.68

(Table continues.)

Table 2 (Continued)

Items Compared			r_{xy}
3.	UMI 3: (Rate)	The instructor's ability to *inspire* interest in the course material.	
	Psych. Item 25:	As a result of the instructor's efforts, you became more interested in the subject.	**.79**
	Psych. Item 22:	As a result of the instructor's efforts, students became motivated to do their best.	**.78**
	Science Item 2:	Instructor presented material in an interesting manner.	**.76**
	Psych. Item 25:	paper-format in 2006; online-format Item 25 in 2007.	.68
	Psych. Item 22:	paper-format in 2006; online format Item 22 in 2007.	.51
4.	UMI 4: (Rate)	The *fairness* of the instructor's assessment of learning (exams, essays, tests, etc.)	
	Psych. Item 2:	Evaluation procedures were fair and reasonable to students. *(nonsignificant)*	**.06**
	Psych. Item 2:	paper-format in 2006; online-format Item 2 in 2007. *(nonsignificant)*	–.11
5.	UMI 5: (Rate)	The instructor's *concern* for students' learning.	
	Psych. Item 3:	The instructor was patient when students requested course-related assistance.	**.68**
	Science Item 3:	Instructor was receptive to questions.	**.62**
	Psych. Item 3:	paper-format in 2006; online-format Item 3 in 2007.	.79
6.	UMI 6: (Rate)	The *overall quality* of the instructor's teaching.	
	Psych. Item 31:	Considering everything, how would you rate your instructor?	**.74**
	Science Item 6:	Instructor taught effectively.	**.71**
	Psych. Item 31:	paper-format in 2006; online-format Item 31 in 2007.	.74
7.	***Mean** of the 6 UMIs* (2007) and ***Mean** of the 7 corresponding Psych. Items (2006):*		***.75***
	***Mean** of the 6 UMIs (2007) and Mean of the 4 corresponding Science Items (2006):*		***.70***
	Mean of the Psych. items online-format, 2007 vs. Mean of the Psych. items paper-format, 2006:		.77
8.	***Mean*** r_{xy} between UMI 2007 online-format and Psych. item 2006 paper format:		$\bar{r}_{xy} =$ ***.58***
	Excluding UMI 4 and Psych. Item 2:		$\bar{r}_{xy} =$ ***.67***
	Mean r_{xy} between UMI 2007 online-format and Science item 2006 paper format:		$\bar{r}_{xy} =$ ***.70***
	Mean r_{xy} between Psych. Item 2007 online format and Psych. Item 2006 paper format:		$\bar{r}_{xy} =$ ***.56***

Two points should be raised about the results in Table 2. First, with the time lag involved between measurements (one calendar year), the correlation coefficients could, perhaps, be better understood as long-term stability coefficients than as test-retest reliability coefficients. In assessing

reliability, however, it is normally the latter that are used. There is a negative correlation between magnitude of retest correlation and the time elapsed between administrations (one estimate of this I have seen is –.34, although the relationship is not really linear). Meta-analyses of long-term stabilities have shown considerable reductions in magnitude of correlation over extended times. Viswesvaran and Ones (2000), for example, showed average 19.5-month stabilities for the Big Five personality traits of .73 (on the basis of 170 samples comprising 41,000 subjects), whereas their shorter-term test-retest reliabilities have been estimated to be closer to .85. Schuerger and Witt (1989) presented (on the basis of 79 samples) 12-month stabilities for individually-assessed IQ scores of .85, compared with 1-month values (test-retest reliabilities) of .92. Using these proportions, we might multiply the tabled values by approximately 1.08 to get closer to conventionally-estimated test-retest values.

A second point is that the UMIs had slightly different wording than did the corresponding items used in Science (prior to the Fall, 2007 term) and in Psychology. Interestingly, this fact seems not, in general, to have lowered the stability coefficients obtained, as can be seen from comparisons between the UMI vs corresponding-item correlations, on the one hand, and the corresponding-item vs corresponding-item correlations, on the other, over the 12 months. In the majority of cases, the former correlations were, contrary to expectations, slightly higher than the latter. Also worth noting is that, not only was the wording slightly different, but the administration format was different as well—paper-format for the 2006 results, online for the 2007. Thus, there was this additional source of error variance operating to reduce these correlations (in addition to the time factor).

Clearly, then, the values in Table 2, although the best we can arrive at with currently-available data, can be seen as underestimates of the true test-reliability of the UMIs and their composite (mean) score. With the exception of two UMIs, the long-term stability estimates are quite impressive, and with the necessary corrections for the extraneous factors noted above, would be well within the completely-acceptable range. On the basis of these results alone, we might expect test-retest estimates for individual UMIs near or above the .80 mark for four of the UMIs, and in the mid-.80s for the composite score. Individual item reliabilities are, of course, generally low, and the values estimated for the UMIs would be near the upper limit of the range of expected values. The value estimated for the composite would compare favourably with test-retest reliability estimates for inventory scales consisting of far more items than the six here.

The two exceptions to the above—generally favourable—assessment are UMIs 1 (clarity of expectations) and 4 (fairness of evaluation). However, it can be seen from Table 2 that this phenomenon exists (and, if anything, to a greater degree) with the corresponding Psychology Department items, and, for this reason, cannot be seen as a problem specifically in the UMI items. (We should note that in the Psychology inventory, this item actually appears in reflected form, as "Evaluation procedures were *un*fair and *un*reasonable to students," so as to avoid response sets; this fact, however, cannot be expected to produce the results obtained here.) It would appear that there is something in the assessment of clarity of expectations and, particularly, assessments of *fairness of the evaluation procedures* that is, perhaps, somewhat transitory or otherwise unstable. In this connection, I have noticed that tests and other evaluation procedures in which individual students might do poorly are often seen as unfair by these same students. Since, however, we are using the class/instructor unit as the unit of analysis, it is not entirely clear how this phenomenon would affect these correlations, but just asking students about the fairness of evaluation procedures might well yield data that is contaminated by their own performance. In other words, it

might prove impossible to get highly-stable assessments of perceived fairness. With respect to UMI 1, however, it seems at least possible that the clarity of "the *clarity* of the instructor's expectations of learning" is not satisfactory! These two UMIs likely deserve further study and possibly slight rewording, and recommendations for further work with the UMIs are made later in this report.

(b) Internal Consistency. Internal consistency reliability, most commonly assessed by Cronbach's coefficient alpha, indexes the extent to which the items of a composite are related to one another. The logic is that, if a set of *k* items are all highly correlated, then they are likely all measuring a single construct. A high internal consistency value indicates that the *k* items are yielding a consistent assessment of the object of measurement (in this case the course/ instructor unit) and that their composite represents a good basis for describing the object of measurement. A low value would indicate that there do not appear to be good reasons to aggregate the information obtained in the individual *k* items. Internal consistency estimates (like all estimates of test reliability) are partly a function of *k,* the number of items.

In addition to assessing the internal consistency of a set of *k* items, we can obtain, from this assessment, an estimate of the reliability of a single item—that is, any one (generally) of the *k* items. These internal-consistency estimates are presented in Table 3 below, where we have calculated them for both the aggregate of all six items, as well as for the first five (omitting the more general, omnibus UMI 6). From both of these estimates, we have estimated the reliability of a single UMI. We note in this connection that these latter estimates refer to the reliability of a single item, in general, not to any one of the UMIs specifically.

Table 3

Estimates of Internal-Consistency for the UMI Total Scale (Comprising Six UMIs or Five, Omitting UMI 6) for Nine Administrative Units, and Estimates of the Reliability of a Single Item

	Internal Consistency Estimate		***Est. of Reliability***
Admin. Unit (# Classes in Paren's)	***6-Item Sum***	***5-Item Sum***	***of a Single item***
Psychology (41)	.93	.89	.66
Land and Food Systems (52)	.96	.94	.78
Law (104)	.94	.92	.71
Education (240)	.96	.95	.80
Science (620)	.97	.95	.82
Pharmaceutical Sciences (74)	.97	.95	.82
Business (144)	.93	.91	.68
Forestry (54)	.96	.94	.78
Arts (672)	.96	.94	.78
Mean (2,001)	**.96**	**.94**	**.78**

(Table continues.)

Table 3 (Continued)

Weighted-Average Correlations between Each UMI and the 6-item Composite UMI Mean Measure

UMI	***6-Item Composite***
1	.91
2	.90
3	.88
4	.76
5	.87
6	.96

We should, at the outset, note that items like the UMIs used to evaluate an object of measurement (like a course/instructor unit) can be expected to manifest "halo effect" in which their intercorrelations are somewhat inflated because of an overall feeling about the class that influences evaluations of the more specific aspects referenced by each item. This effect, which is unavoidable in such measurements, also inflates the magnitude of the internal-consistency estimates of such instruments, and thus should be seen as accounting, to some extent, for the exceedingly high alpha coefficients in Table 3. Since UMI 6 is an overall, or global, assessment item, we evaluated the internal consistency of the remaining five both with and without UMI 6 present in the analysis. As can be seen, the presence or absence of UMI 6 in the analyses made very little difference, although the difference is in the predictable direction.

Still, the values in Table 3 are impressively high (all, of course, statistically significant), even taking into account halo effect. This is best seen from the means, given at the bottom of the first panel of Table 3, over the nine administrative units and 2,001 instructor/class units.

Although there will be a general overlay of positive correlation operating among items like the UMIs, there is also, with good inventories, some pattern of differences among the correlations that indicate that some discriminant validity also exists in the data. Discriminant validity is the property of (in this case) items correlating somewhat differently with differing underlying themes running through the inventory. Put more simply, an item with discriminant validity will correlate more highly with the theme it was designed to measure than it will with other themes. Our hope, of course, with inventories such as that containing the six UMIs, is that the items will possess some discriminant validity and that this will not be completely submerged under halo effect.

With a small number of items, it is not possible to search for the underlying themes among the UMIs, since, for the most part, only one item has been included for each theme. In longer inventories, these underlying themes can be identified by factor analysis. I noted earlier the factor analysis of the Psychology Department inventory, in which three large, molar themes were identified. The 32 items aligned themselves with these themes and manifested larger correlations with the intended theme than with the other two themes, thus displaying some degree of discriminant validity. Since these items are very similar to the UMIs, we can, I believe, assume that the UMIs are adequately tapping their intended theme domain, and that, had the inventory been longer, these themes would have emerged in the form of common factors.

It would perhaps be helpful for the reader to have some perspective on typical values of coefficient alpha for various kinds of instruments used in the behavioural sciences. Cognitive-ability measures often have alpha coefficients in the .80 to .90 range. Personality scales tend to have slightly lower internal consistencies, typically ranging from .70 to the mid- to high-.80s.

One other finding—presented in Table 3—concerns the extent to which each UMI correlates with the whole (the aggregate of the six UMIs, or the UMI Mean). These correlations are given in the lower panel of Table 3 and represent the weighted-average r_{xy} of each item (across the nine administrative units) with the total or mean UMI score with that item removed from the total for the correlation. This latter adjustment is to prevent part-whole correlation from artificially inflating the obtained correlation coefficients. Not surprisingly, the largest item-total correlation is for UMI 6, the overall or global assessment item. Of the remaining five items, four manifest fairly similar correlations with the whole, but one of them appears to be significantly lower in this regard—UMI 4 (the *fairness* of the instructor's assessment of learning—exams, essays, tests, etc.). Interestingly, this same item manifested by far the lowest long-term stability of the six UMIs, as seen in Table 2.

I will refrain from speculating about the reasons for this phenomenon, although I will suggest that "fairness" may be too subjective an attribute of which to obtain good, replicable measurement. One suggestion in this regard would be to consider rewriting this item to read something like: "The extent to which assessment of learning (exams, papers, etc.) was consistent with what students were led to believe" (if that is the theme about which evaluation is desired) or "The extent to which the number and difficulty of learning assessments (exams, papers, etc.) were appropriate for the course" (if *that* is what we want to measure). Some rewording of the items is one of a number of recommendations made at the end of this report.

(c) Inter-Rater Reliability. One form of reliability often considered in connection with ratings (like those students provide on their instructor's teaching performance) is inter-rater (or inter-judge) reliability, which indexes the extent to which raters (generally a small number) are consistent in their evaluations (or, we might say, exhibit internal consistency). Since the concern here has been with means—across *large* numbers of ratings for instructors—and because there is a lawful relationship between the number of data points composing a mean rating and its reliability, the topic of inter-rater reliability has been ignored so far. I do not consider inter-rater reliability to be of great concern in the present context, but brief consideration of it might be helpful.

I have experience with what is termed *360° assessment* in industry. This form of assessment—generally of management personnel—is very widely carried out in all businesses and organizations these days, and is termed 360° assessment because it involves ratings of managers from their direct reports, peers, and bosses, and hence, represents assessment from all directions. In application, each manager is normally evaluated by anywhere from 4 to 6 or so raters, and the mean of these ratings is taken as that manager's "score" on a particular dimension. With 360° assessment, we really do have concerns about inter-rater reliability because of the small number of raters. In our analyses over many years, I have seen management assessment results with inter-rater reliability coefficients in the low .50s or below (being based on only two or three ratings for each manager), constituting real cause for concern about the value of the measurements.

Empirical assessment of inter-rater reliability. We can view our student evaluation process as very similar to 360° assessment in many ways (except that our process is more like the "bottom-up" facet of the full 360° process). What makes our form of evaluation greatly superior to 360°

assessment in almost all other organizations, however, is the large number of respondents that most classes comprise. To confirm that this fact yields acceptable levels of inter-rater reliability, we conducted such an analysis using standard analysis of variance techniques. We took 31 sections of courses in Psychology with at least 50 student raters in each (2007 online administration) and assessed the inter-rater reliability for one of the UMIs—UMI 6, the global item—for a class size standardized at 50. (In these analyses, the individual rater became the unit of analysis.) Our obtained value was **.93**. The corresponding value for a class size of 25 is .87, and for a class of 15 raters, .80. These values fully confirmed our expectations in this regard (and were consistent with what we would expect with 360° assessment in industry with correspondingly large samples), and thus no further examinations of inter-rater reliability were undertaken.

At 10-15 student raters, our mean item scores possess adequate inter-rater reliability. With 30 or 40 student respondents, these values will, as we have seen, approach the .90s and be more than adequate. With very small classes, however, comprising fewer than 8–10 students, we do need to be aware of the potential for low inter-judge consistency in the ratings.

(d) Summary of reliability assessments. The evidence we have at present about the performance of the UMIs in this their first use is very encouraging. This is not surprising since they appeared, from the outset, to be well-written statements about key aspects of effective teaching, and were thus expected to perform as other good teaching-evaluation items have done in the past. Years ago, I supervised extensive analyses of the 32 items on the Psychology Department inventory and found similar item-performance results. The UMIs, as noted, are very similar to those on the Psychology inventory and could have been expected to perform at least as well. The six UMIs appear to have good long-term stability (with the one exception noted earlier, this exception being found also for the corresponding Psychology-inventory item), and are internally consistent.

A needed follow-up study to the present analyses is one involving both short-term and long-term stability of the UMIs, now possible with subsequent administrations of these items. It seems very likely that some fall-to-spring results will be able to be obtained to assess 3- to 4-month stability, although these will be far fewer than fall-to-fall or spring-to-spring data sets. In the present analyses, we have had to correlate the UMIs with similarly-worded items administered on paper one year earlier. With the present data on the Fall, 2007 administration available, more refined assessments of the item stabilities will be possible—assessments in which both the items and the administration format are identical, with the time lag as the only apparent source of error.

Validity

The validity of a measuring instrument refers to the extent to which the instrument is actually measuring what it was intended to. Earlier, what might be termed the *content validity* of the UMIs was discussed via a consideration of the extent to which they were—at least on their face—central to conventional notions of teaching and comprehensive in the sense of being sufficiently representative of the overall practice of teaching.

Other forms of validity evidence exist, and in the present context, perhaps the most meaningful would result from an examination of the extent to which scores on the UMIs correlate with scores on other items or aggregates that are putatively measuring similar content. (It should be noted here that some might consider these correlations as indices more of *alternate-forms reliability* than of validity. There is often a fine line between the two, and in the present discussion, I will, perhaps

a little arbitrarily, characterize these correlations as providing evidence of validity.) We have already seen some correlational evidence in the earlier demonstration that the UMIs do correlate in expected ways with the three molar factors identified in the Psychology Department factor analysis. In this section, additional results are presented that bear on the question of just what the UMIs are measuring. This kind of validity evidence has in the past been referred to as *construct validity*, since we are interested in the extent to which the measures being evaluated assess identifiable constructs—in the present case, the important themes underlying teaching.

An effective way of establishing validity would be to correlate UMI scores with independently-determined assessments of the various aspects of teaching. This research design is obviously not available in the present case. Instead, we have UMI-based assessments obtained at the same time as those obtained by other instruments and provided by the same students. The long-term stability results presented earlier in Table 2 do avoid the problems of having responses obtained at the same time, in the same context, and with the same halo effects operating—all of which contribute to inflation of correlations due to what is called "method variance." Still, these long-term correlations are attenuated by other problems, such as the passage of time introducing error and the possibility of function fluctuation (or changes incorporated by the instructor). Still, these correlations can be seen as representing lower-bound validity estimates for UMIs 2, 3, 5, and 6, and their aggregate.

(a) Some empirical results concerning validity. In Table 4 below, some correlates are presented of each of the UMIs from the data collected at the same time (Fall, 2007) via other inventories administered along with the UMIs. In all cases, therefore, we have—between the UMIs and other items—the same time of administration, administration format, and particular set of students completing the evaluations.

Table 4

Correlations between the UMIs and Other Items Measuring Similar Themes at the Same Time (Fall, 2007 Administration), across Several Administrative Units; Numbers of Classes for Each Unit: Psychology: 41; Land and Food Systems: 52; Law: 104; Education: 237; Pharmaceutical Sciences: 74; Forestry: 54; Arts: 673

A. Correlations between Individual UMIs and Similar Individual Items

1. UMI 1: (Rate) The *clarity* of the instructor's expectations of learning

r_{xy}	*Other Item*	*Wording of Other Item*
.80	Psych. Item 28	The course requirements were clearly outlined to students.
.74	Law Item 18	The instructor stated clearly the basis for evaluating students.
.77	Law Item 33	The instructor has made course objectives and requirements clear.
.91	Educ. Item 6	Course objectives were made clear.
.91	Educ. Item 20	Course requirements were clear.
.94	Pharm. Item 1	The instructor provided clear learning objectives.
.80	Arts Item 6	The course requirements were clearly outlined to students.
.82	Weighted Mean of Above Correlations	

(Table continues.)

Table 4 (Continued)

A. Correlations between Individual UMIs and Similar Individual Items (Continued)

2. UMI 2: (Rate) The instructor's ability to *communicate* the course content effectively

r_{xy}	*Other Item*	*Wording of Other Item*
.86	Psych. Item 13	The instructor spoke with effective pacing, pitch, clarity, and volume.
.83	L&F Sys. Item 7	The course included clear and appropriate illustrations and/or practical applications of the subject.
.87	Law Item 11	The instructor's class presentations are effective.
.79	Law Item 15	The material is presented logically and systematically.
.81	Law Item 16	When students are confused the instructor tries to clarify.
.89	Law Item 17	The instructor communicates clearly in class.
.87	Educ. Item 26	The instructor used examples and illustrations that helped clarify topics.
.92	Educ. Item 29	The instructor communicated clearly.
.84	Pharm. Item 2	The instructor provided the material in an organized manner.
.85	Pharm. Item 3	The instructor taught at a level appropriate to students' abilities.
.93	Pharm. Item 4	The instructor provided the material in a clear and understandable fashion.
.89	Forestry Item 1e	(Rate) Effectiveness of classroom presentation.
.85	Arts Item 10	The instructor communicated at an appropriate level for the class.
.93	Arts Item 11	The instructor was articulate and communicated effectively.

.87 Weighted Mean of Above Correlations

3. UMI 3. (Rate) The instructor's ability to *inspire* interest in the subject

r_{xy}	*Other Item*	*Wording of Other Item*
.90	Psych. Item 22	As a result of the instructor's, students became motivated to do their best in this course.
.94	Psych. Item 25	As a result of the instructor's efforts, you became more interested in the subject.
.85	L&F Sys. Item 8	The course stimulated my interest in the subject.
.87	Law Item 2	Considering the nature of the subject matter, the instructor has aroused interest for and enthusiasm in this subject.
.84	Law Item 3	As a result of taking this course, I have become more interested in the subject.
.88	Educ. Item 2	The instructor was enthusiastic about the subject matter.
.90	Forestry Item 1c	(Rate) Ability to engage class.
.91	Arts Item 15	As a result of the instructor's effort, you became more interested in the subject.

.90 Weighted Mean of Above Correlations

(Table continues.)

Table 4 (Continued)

A. Correlations between Individual UMIs and Similar Individual Items (Continued)

4. UMI 4. (Rate) The *fairness* of the instructor's assessment of learning (exams, essays, tests, etc.)

r_{xy}	*Other Item*	*Wording of Other Item*
.89	Psych. Item 2	Evaluation procedures were unfair and unreasonable to students. (Reflected.)
.88	Law Item 19	The basis for evaluation of students as stated is being followed.
.37	Law Item 25	Where mid-term evaluations are used, the evaluation procedures are fair.
.92	Educ. Item 8	Evaluation procedures were fair.
.94	Pharm. Item 10	The instructor's examination questions or other evaluations were consistent with the stated learning objectives.
.84	Arts Item 2	Evaluation procedures were fair and reasonable.
.85	Weighted Mean of Above Correlations	
.87	Weighted Mean without Law Item 25	

5. UMI 5. (Rate) The instructor's *concern for* students' learning

r_{xy}	*Other Item*	*Wording of Other Item*
.85	Psych. Item 3	The instructor was patient when students requested course-related assistance.
.84	L&F Sys. Item 10	Out-of-class assistance was available.
.73	Law Item 10	The instructor deals effectively with questions raised in class.
.67	Law Item 13	The instructor is available to help students outside of class.
.71	Law Item 16	When students are confused the instructor tries to clarify.
.92	Educ. Item 1	The instructor was interested in whether the students learned.
.78	Educ. Item 13	The instructor was available for help outside of class or by email/website.
.80	Educ. Item 14	The class atmosphere was conducive to learning.
.77	Pharm. Item 7	The instructor provided sufficient time for outside-of-class consultation.
.77	Forestry Item 1g	(Rate) Availability for discussion out of class.
.75	Arts Item 7	The instructor was helpful when students requested course-related assistance outside of class.
.69	Arts Item 9	The instructor was readily available to students either through regular Office hours or by appointment.
.77	Arts Item 12	The instructor attempted to answer all questions to the best of his/her ability.
.77	Weighted Mean of Above Correlations	

(Table continues.)

Table 4 (Continued)

A. Correlations between Individual UMIs and Similar Individual Items (Continued)

6. UMI 6. (Rate) The *overall quality* of the instructor's teaching

r_{xy}	*Other Item*	*Wording of Other Item*
.97	Psych. Item 31	Considering everything, how would you rate your instructor?
.94	Law Item 1	Considering all facets of this instructor's teaching, I would rate it as:
.94	Pharm. Item 11	The instructor created a positive classroom atmosphere that promoted learning.
.96	Arts Item 20	Considering everything, how would you rate your instructor?

.96 Weighted Mean of Above Correlations

7. .86 Unweighted Mean of Means over the Six UMIs

B. Correlations between the Aggregate (Sum) of the Six UMIs and Corresponding Aggregates of Similar Items in the Various Administrative Units

(Administrative Unit is Given with the Number of Items in the Unit Aggregate in Parentheses)

.95 Psychology (7)
.91 Land and Food Systems (4)
.93 Law (13)
.96 Education (9)
.96 Pharmaceutical Sciences (7)
.92 Forestry (3)
.95 Arts (9)

.95 Weighted Mean of Above Correlations

Note: All tabled correlation coefficients statistically significant ($p < .0001$).

The correlations in Table 4 are, for the most part, very high, and, as noted earlier, part of the reason is the presence of "method variance." Nonetheless, most of the correlations are at about the maximum that item reliabilities could possibly attain, and thus suggest that the UMIs are measuring—for all intents and purposes—the same themes considered important by the various administrative units and measured by the specific administrative-unit items.

(b) Summary of validity assessment. When the information in Table 4 is combined with that provided earlier in this report concerning item content, there is strong evidence that the UMIs provide mainline teaching assessment—valid assessment of what have been widely identified as central aspects of effective teaching. They are, in all the fundamental ways, parallel to items that the various administrative units on campus have been using for decades, and do not in any way represent a departure from what have always been considered the important themes to assess at UBC.

Brief Summary of What We Know about the Six UMIs

From the examinations of the item content and of their means, variances, and distributional properties, along with the assessments of reliability and validity, it is my opinion that the six UMIs more than adequately perform the function intended for them. They are as strong technically as existing items being used at UBC in the various administrative units, and they are sufficiently general to function well across these units and provide the University with a uniform component (the University Module) that assesses effective teaching. This is not to say that they cannot be slightly improved and, perhaps, augmented by two or three additional UMIs, and I will comment on possible improvements later in this report when making recommendations. I do believe, however, that they are adequate, as is, for their intended purpose.

Online Administration of the UMIs

The University's long-term plan calls for online administration of the UMIs, and they were presented online in some administrative units in the Fall, 2007 term. In these cases, an opportunity was created to examine differences in response rates and mean scores between the two administration formats.

Effects of Administration Format on Response Rates

In Table 5 below, mean response rates are presented for administrative units that gave the UMIs online this past fall, and these are compared with those obtained in these units in the 2006-07 administration, in which the paper format was used.

Table 5

Mean Response Rates to the UMIs, and Other Inventories Administered at the Same Time, for Four Administrative Units Using Online Administration in Fall, 2007, and Corresponding Response Rates for These Same Units Using Paper Administration in 2006-07

Admin. Unit (No. classes in Paren's)	Mean Resp. Rate Online	Mean Resp. Rate Paper
Psychology (41 Online; 99 Paper)	63.30%	63.31%
Land and Food Systems (52 Online; 32 Paper)[a]	60.98	79.24
Law (104 Online; 82 Paper)	65.18	65.91
Science (359 Online: 624 Paper)	66.52	66.40
Wt'd Mean Resp. Rates (556 Online; 837 Paper)	**65.51**	**66.48**

[a]Some of the discrepancy between the two response rates in this case is due to the fact that students who were not registered for the courses (or were registered in cross-listed sections) were able to complete paper evaluations, but only those actually registered for the courses (and were LFS students) completed the online forms.

Of the results reported in Table 5, only the difference between response rates for Land and Food Systems reached statistical significance. When the four administrative units are pooled, the difference (65.51% vs. 66.48%) is nonsignificant, representing an absolute difference of less than 1%, or a relative decline from the 66.48% for paper administration of about 1.5%.

There thus does not appear to be any strong evidence that response rates will be affected by a switch to online administration. My belief in this connection is that, as the number of students who both own computers and are familiar with the Internet reaches 100% (which one would expect to be the case in the very near future), that there will be no decline in response rates due to online administration. In fact, it would not be surprising to see response rates rise. Findings similar to the present ones were, as noted earlier, reported by Hardy (2003), Heath et al. (2007), Johnson (2003), and Kelly, et al. (2007).

Administrative units in a number of universities have used incentives to encourage students to complete the online inventories. In some, where students identify with a particular cohort or year in the program (and all students in a particular class are in the same year), competitions with prizes can be set up between the years. This practice has been used, for example, in the UBC Faculty of Pharmaceutical Sciences, where these conditions obtain. Other incentives include earlier access to term grades than would otherwise be the case or a point or two added to the student's final grade for participating. (Although used at some colleges and universities, this latter practice may be less than ideal, compromising the integrity of the grades awarded.) In addition, a department could award a prize to a participating student, picked at random from all those who completed the inventory. Follow-up e-mails to students who have not completed the evaluations could perhaps be improved and used more effectively than at present. At this point it is not clear just how important such incentives are—or will be in the future—but they may be worth consideration if low response rates are of concern.

The conclusions reached by Layne et al. (1999) are worth repeating here. In their analysis of paper vs. online student evaluations, they noted that, although students reported a preference for the online form, there was some suspicion about anonymity of online responses. Any concerns about anonymity, if present in connection with the UBC online administration of student evaluations, must, of course, be eliminated.

Effects of Administration Format on Scores

Do online assessments result in higher mean scores for instructors? Lower mean scores? There has been speculation around this, with some hypothesizing that only the strongest students will actually respond in the online format, thus, it is reasoned, awarding slightly higher scores to the instructor. Our data related to this question are very limited at present. We begin with some results calculated on Department of Psychology data collected in Fall, 2006 and Spring, 2007, when the Psychology items were administered in paper format and again in Fall, 2007, when the same items were administered online. In Table 6 below, two different sets of data are presented: (a) results on the three Psychology factors for all paper-administered 2006-07 data (both terms; 99 classes) and for all online-administered Fall, 2007 data (41 classes), and (b) results on items that correspond to the UMIs from the Psychology inventory for the paper-administered Fall, 2006 instructor/course combinations on which identical (same instructor/course combinations) data also exist in online format for Fall, 2007. The latter, although providing an excellent comparison, comprise only 18 instructor/course units.

Table 6

Factor, Item, and Aggregated Means for Paper Administration (2006-07; n *= 99 classes) and Online Administration (Fall, 2007;* n *= 41 classes) of the Psychology Department Inventory*

1. Factor and Item Means

Factor or Item	Paper (2006-07)	Online (Fall, 2007)
Factor 1 – Instructor Competence	4.10	4.10
Factor 2 – Respect for Students	4.44	4.42
Factor 3 – Academic Standards & Motivation of Students	3.91	3.92
Mean of the Three Factors	**4.15**	**4.15**
Item 2 – Fairness of evaluation procedures	3.91	3.79
Item 3 – Patience re course-related assistance	4.14	4.13
Item 13 – Effective communication skills	4.08	4.18
Item 25 – Inspiring interest in students	3.65	3.74
Item 31 – Overall assessment of instructor	4.11	4.11
Item 32 – Overall assessment of course	3.73	3.85
Mean of Above Six Items:	**3.94**	**3.97**

2. Item and Aggregated Means for the Same 18 Instructor/Course Combinations in Fall, 2006 and again in Fall, 2007

Item	Paper (2006)	Online (2007)
2 – Fairness of evaluation procedures	3.87	3.76
3 – Patience re course-related assistance	4.13	4.14
13 – Effective communication skills	4.20	4.11
25 – Inspiring interest in students	3.80	3.77
31 – Overall assessment of instructor	4.18	4.16
32 – Overall assessment of course	3.78	3.87
Mean of Above Six Items:	**3.99**	**3.97**
Overall Mean of Above Three Means:	**4.03**	**4.03**

None of the pairwise comparisons in Table 6 are statistically significant—via independent-samples comparisons for the factor and item means in the top panel of the table and by paired-comparison analyses for the items in the bottom large panel. Clearly, there was absolutely no effect whatsoever for online vs. paper administration of the student-evaluation items in these data collected in the Psychology Department with the Psychology inventory. It does not seem like a

stretch to assume that the same close-to-identical results would have occurred with the UMIs had paper-format data been available for them.

Further, in the interests of completeness, we compared the variances of the factors and items in Panel 1 of Table 6. These were independent-samples tests involving results based on 99 classes from 2006-07 and in paper format, and 41 classes in Fall, 2007 in online format. None of the pairs of variances yielded a significant difference, even at the .05 level.

To augment the results in Table 6, we obtained similar data from the Faculties of Land and Food Systems and Law. The Land and Food Systems scales are in a 1–5, Strongly Disagree – Strongly Agree format, but the Law scales run from 1–7, with anchor points running from Disagree Very Strongly to Agree Very Strongly, so that the metric is different from that of the scales considered earlier. This fact, however, is immaterial to the present question. In Table 7 below, results are presented on some items close in content to the UMIs from both their 2006-07 paper and their Fall, 2007 online administrations.

Table 7

Item and Aggregated Means for Paper (2006-07) and Online Administration (Fall, 2007) of Selected Items from the Inventories Used by the Faculties of Land & Food Systems and Law

1. Land and Food Systems (5-Point Scale)

Item	**Paper (2006-07; *n* = 32)**	**Online (2007; *n* = 52)**
Item 1 – Instructor was well prepared	4.38	4.34
Item 2 – Instructor knows the subject well	4.57	4.55
Item 3 – Instructor was interested in teaching	4.34	4.34
Item 8 – Course stimulated my interest in the subject	3.84	3.88
Item 9 – Evaluation feedback was available	3.82	4.11
Item 10 – Out-of-class assistance was available	4.09	4.10
Mean of Above Six Items:	**4.17**	**4.22**

2. Law (7-Point Scale)

Item	**Paper (2006-07; *n* = 82)**	**Online (2007; *n* = 104)**
Item 1 – Overall rating of instructor's teaching	5.67	5.70
Item 2 – Instructor has aroused interest and enthusiasm	5.64	5.86
Item 13 – Instructor available outside of class	5.78	5.84
Item 14 – Students are treated respectfully	6.39	6.39
Item 17 – Instructor communicated clearly in class	5.72	5.88
Item 18 – Instructor stated clearly basis for evaluation	5.57	5.84
Item 19 – Stated basis for evaluation being followed	5.78	5.90
Mean of Above Seven Items:	**5.79**	**5.91**

With respect to the results in Table 7, only one of the mean differences of the 14 presented was statistically significant at the .01 level: that for Law Item 18. There is little, if any, evidence from Table 7 that the two formats can be expected to result in different levels of ratings. If we were to take the differences of the aggregated means at face value, these would represent effect sizes of only .10 – .15, which would be considered extremely small. When the results of Tables 6 and 7 are combined, it is clear that, at least on the basis of available data, there is no reason to believe that the change from paper to online format will have any systematic effect on mean scores received by instructors. These conclusions, based on the preceding data analyses, are consistent with those reached by Layne et al. (1999), Hardy (2003), and Heath et al. (2007) and noted earlier.

Again, for completeness, we did some comparisons of variances. For the six items administered in Land and Food Systems by paper format in 2006-07 and online format in 2007, none of the variance differences reached statistical significance. These results, along with those reported above, for the Psychology Department factors and items, suggest that not only can we expect mean score levels to remain similar from paper to online administration, so too can we expect item and aggregate variances not to change.

Brief Summary of Results Concerning Administration Format

The preceding results of the analyses of response rates and score values suggest that the shift from paper to online administration can be expected to have virtually no effect on either variable. It is important to note, however, that the available data base is very limited at present, and more analyses, based on much larger samples and including many administrative units, are needed to arrive at definitive conclusions about administration format at UBC. Still, the literature is encouraging about both response rate and level of ratings given by students when inventories are administered in an online format. Further, as we have seen, the open-ended comments tend to be both more frequently provided and lengthier in online administration than with the paper format (Heath et al, 2007; Johnson, 2003). Finally, as noted earlier, we can now expect that virtually all UBC students will have either their own computer or access to one, and that they will be sufficiently computer- and Internet-literate to complete the teaching evaluations.

A Very Brief Look at Some Correlates of Student Ratings of Teaching

In the process of reviewing the performance of the UMIs and their online administration, it was possible to collect a small amount of data that may be of interest to some. One question of interest was whether there is a substantial correlation between the scores that instructors receive on their teaching effectiveness and the mean grades they award to their classes (the latter possibly functioning, to some extent, as a determinant of the former). We might expect, for example, that there would be such a positive correlation—with higher class grades accompanying higher evaluations of teaching. Another question of interest was whether there is a substantial correlation between the response rates on the student evaluation inventory and the mean grades that the instructor assigns to the class. We might hypothesize, for example, that the response rate would be higher in classes in which the grades were higher.

There are, of course, several extraneous factors that might affect student-evaluation responses, and some of these have been mentioned in the introduction to this report. Class size is certainly one of these factors. Some feel that the instructor's gender may be such a factor, along with his/her race or ethnicity. These factors may well be too subtle and complex in their operation to attempt to

understand—insofar as their effects at UBC—without a thorough and well-designed study. There are relevant findings from the literature on teaching evaluation, of course, and whether further investigation into these factors at UBC is warranted is beyond the scope of this report. Since we did have easy access to gender data in the Psychology Department, however, we did conduct a very small analysis of the possible effects of instructor's gender in that department.

With regard to this last point, we compared the class means on the six UMIs (and their sum) for the 41 classes evaluated in Fall, 2007—23 taught by women and 18 by men. On none of the UMIs was there a statistically significant difference in the mean scores. For the sum of the six UMIs, the means were: (a) 23 female instructors: *3.97*; (b) 18 male instructors: *4.04*—again a nonsignificant difference. These findings are presented for interest only and not as, in any way, definitive. As noted, the matter of instructor gender is complicated, involving as it undoubtedly does, also the gender of the students and other contextual factors, and the conflicting findings in this connection have been noted earlier (e.g., Feldman, 1993; Kierstead et al., 1988; Sinclair & Kunda, 2000).

In Table 8 below, some results are presented on class grades, mean levels of ratings, and student response rates.

Table 8

Partial Correlations between, in turn, pairs of: (a) Mean Class Grades, (b) Mean UMI Scores, and (c) Mean Student Response Rates for Nine Administrative Units and more than 1,700 Instructor/Course Combinations in Fall, 2007

		Correlation between:		
Administrative Unit	**(No. of Classes)[a]**	**Class Grades & Mean UMI Scores**	**Class Grades & St. Resp. Rates**	**Mean UMI Scores & St. Resp. Rates**
Psychology	37-41	–.10	–.05	.32
Land and Food Systems	41-52	.22	.09	.12
Law	91-104	.16	.27	.12
Education	123-248	.21	.18	.08
Science	551-620	.29	.15	.07
Pharmaceutical Sciences	68-73	.38	.18	.18
Business	143-144	.19	–.03	–.08
Forestry	50-52	.28	–.08	–.19
Arts	645-672	.29	.19	.12
Wt'd-Mean Correl'ns:	1,749-2,006	**.27**	**.15**	**.11**

Notes: The correlations in this table are *partial r*s, with the effects of *year of class* (1st-, 2nd-, 3rd-, 4th-year, graduate) partialed out of both variables correlated in each case. The first column of correlations provides the partial *r*s between a class's mean grade and that instructor's score on the UMI Mean variable (the aggregate of the six UMIs), with year of class partialed out. The second column gives the partial *r*s between a class's mean grade and the proportion of students in that class that completed the student-evaluation inventory, with year of class partialed out. The third column contains the partial *r*s between the instructor's score on the UMI Mean variable and the proportion of students in that class that completed the inventory, with year of class partialed out.

[a] The administrative-unit *n*s differed somewhat from one of the three analyses to another.

The correlations in the first column in Table 8 (between grades and course ratings) are fairly low, although those for Education, Science, Pharmaceutical Sciences, and Arts do reach significance at the .01 level, and the aggregated value of .27 would, of course, also. Nonetheless they *are* low, and they are not easy to interpret with any certainty. How, for example, should we explain these correlations? Are the better-taught classes those that actually achieve at a higher level, thus earning for their students slightly higher grades? This interpretation would be consistent with the conclusions of Marsh and Roche (2000). Or is the causal pathway the other way around: those classes in which students perceive the grades as likely to be higher at the end of the term are those that, as a result, get slightly higher ratings from the students on the UMIs? This interpretation would be more consistent with the findings of Greenwald and Gillmore (1997a, 1997b) and, given the negative correlation between perceived workload and expected grades noted by these authors, the results might suggest that perceived workload is an extraneous factor affecting evaluation results, probably to a small degree. (It must be remembered in this connection that students know neither their course grades nor the class average when filling out the teaching evaluations, although they will, of course, be aware of their intermediate grades and may have some idea of these for the class as a whole.) In the present research, we did not have information on perceived workload, and thus cannot address this possible phenomenon directly. In any case, the average correlation of .27 is quite low, and, although interesting and deserving of further research, workload and leniency of grading should not be considered factors that can be seen as seriously compromising the results obtained in student evaluations of teaching. It is noted here that this value and the fact that eight of the nine reported in Table 8 are lower than .30 place the present results very close to those noted by other researchers for this correlation (Marsh, 1980, 1983, Marsh & Roche, 2000; Stumpf & Freedman, 1979).

The correlations between student response rates, on the one hand, and (a) class grades and (b) rated quality of teaching (via the UMI Mean measure), on the other (in the second and third columns of correlations in Table 8) , are even less potent, and, although for a few administrative units these are statistically significant ($p < .01$) because of the very large *n*s, (Law, Education, Science and Arts for grades and response rates; and Arts for course ratings and response rates), they represent very tiny correlational effect sizes, suggesting an almost complete absence of any relationship between the pairs of variables. I do not feel that the results in Table 7 indicate potential problems with the use of the UMIs and their online administration, suggesting as they do that whether or not students choose to complete evaluations on their instructors is close to being independent of the course grades and the students' perception of the quality of the teaching.

A detailed examination of the various contextual factors affecting student ratings of teaching does not fall within the terms of reference of the present evaluation. Here the concern has been with the adequacy of the UMIs and of the online administration of these items. My conclusion is that the various extraneous factors at play, although undoubtedly present and worthy of consideration, do not undermine the use of the UMIs or their online administration. All the administrative units at UBC have, for decades, been using evaluation inventories, and these extraneous factors have been operative (or not) during that period. There is nothing in the new initiative involving the UMIs or their online administration that should provoke greater concern in this regard.

Some Comments on the Posting of Teaching-Evaluation Results on a University Website

As noted, our concerns so far have been with the new UMIs and the effects of administering them online, questions amenable to empirical analysis and thus our primary focus in this report. Another

recommendation from Senate, however, and one that the SEoT Committee is interested in implementing, is the posting of results for individual instructors on a secure University website. It should be noted that what follows in this section, therefore, is not based on recently-obtained data, but rather reflects my own experiences and views in connection with posting of results.

Rationale for Posting Student-Evaluation Results

The University, acting on the Senate's recommendations, wishes to provide students with teaching-evaluation information to aid them in planning their programs of study. It is my understanding that students now use—to a very great extent—the *ratemyprofessors.com* website for this purpose. Examination of this website reveals that the numbers of respondents for each class are generally very small (maybe 1% - 5% of the class or less) and that the items used to rate the courses are less than optimal. Regarding the latter, students are asked to rate the instructor on (a) Easiness of the course, (b) Helpfulness of the instructor, and (c) Clarity of the instruction. The average of (b) and (c) forms an overall rating for the instructor. Thus, inadequate ratings are provided by far, far too few respondents for the results to be meaningful and useful. The University's position has been that, if students will otherwise use faulty information to make their course and class choices, then it has an obligation to provide more reliable and representative data for this purpose.

Features of the Proposed Posting Initiative

1. Perhaps the most important feature of this proposal is the option for any instructor to decline having his/her results posted. This seems like a valuable and necessary condition. The opinion has been expressed, however, that, despite this opt-out feature, instructors will feel pressure to agree to allow their results to be posted. I believe that this can be greatly reduced, if not completely eliminated through a few carefully-considered steps. A number of years ago, the Psychology Department was asked to post teaching-evaluation results. We developed our own website for this purpose and posted results. We used an "opt-in," rather than "opt-out" process in which instructors were contacted each spring (if they met the conditions noted below) after they had received their evaluation results, but before registration had begun for the following academic year and were asked whether they would agree to having their results posted. The conditions were as follows:

   (a) Only courses that were to be offered in the upcoming academic year were posted. We felt that this was consistent with the purposes of providing students with a basis for planning their academic program and that posting more would not effectively serve this purpose.

   (b) Instructors teaching a course for the first time in the upcoming year were exempted.

   (c) Courses having only a single section that were required of Major or Honours students and not open to any others were exempted. In this case, Major and Honours students did not have a choice about the course, and no other students were allowed to take it.

   (d) Courses with fewer than 15 students in the past were excluded. The logic here was that the number of respondents in such classes would be too small to provide results having sufficiently high inter-rater reliability.

We found that nearly all of those eligible (according the above conditions) agreed to have their results posted. We matched the course to be posted with the most recent section of that course that the instructor had taught and used the results from the latter. For instructors who had taught more than one section of the course in the same (and most recent) year, we took the average ratings across the sections. We posted results on the three departmental factors (described earlier) and provided department norms on those factors along with the instructors' mean scores.

We also provided a fairly lengthy introduction to this webpage, in which we explained the meaning of the factors (providing all the item content for each) and went to considerable lengths to explain why a particular instructor's results might be missing. This latter point relates to the concerns expressed (and briefly noted above) by some instructors that declining to have their results posted will signal poor evaluations. We felt, in Psychology, that by listing a number of reasons that a particular course might not be listed in the webpage that year, students would understand that many factors could account for missing results. We also cautioned students specifically not to assume that the absence of results for a course/instructor indicated anything whatsoever about the quality of the course.

In my opinion, the posting of student-evaluation results does fulfill the University's desire to provide students with better information than they could get from other sources and can be done in ways that minimize pressure on instructors who do not wish to have their results shown. With sufficient attention paid to the latter point, I feel that the process can be made minimally stressful for those who choose not to opt in.

2. It is my opinion that a number of other concerns raised by instructors need to be considered if they have not, in fact, already been considered and effectively dealt with. These concerns relate to: (a) preventing republishing of the results by students or others; (b) security of the data on the website and having safeguards in place to prevent manipulation or tampering; (c) the length of time results will remain available on the website; and (d) having safeguards guaranteeing that only legitimate class members have access to the inventory for an instructor and that they can complete it only once.

   These concerns all appear to be amenable to satisfactory resolution. Some of them will simply require a decision by the University; others will require some advanced web programming and implementation of security devices. Other matters, like those considered in the Psychology Department and discussed above, will need attention from SEoT and the University. My overall feeling, however, is that through careful and well-thought-through planning, the posting of student-evaluation results can serve a useful purpose and be done with minimal potential for undesirable outcomes.

SECTION 4 – SUMMARY AND RECOMMENDATIONS

Summary

Certainly, one finding that is unmistakable from the results of the present analyses is the generally high regard that UBC students have for the quality of the teaching they receive. Although the task was not to examine the results for the general quality of teaching they portrayed, it seems worthwhile to mention this, as the levels of the item and aggregated means we have seen have been striking. For the most part, means at or above 4.0 (in some cases, well above 4.0) on the 5-point scales have been routinely found (Tables 6 and 7, and particularly Table 1), and this indicates a mean level of perceived quality falling, for the most part, between "Good" and "Excellent" (or in some cases "Very Good"). This is particularly true with respect to the global items like UMI 6, that ask for a rating of the overall quality of the instructor's teaching, and the aggregated results. These results speak very well for the general level of teaching at UBC.

The addition of the University-Module Items (UMIs) to the current student-evaluation materials used in all administrative units on campus can provide additional uniform and useful teaching-evaluation information. In addition, the shift to online administration is a necessary advancement over the existing process and one that can be accomplished without any reduction in the quality of the data obtained. It is felt that the posting of results obtained with the UMIs will accomplish the Senate's goals in providing students with useful course-selection information and can be accomplished fairly and with minimal harm.

Some recommendations follow.

A General Recommendation

Although the UMIs have been shown to be satisfactory for the assessment of mainline facets of effective teaching, it is recommended that each administrative unit—department, faculty, or school—augment the University Module with its own items that reflect the particulars of instruction in that unit. This recommendation is not new or original with this report, having been made, as noted earlier, in Section 1, by the SEoT Committee and reflected in the Senate Policy. Nonetheless, as it appears that not all administrative units on campus were able to augment the UMIs in the first round of their implementation, it is felt that this recommendation should be stressed.

More Specific Recommendations – I. Short-Term

These are recommendations concerning the use of the new student-evaluation system over the next year.

1. Spring term, 2008. It is recommended that the University proceed with the six UMIs administered online.

2. Spring term, 2008. In order to enhance response rates, it is recommended that the University take steps to provide greater assurances about the anonymity of online student responses, with clear and persuasive statements to this effect.

3. Spring, 2008. It is recommended that the University website on which results are to be posted be thoroughly tested to ensure that the full range of safeguards against fraudulent responding and tampering are in place, and that all information posted is completely secure, with the raw data available only to those closely involved in the scoring and operation of the website. Steps should also be taken to offer clear and credible assurance to students that their responses will be kept completely anonymous.

4. Spring, 2008. I would recommend that SEoT develop a set of principles governing the posting of student-evaluation results. I am referring here to the matters raised earlier in this report about presentation of results, guidelines for inclusion of results, and opting-in procedures.

5. If the tasks outlined in Points 2 and 3 above can be successfully completed in time, it is recommended that the University post the results (subject to the inclusion principles developed) so that students registering for the 2008-09 academic year can benefit from them.

6. Spring-summer, 2008. It is recommended that SEoT begin holding focus groups with the various constituencies on campus in order to improve the process. Ideally, these would be completed in time to have a revised process in place for the Fall, 2008 evaluations. The specific matters needing attention, in my view, follow.

   (a) Wording of the UMIs. Although I feel that the individual UMIs do tap the important facets of effective classroom teaching, and are, therefore, adequate in their present form for use, I also feel that they can be improved. For the most part, the intended content of the item should be preserved, but the wording of some can be sharpened and generally improved. (I should note that, in this recommendation and the next, the suggestions are not based on empirical data, such as a survey of opinions of the items, but are instead my own opinions.)

   (b) Additions to the UMIs. I understand the importance of keeping the number of these items to a minimum in order to prevent the overall student-evaluation inventories from becoming too long. Nonetheless, the addition of two or three items would enable the sharpening and clarifications noted above in (a) without loss of item content. It seems in some cases that clarity was slightly sacrificed to keep the number of items low. One example is UMI 5: *(Rate) The instructor's concern for students' learning*. It seems likely that some students would read this with respect to in-class behaviours, whereas others would focus on accessibility outside of class. Two sharply-written, unambiguous items would help to solve this problem. Two items that it is felt would address this problem are presented below. These are merely suggestions; final forms of any new items would be the result of suggestions from a focus group assembled for this purpose.

   Present form (Rate): 5. The instructor's *concern* for students' learning.

   Alternate forms (Rate):

   5a: The instructor's willingness to explain material and answer questions in class.

   5b: The instructor's availability either through regular office hours or by appointment.

(c) Revision of existing inventories to accommodate the UMIs. One way to prevent the whole student-evaluation process becoming too lengthy is to remove the items on the existing administrative-unit inventories that tap the same themes as the UMIs. This recommendation is primarily directed to the individual academic units on campus (i.e., the faculties and departments).

(d) Incentives to improve response rates. This is perhaps of lesser importance than the recommendations noted above because the evidence we have now suggests that the response rates are not significantly different from those found with the paper format. Nonetheless, response rates in the 60%-70% are lower than ideal.

(e) Greater use of instructor-initiated items for formative evaluation. Although we do not have data that would support this, it is suspected that this fourth of the "four modules" is probably underused by instructors. Discussions would ideally centre around ways in which questions related to course improvement could be quickly and easily presented online in a secure way so that only the instructor had access to the responses. Given the ease of online presentation and the almost-universal access to computers by students, instructors should be encouraged to consider formative evaluation initiatives via this technology. In addition, for those instructors who wish to obtain information on matters not presently covered by the UMIs or additional faculty or departmental items, the opportunity to augment their present evaluation inventory with items of specific interest to themselves should be explored and use of this fourth module encouraged.

(f) Matters concerning posting of results. Principles of inclusion/exclusion, and website wording and presentation need discussion, with possible refinements added to the way these matters have been handled in Spring, 2008. Department-wide policies could also benefit from focused discussion.

More Specific Recommendations – II. Longer-Term

These recommendations concern longer-term maintenance of the student-evaluation program, along with institutional research and revisions based on this research.

1. It is recommended that the University set up a standing committee to monitor and refine the student-evaluation program.

2. It is recommended that programmatic institutional research be conducted with the new system with an eye to understanding its performance characteristics better and improving these. Particular attention should be paid to the following:

   (a) Development of Norms

   This is an important task that should begin as soon as possible. Normative data can be collected from a single administration that would provide useful campus-wide norms for those receiving and using the evaluation results. Individual instructors can benefit from knowing how their means on the UMIs compare with those of the population of all university instructors. For example, it is of limited usefulness to know that one obtained a mean score of 3.73 on UMI X. If, however, one also knows that, across campus, the

overall mean if 3.97, this makes more sense and provides needed perspective. After several terms of administrations of the UMIs, more specialized norms can be developed for individual administrative units, so that more refined information will become available against which to compare an individual instructor's results. Thus, an instructor would be able to compare her/his standing on a UMI with that of (a) all instructors at UBC, (b) all instructors in his/her faculty, and, say, (c) all instructors in her/his department. To the extent that these normative values differ, this more finely-grained information will prove useful.

As an example, in the Psychology Department many years ago, we developed departmental norms for our items and factors. These were originally based on something on the order of 400 instructor/course units. In time our norm base increased. Now, when an instructor receives his/her student-evaluation information, s/he can compare the item and factor means received with those for the Department as a whole, with the latter printed alongside that instructor's results.

In addition, we calculated means for (a) all first-year courses, (b) all 200-level courses, (c) all 300-level courses, and (d) all 400-level courses in the Department. For some multi-section courses, we obtained normative means for all the sections. These increasingly fine cuts of the norms have provided useful information to instructors in the Department.

(b) Examination of the Scale Points of the UMIs

In the process of preparing this report, I encountered the view from one person who is involved in student evaluations at UBC that the wording of the scale anchor points made it difficult to capture differences at the top end of the scales. In the Psychology Department and Faculty of Arts inventories, for example, the upper two anchor points are Good and Very Good. As we have seen, with the UMIs, the corresponding words are Good and Excellent. The thinking was that, since most of the scale means across campus are now at or above 4.0 (Good), more room was needed at the top end to discriminate between the *good*, *very good*, and *truly outstanding* instructors. Although potentially problematical, such scale revision might well be worth considering. It is therefore recommended that the matter be the central topic of a focus group, and, if this revision is perceived as worth pursuing after due consideration, the necessary psychometric work be done to make this change.

(c) Consideration of the Effects of Perceived Workload and Anticipated Grades

In a useful document entitled *Student Evaluations of Teaching: A Research Perspective* (2007), Gary Poole summarized some research on grading leniency performed by Greenwald and Gillmore (1997a, 1997b) that presented negative correlations between perceived workload in a course and expected grade. The direction of causation was not definitively established in this work, and differences in this phenomenon have different implications. The authors' conclusions were that these extraneous factors should be included in data gathered in connection with student evaluations. In the present analyses, we were able to examine, after the fact, the relationship between *actual* grades awarded and teaching evaluation scores, but the question of how perceptions of

grades and of workload affect student evaluations was left unanswered. It is recommended that further research be conducted on these factors and, if they are found to be important contaminants of evaluation results, some thought be given to the possibility of removing their effects from these results. As noted earlier, extraneous factors like these likely play a relatively small role in determining evaluation results, and the evaluation process is likely not seriously compromised by their presence. It would, nonetheless, be worthwhile to examine these factors, and, as the student evaluation process becomes more refined in the future, to consider ways to control for such contaminating factors.

(d) Short- and Long-Term Stability of the UMIs

(i.) The long-term study can follow the design of the one we used earlier (Table 2) to examine UMI stability via correlations with Psychology and Science items administered in Fall, 2006. As noted earlier, these correlations were attenuated because of slightly-differing item content. The ideal long-term stability study would involve as many administrative units as possible, with the online results obtained in Fall, 2008 and those parallel results obtained in Fall, 2009. Actually, the research can begin after the Fall, 2008 administration with classes in Psychology and Science because we have online UMI results for these units from Fall, 2007. A more comprehensive study, however, would include many more administrative units and would run Fall, 2008 to Fall, 2009. A parallel study could run from Spring 2009 to Spring, 2010.

(ii.) A more conventional test-retest study could be conducted over the period of Fall, 2008 to Spring, 2009. The number of parallel sections would be far less than with either the fall–fall or spring–spring sessions, but if classes were taken from all across campus, a sufficiently-large number would be available for a solid study of shorter-term stability.

(e) More Comprehensive Generalizability Studies

Data like these lend themselves to more comprehensive multi-faceted generalizability studies. To determine the extent of dependability over both time and course taught, for example, a study could be run with instructors who were teaching more than one course in a term, with this condition repeated a year later.

(f) Validity Studies

To better understand what precisely the UMIs are measuring, there are a number of ways construct validity could be evaluated once the program was in full swing. The number of UMIs is too small to allow for a factor analysis of them alone, but factor analyses could be conducted with the UMIs together with other items used in the various administrative units. If, for example, such a study were conducted in the Faculty of Arts—with the additional items administered by departments in the faculty—well-represented factors would emerge, locating the UMIs more precisely in the nomological net than they are presently.

(g) Experiments with Improving Response Rates

Studies on this topic would be useful. For greatest statistical power, these would ideally be repeated-measures studies, in which classes employing different incentives (ranging from none to several alternatives) over time would be compared with respect to their student-participation rates. Studies like this could take considerable time, in that a particular instructor/course unit would have to be available with more than one form of incentive. Some shorter-term repeated-measures options exist, however. Multi-sectioned courses taught by the same instructor could be used, in which no incentives were used in one section, for example, with each of two incentives, say, used in the second and third sections. For two-section courses by the same instructor in the same term, a two-incentive comparison would be possible.

Between-instructor studies would also be possible. In these, some instructors would use no incentives, whereas others would use one or more incentives, with response rates compared. There would be no reason not to include other factors in these studies, such as year of the course, gender of the instructor, etc., to determine whether these factors interacted with student response rates.

(h) Studies on Instructor Concerns

Instructors have identified some concerns about student evaluations of teaching, and these could be examined empirically. Are there administrative-unit differences in mean scores on the UMIs? Gender differences? Class-size effects? Interactions between these factors? How are the results being used? Are the proportions seeking assistance from TAG increasing? What are the experiences of instructors with having the results posted?

(i) Development of Follow-Up Procedures

To maximize the developmental benefits of teaching evaluation, it is recommended that some work be done in developing modules that would connect evaluation results to specific developmental activities like those offered through TAG. I would see this work as a long-term investment by the University and could be carried out, for example, by TAG personnel. In addition, individual administrative units might wish to develop policies for the use of evaluation results to enhance teaching effectiveness. Individual faculty members could be asked to debrief instructors after an evaluation and suggest remediation options if needed.

REFERENCES

Agreement on Conditions of Appointment for Faculty. (Last reviewed 4 October 2007). Retrieved March 1, 2008 from University of British Columbia, Human Resources Web site: http://www.hr.ubc.ca/faculty_relations/agreements/appointmentfaculty.html#4

Aleamoni, L. M. (1976). Typical faculty concerns about student evaluation of instruction. *National Association of Colleges and Teachers of Agriculture Journal, 20*, 16-21.

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

Appendix C – Effective Teaching Principles. (n.d.). Retrieved March 3, 2008 from University of British Columbia Centre for Teaching and Academic Growth Web site: http://www.tag.ubc.ca/resources/evaluation/appendixc.php

Arts Course Evaluations. (n.d.). Retrieved March 1, 2008 from http://evals2.arts.ubc.ca/search.php

Ballantyne, C. (2003). Online evaluation of teaching: An examination of current practice and considerations for the future. *New Directions in Teaching and Learning, 96,* 103-112.

Barnett, C. W., & Matthews, H. W. (1997). Current procedures used to evaluate teaching in schools of pharmacy. *American Journal of Pharmaceutical Education, 62*, 388-391.

Barnett, C. W., & Matthews, H. W. (1997). Student evaluation of classroom teaching: A study of pharmacy faculty attitudes and effects on instructional practices. *American Journal of Pharmaceutical Education, 61*, 345-350.

Cashin, W. E. (1990). Students do rate different academic fields differently. *New Directions for Teaching and Learning, 43,* 113-121.

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

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education, 13*, 321-341.

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

Dee, K. C. (2007). Student perceptions of high course workloads are not associated with poor student evaluations of instructor performance. *Journal of Engineering Education, 96*, 69-78.

Feldman, K. A. (1993). College students' views of male and female college teachers: Part II: Evidence from students' evaluations of their classroom teachers. *Research in Higher Education, 34*, 151-191.

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.

Felton, J., Koper, P. T., Mitchell, J., & Stinson, M. (2008). Attractiveness, easiness and other issues: Student evaluations of professors on Ratemyprofessors.com. *Assessment & Evaluation in Higher Education, 33*, 45-61.

Giesey, J. J., Chen, Y. N., & Hoshower, L. B. (2004). Motivation of engineering students to participate in teaching evaluations. *Journal of Engineering Education, 93*, 303-312.

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

Greenwald, A. G., & Gillmore, G. M. (1997a). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 1997,* 1209-1217.

Greenwald, A. G., & Gillmore, G. M. (1997b). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology, 89*, 743-775.

Guide to Promotion and Tenure Procedures at UBC 2007/08. (2007). Retrieved March 1, 2008 from University of British Columbia, Human Resources Web site: http://www.hr.ubc.ca/files/faculty_relations/pdf_files/sacguide0708.pdf

Hardy, N. (2003). Online ratings: Fact and fiction. *New Directions for Teaching and Learning, 96*, 31-38.

Harrison, P. D., Moore, P. S., & Ryan, J. M. (1996). College student's self-insight and common implicit theories in ratings of teaching effectiveness. *Journal of Educational Psychology, 88*, 775-782.

Heath, N. M., Lawyer, S. R., & Rasmussen, E. B. (2007). Web-based versus paper-and-pencil course evaluations. *Teaching of Psychology, 34*, 259-261.

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

Hoffman, K. M. (2003). Online course evaluation and reporting in higher education. *New Directions for Teaching and Learning, 96*, 25-29.

Howell, A. J., & Symbaluk, D. G. (2001). Published student ratings of instruction: Revealing and reconciling the views of students and faculty. *Journal of Educational Psychology, 93*, 790-796.

Johnson, T. D. (2003). Online student ratings: Will students respond? *New Directions for Teaching and Learning, 96*, 49-59.

Johnson, T. D., & Ryan, K. E. (2000). A comprehensive approach to the evaluation of college teaching. *New Directions for Teaching and Learning, 83,* 109-123.

Kelly, H. F., Ponton, M. K., & Rovai, A. P. (2007). A comparison of student evaluations of teaching between online and face-to-face courses. *Internet and Higher Education, 10*, 89-101.

Kierstead, D., D'Agostino, P., & Dill, H. (1988). Sex role stereotyping of college professors: Bias in student ratings of instructors. *Journal of Educational Psychology, 80,* 342-344.

Kwan, K. P. (1999). How fair are student ratings in assessing the teaching performance of university teachers? *Assessment & Evaluation in Higher Education, 24*, 181-195.

Langbein, L. I. (1994). The validity of student evaluations of teaching. *Political Science and Politics, September,* 545-553.

Layne, B. H., DeCristoforo, J. R., & McGinty, D. (1999). Electronic versus traditional student ratings of instruction. *Research in Higher Education, 40*, 221-232.

Llewellyn, D. C. (2003). Online reporting of results for online student ratings. *New Directions for Teaching and Learning, 96*, 61-68.

Marsh, H. W. (1980). The influence of student, course, and instructor characteristics on evaluations of university teaching. *American Educational Research Journal, 17*, 219-237.

Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology, 75*, 150-166.

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

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

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

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

McKeachie, W. J. (1987). Student ratings: The validity of use. *American Psychologist, 52*, 1218-1225.

Moore, T. (2006). Teacher evaluations and grades: Additional evidence. *The Journal of American Academy of Business, 9*, 58-62.

Murray, H. G. (1987). Acquiring student feedback that improves instruction. *New Directions for Teaching and Learning, 32*, 85-96.

Murray, H. G. (1997). Does evaluation of teaching lead to improvement of teaching? *International Journal for Academic Development, 2*, 8-23.

Ogier, J. (2005). Evaluating the effect of a lecturer's language background on a student rating of teaching form. *Assessment & Evaluation in Higher Education, 30*, 477-488.

Provostial guidelines for developing written assessments of effectiveness of teaching in promotion and tenure decisions. (2003, May 14). Retrieved February 27, 2008 from University of Toronto Governing Council Policies and Procedures Web site: http://www.utoronto.ca/govcncl/pap/policies/teaching.html

Pounder, J. S. (2007). Is student evaluation of teaching worthwhile: An analytical framework for answering the question. *Quality Assurance in Education, 15*, 178-191.

Regulations relating to the employment of academic staff. (Last revised 10 April 2006). Retrieved March 1, 2008 from McGill University Administration and Governance Secretariat Web site: http://www.mcgill.ca/files/secretariat/1RegulationsRelatingtotheEmploymentofAcademicStaff.pdf

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

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment & Evaluation in Higher Education, 30*, 387-415.

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

Santhanam, E., & Hicks, O. (2002). Disciplinary, gender, and course year influences on student perceptions of teaching: Explorations and implications. *Teaching in Higher Education, 7*, 17-31.

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

Schuerger, J. M., & Witt, A. C. (1989). The temporal stability of individual tested intelligence. *Journal of Clinical Psychology, 45,* 294-302.

Sinclair, L., & Kunda, Z. (2000). Motivated stereotyping of women: She's fine if she praised me but incompetent if she criticized me. *Personality and Social Psychology Bulletin, 26*, 1329-1342.

Spencer, K. J., & Schmelkin, L. P. (2002). Student perspectives on teaching and its evaluation. *Assessment & Evaluation in Higher Education, 27*, 397-409.

Sorenson, D. L., & Johnson, T. D. (Eds.). (2003). Online student ratings of instruction. *New Directions for Teaching and Learning, 96.*

Stumpf, S. A., & Freedman, R. D. (1979). Expected grade covariation with student ratings of instruction: Individual versus class effects. *Journal of Educational Psychology, 71*, 293-302.

Summary of course evaluation systems for G13 universities. (2007, May 17). Teaching and Learning Services, McGill University.

Tang, T.L.-P. (1997). Teaching evaluation at a public institution of higher education: Factors related to overall teaching effectiveness. *Public Personnel Management, 26*, 379-389.

The Teaching Dossier. (n.d.). Retrieved March 3, 2008 from University of Guelph, Teaching Support Services Web site: http://www.tss.uoguelph.ca/resources/idres/packagetd.html

Viswesvaran, C., & Ones, D. S. (2000). Measurement error in "Big Five Factors" personality assessment: Reliability generalization across studies and measures. *Educational and Psychological Measurement, 60,* 224-235.

Wisdom through reflective practice. (n.d.). Retrieved March 3, 2008 from University of British Columbia Centre for Teaching and Academic Growth Web site: http://www.tag.ubc.ca/programs/series-detail.php?series_id=277