# SOME RESULTS FROM TWO YEARS OF ONLINE ADMINISTRATION OF THE UNIVERSITY MODULE ITEMS (UMIS) IN THE EVALUATION OF TEACHING IN THE FACULTY OF ARTS AND FACULTY OF SCIENCE AT UBC

June, 2010

Student evaluation of teaching at UBC has been carried out in almost all faculties for the past two years by means of the online system: online presentation of the items to student respondents and online feedback of results to instructors. The exceptions have been Business, some departments in Medicine, and, until 2010, Education and Forestry (with Education beginning use of the online system in Summer, 2010, and Forestry scheduled to begin in Fall, 2010). Further, the departments that have used the online system have used the set of six evaluation items known as the University Module Items, or UMIs. We thus now have sufficient information to critically evaluate, on the basis of a large sample of instructors, these UMIs in terms of their performance characteristics, at the same time as investigating how the online system compares in certain ways with the earlier pencil-and-paper format. In this report, we describe three aspects of the present online administration of the UMIs, on the basis of data from the Faculties of Arts and Science: (a) possible effects of online item presentation on student-response rates, (b) stability over time of scores on the online-administered UMIs, and (c) UMI score levels obtained via the online format, and how these compare with those obtained earlier with the pencil-and-paper administration mode. These performance aspects of the new online system with the UMIs are treated in turn below. We note here that we have *not* included in these analyses [with respect to characteristics (b) and (c) above] items other than the UMIs, such as faculty- and department-specific items.

## RESPONSE RATES: BEFORE AND AFTER THE CHANGE TO ONLINE ADMINISTRATION

On the basis of a large dataset of student responses to the UBC University Module Items (UMIs) administered via both the pencil-and paper and online inventories, we can now compare the two administration formats with respect to their associated ***response rates*** (*defined as the percentage of students in a class that complete the student-evaluation inventory for that class*). 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 and 2009-10 academic-year online administrations of the UMIs. 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 and 2009-10 academic-year online administrations of the UMIs.

### *(a) Potential Bias in the Paper Results*

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

*(b) 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**

*Average Yearly* Student-Response Rates for Paper Administration and for Online Administration for the Faculty of Arts and Faculty of Science*

| Faculty | Administration Format | No. of Sections | Percentage of Students Responding |
|---|---|---|---|
| **ARTS** | Pencil & Paper 2007-08 Terms 1 & 2 | 1,898 | 75.97% |
| | Online 2008-09 Terms 1 & 2 | 2,691 | 59.70% |
| | Online 2009-10 Terms 1 & 2 | 2,328 | 56.32% |
| | Mean Online over 2 Academic Years | 5,019 | 58.13% |
| **SCIENCE** | Pencil & Paper 2006-07, Terms 1 & 2 | 783 | 66.25% |
| | Online 2008-09 Terms 1 & 2 | 1,342 | 62.90% |
| | Online 2009-10 Terms 1 & 2 | 1,212 | 60.37% |
| | Mean Online over 2 Academic Years | 2,554 | 61.70% |

**Note:* The percentages of students responding reported above are the ***unweighted*** *mean response percentages* of the number of sections indicated in each line. That is, the percentage of student response is weighted equally for all sections.

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, it appears like about a 17–18% drop in the Faculty of Arts, but only about a 4.5% drop in the Faculty of Science. Given the uncertainty surrounding our paper data, however, we are reluctant to attempt much of an interpretation of these apparent reductions in response rates.

It might be of interest to readers 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 58–62% rank near the top. Many of these universities have seen similar reductions in response rates from paper to online administration and resulting online response rates in the 30–40% range, although there is evidence that, by altering administration methods and including incentives, this need not be the case, and much higher response rates are possible.[1] We (the Student Evaluation of Teaching—SEoT—Implementation Committee) are actively reviewing and considering methods by which our student-response percentages can be raised.

[1] Examples of administrative interventions have included requiring inventory completion before being allowed to take the final exam and delayed receipt of grades if the inventories were not completed. Just two examples of the many incentives used are to award bonus points to students who complete the inventories and to register those who complete them into a draw for a prize.

*(c) Relationships between (i.) Student-Response Rate and (ii.) 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 is 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 and: (a) Mean Scores on the Six UMIs and (b) Class Size for the Two Faculties, Two Administration Modalities, and Several Aggregations*

| | | **Correlation between Student-Response Rate and:** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | **Scores on:** | | | | | | **Class** |
| **Faculty** | **Administration Format/Year** | **UMI1** | **UMI2** | **UMI3** | **UMI4** | **UMI5** | **UMI6** | **Size** |
| **ARTS** | Paper 2007-08, Terms 1 & 2 (1,898 Sections) | .182 | .188 | .183 | .155 | .231 | .189 | –.439 |
| | Online 2008-09 Terms 1 & 2 (2,691 Sections) | .066 | .092 | .109 | .095 | .115 | .092 | –.173 |
| | Online 2009-10 Terms 1 & 2 (2,328 Sections) | .099 | .110 | .133 | .134 | .189 | .120 | –.240 |
| | Mean Online over 2 Full Acad. Years (5,019 Sections)[a] | .082 | .101 | .121 | .114 | .151 | .105 | –.207 |
| **SCIENCE** | Paper 2006-07, Terms 1 & 2 (783 Sections) | (------ | ------ | *UMIs not used* | | ------ | ------) | –.402 |
| | Online 2008-09 Terms 1 & 2 (1342 Sections) | .086 | .050 | .060 | .099 | .120 | .062 | –.155 |
| | Online 2009-10 Terms 1 & 2 (1,212 Sections) | .142 | .093 | .154 | .098 | .210 | .125 | –.131 |
| | Mean Online over 2 Full Acad. Years (2,554 Sections)[b] | .113 | .071 | .105 | .099 | .165 | .092 | –.145 |

**Table 2 *(Continued)***

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Mean Online over Both Faculties and Both Online Years Combined** (7,573 Sections)[c] | .093 | .091 | .116 | .110 | .157 | .101 | –.188 |

[a]None of the year-to-year *differences* in correlation coefficients (Faculty of Arts) between the two online years is statistically significant *(p* < .01) except that for UMI5 *(p* = .008). For this reason, we have calculated a weighted average (for the two online years, appearing above this row) for each correlation coefficient, and these weighted averages appear in this row.

[b]None of the year-to-year *differences* in correlation coefficients (Faculty of Science) between the two online years is statistically significant (*p* < .01). Thus, we have calculated a weighted average (for the two online years, appearing above this row) for each correlation coefficient, and these averages appear in this row.

[c]None of the corresponding correlational *differences* between the two faculties is statistically significant (*p* < .01) except that for Class Size *(p* = .009). For this reason, we have calculated a weighted average (aggregated over the two faculties and two online years—Row 4 for Arts and Row 4 for Science) for each correlation coefficient, and these weighted averages appear in this row.

The results in Table 2 are interesting. With respect to the correlations involving the UMIs, we see two clear consequences of the shift from paper to online administration. The first is the significant reduction in these correlations when going from those associated with paper administration to those associated with online administration. In Table 2, the six correlations corresponding to paper administration are in Row 1 for the Faculty of Arts. Over the six, the mean correlation is about .19. These correlations can be compared with those for two aggregations of correlations associated with online administration, with the results almost identical for these two comparisons. For the first comparison, we might use Row 4 for the Faculty of Arts. These are aggregated correlations over the two years of online use in the Faculty. The mean of the six correlations in that row is about .11. For all of the UMIs except UMIs 3 and 4, the differences are statistically significant *(p* < .01), and for UMI3, the difference is close to this stringent statistical criterion *(p* = .019).

If, instead of making the comparison of correlations associated with paper administration with those for the Faculty of Arts only, we compare them with the fully-aggregated results in the last row in Table 2, the results are almost identical, with the only difference being that now the difference involving UMI3 does meet the .01 criterion, leaving UMI4 the only item for which the difference does not reach statistical significance. Again, the mean of the six correlations in this row is about .11. Thus, we can see that, with online administration, there is a statistically significant and fairly substantial reduction in the correlation between student-response rate and scores received on the UMIs over that found with paper administration (on average, .11 vs. .19).

The second, related, observation of interest concerning correlations with the UMIs is the actual size of the correlations between student-response rate and the UMIs under online administration. These correlations are now in the essentially-negligible category, averaging, as noted, around .11. This should be seen as good news, since the change to online format has been accompanied by near-removal of 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 very little effect on the mean UMI scores that instructors receive.

Another interesting result seen 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 inventory and the class size: –.439 for Arts and –.402 for Science (the two correlations not significantly different), over a total of 2,681 sections. 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. Results based on our most complete data for online administration appear in the last row of Table 2, where we see that, over both faculties and two years of online administration (a total of 7,573 sections), the mean correlation between student-response rate and class size is –.188. This reduction in the strength of relationship of, on average, .240 is large and highly statistically significant. Thus, although this reduced correlation does not reflect a complete absence of relationship between student-response rate and class size (it is still true that the student-response percentage is higher in smaller classes), the magnitude of the effect of this extraneous variable (for interpretive purposes) is, with the change to online administration, small enough to be considered almost irrelevant. As with the reduction in the relationships between student-response rates and mean UMI scores, this result with class size is a welcome one. It means that the results received by instructors of large classes are based, in general, on *nearly* the same *percentage* of student respondents as for those instructors of smaller classes.

*(d) Summary of Results Involving Student-Response Rate*

We now have evidence of a somewhat-reduced student-response rate following the change from paper inventory administration to online administration. We cannot be certain of the actual magnitude of the reduction because the paper response-rate results are not free of bias. The Faculty of Science results suggest that the decrease may be on the order of 4–5%, although that in the Faculty of Arts may well be twice that or more. Efforts on the part of the SEoT Implementation Committee will continue to be focused on improving our online-based student-response rates that are currently around 60%.

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 on mean UMI scores received by instructors are significantly lower. This means that, although we shall continue to do whatever we can to raise these response rates, instructors need not feel that their obtained UMI results have been greatly and adversely affected by low response rates and that these results are, because of this fact, an unfair indication of their teaching performance. Second, the negative association between class size and student-response rates appears to be substantially lower with the change to online administration. This means that instructors of very large classes can expect not to be faced any longer 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.

It is worth noting here that the results displayed in Tables 1 and 2 should be seen as very solid and unlikely to change to any extent on the basis of sampling error. We now have online-

administration data on 7,573 course sections, over both faculties, on which our results are based, comprising two full academic years' use of the online system.

## LONG-TERM STABILITY OF UMI MEAN SCORES

In order to assess the long-term (defined here as one-year) stability of the UMI mean scores, we assembled a sample of instructor/course units that were identical in Terms 1 and 2, 2008-09 and Terms 1 and 2, 2009-10. That is, we isolated 919 sections in the Faculty of Arts and 461 sections in the Faculty of Science in which the same course was taught by the same instructor in the two full academic years, separated by one year. The purpose was to examine the extent to which the rank-ordering of the UMI scores remained constant from one year to the next. We would, for example, have 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.

As a reminder to the reader of the content of the six UMIs, they are given below in Table 3.

**Table 3**

*The University Module Items (UMIs)*

| UMI | Content |
|---|---|
| UMI1 | The instructor made it clear what students were expected to learn. |
| UMI2 | The instructor communicated the subject matter effectively. |
| UMI3 | The instructor helped inspire interest in learning the subject matter. |
| UMI4 | Overall, evaluation of student learning (through exams, essays, presentations, etc.) was fair. |
| UMI5 | The instructor showed concern for student learning. |
| UMI6 | Overall, the instructor was an effective teacher. |

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

*(a) One-Year Lag Correlations by Faculty*

Below in Table 4 are the one-year correlations for each UMI, along with some summary means of these correlations, for each faculty separately and then aggregated over the faculties.

**Table 4**

*One-Year Correlations (2008-09, Terms 1 & 2 to 2009-10 Terms 1 & 2) by Faculty for Each of the UMIs*

| | Faculty | | |
|---|---|---|---|
| UMI | Arts (919 Sections) | Science (461 Sections) | Wt'd Average (1,380 Sections)[a] |
| UMI1 | .59 | .57 | .58 |
| UMI2 | .64 | .66 | .65 |
| UMI3 | .69 | .70 | .69 |

**Table 4 *(Continued)***

| | | | |
|---|---|---|---|
| UMI4 | .56 | .49 | .54 |
| UMI5 | .70 | .64 | .68 |
| UMI6 | .66 | .64 | .65 |
| Mean over 6 UMIs: | .64 | .62 | .63 |
| Mean excluding UMI4: | .66 | .64 | .65 |

[a]None of the Arts – Science correlational differences for the UMIs and the mean correlations was significantly different from 0, indicating that we can consider the parameters estimated in each case identical, with our best estimate of this parameter correlation being that in this column.

*(b) Issues and Conclusions Concerning the Long-Term UMI Correlations*

There are several conclusions we can draw from the results in Table 4. First, as noted in the footnote to the table, there is considerable similarity between the results obtained in Arts and those in Science. 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, and the differences between this correlation and those for all other UMIs except for UMI1 (also on the low side) are significant. This difference is the reason we have presented two overall means in Table 4, one with and one without UMI4. We must note, however, that beginning in Term 1, 2009-10, a *N/A*, or "Not Applicable" response option was added for UMI4 only. The fact that this option was absent in the 2008-09 administration, but present in 2009-10, may help explain the lower long-term correlation for this item.

Although manifesting a significant difference in terms of their one-year correlations from only 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 (very nearly .65) is similar to what behavioral scientists have found, over the years, for many variables.

We might be tempted to search for the reasons for the one-year correlational differences among the UMIs noted above. 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 learning* (UMI4)? 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. One answer might be that the personal characteristics of instilling inspiration in, and feeling concern for, others (tapped by UMIs 3 and 5) are tendencies that lie deeper in the organism and are more definitive of the individual than are the more mechanical aspects of teaching such as being sure to lay out clear expectations and using what are seen as fair evaluation procedures (UMIs 1 and 4). The more long-term person-defining traits of instructors change very little from one year to another, whereas the more mechanical *skills* (as opposed to deep-seated traits) are both easier, and more likely, to change over time.

This latter form of change over time of measurable behaviors has, in the past, been referred to by measurement specialists as *function fluctuation*. This phenomenon refers to an actual change in the underlying behavior, not in merely the measurement of it, and we hope for such behavioral change in the cases of instructors receiving low UMI mean scores. Thus, although we might expect considerable similarity of mean UMI scores for instructors from year to year, we are not dealing with something assumed to be a constant except for the fact of random measurement error (although these scores do, of course, contain such error).

The biggest issue in connection with the results in Table 4, however, may be 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. Time intervals, however, 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, not far from 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 over time, 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 those present 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 test-retest variety considered in the studies to which we have alluded. Thus, any test-retest correlation—particularly with a long time interval like one year—involving a single personality or ability item would be expected to be considerably lower than those correlations obtained with scale scores.

Second, the stabilities reported in the literature are for scores obtained by either *self-report* (on inventories of personality, attitudes, interests, or other non-cognitive variables) or *performance* on a maximum-performance measure (such as an ability test). The unit of analysis in these studies has been the individual responder. With the UMIs, of course, the instructor's score arises from the aggregation and averaging of ratings by *others*.

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 4.01. 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 reduce students' perceptions of a course and its instructor, or that can account for year-to-year changes in these perceptions. Some of these might be: 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 the course and instructor in ways they would not have without this single student (or group of 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. Even slight year-to-year historical changes in society or the world at large can play a part.

Along with function fluctuation—or true planned change in teaching by the instructor—we have other unintended instructor changes during the term, some of which are negative, such as illness, failure to receive a grant, 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 performance is diminished. On the other hand, positive incidents can occur, such as news of a positive promotion or tenure decision, that could trigger a spurt of pleasure and enthusiasm during a term and 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.

## CHANGES IN MEAN UMI SCORES OVER ADMINISTRATION FORMAT AND 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, in the present analyses, it would be informative to examine and compare with this earlier-noted decline, what might be considered typical one-year fluctuations in mean UMI scores 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 was a group of 707 instructors that taught the same course in the full 2007-08 academic year (and had their teaching evaluated via the paper forms) and in the 2008-09 year (when they had their teaching evaluated by the online system). Arts Cohort 2 was a group of 919 instructors (many, if not most, of whom would undoubtedly have also been in Cohort 1) that taught the same course in both the full 2008-09 and the full 2009-10 academic years and had their teaching evaluated on both occasions by means of the online format.

In the Faculty of Science, Cohort 1 was a group of 275 instructors who taught the same course in both the full 2006-07 academic year (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 5 below) and in the full 2008-09 year (when their teaching was evaluated by means of the online system). In the Faculty of Science, 2006-07 was the last year in which the pencil-and-paper format was used. The number of instructor/course units is lower than might be expected because of the two-year time interval. Science Cohort 2 was a group of 461 instructors (many of whom would also have been in Cohort 1) that taught the same course in both the full 2008-09 and the full 2009-10 academic years and had their teaching evaluated on both occasions by means of the online survey.

The reason for using these particular groups of instructors in these analyses was to control as much as possible for factors extraneous to the intended purposes, such as the effects of different instructors and courses in the two comparison groups which would, of course, introduce additional error into the results. Results for these analyses appear below in Table 5.

**Table 5**

*Changes in Mean UMI Scores Over Administration Format and Time for the Faculty of Arts and the Faculty of Science for Two Cohorts of Instructors Teaching the Same Courses over Two Years*

**FACULTY OF ARTS**

*Cohort 1:* (n = *707) Paper, 2007-08 – Online, 2008-09*

| | 1st Administration 2007-08 Paper | 2nd Administration 2008-09 Online | Change (2nd – 1st) |
|---|---|---|---|
| UMI1 | 4.193 | 4.168 | –.025 |
| UMI2 | 4.301 | 4.148 | –.153 |
| UMI3 | 4.170 | 4.111 | –.059 |
| UMI4 | 4.192 | 4.069 | –.123 |
| UMI5 | 4.284 | 4.251 | –.033 |
| UMI6 | 4.341 | 4.164 | –.177 |

*Cohort 2:* (n = *919) Online, 2008-09 – Online, 2009-10*

| | 1st Administration 2008-09 Online | 2nd Administration 2009-10 Online | Change (2nd – 1st) |
|---|---|---|---|
| UMI1 | 4.131 | 4.162 | .031 |
| UMI2 | 4.133 | 4.160 | .027 |
| UMI3 | 4.089 | 4.121 | .032 |
| UMI4 | 4.026 | 4.078 | .052 |
| UMI5 | 4.226 | 4.217 | –.009 |
| UMI6 | 4.140 | 4.160 | .020 |

**Table 5 *(Continued)***

**FACULTY OF SCIENCE**

*Cohort 1:* (n = *275) Paper, 2006-07 – Online, 2008-09*

| | 1st Administration 2006-07 Paper | 2nd Administration 2008-09 Online | Change (2nd – 1st) |
|---|---|---|---|
| UMI6 (2008-09); Summative Item (2006-07) | 4.188 | 4.096 | –.092 |

*Cohort 2:* (n = *461) Online, 2008-09 – Online, 2009-10*

| | 1st Administration 2008-09 Online | 2nd Administration 2009-10 Online | Change (2nd – 1st) |
|---|---|---|---|
| UMI1 | 4.084 | 4.073 | –.011 |
| UMI2 | 4.025 | 4.006 | –.019 |
| UMI3 | 3.961 | 3.938 | –.023 |
| UMI4 | 3.926 | 3.889 | –.037 |
| UMI5 | 4.158 | 4.103 | –.055 |
| UMI6 | 4.081 | 4.043 | –.038 |

With respect to the paper-to-online differences in Table 5 (Cohorts 1 for each faculty), we should note that, in addition to the administrative change from paper to online presentation, there were small wording (and, for the Faculty of Arts, response-option) changes between the paper and online items. First, the wording in the item stems changed very slightly. An example is UMI4, which in its paper form when used by the Faculty of Arts read:

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

and was responded to on a 5-point Very Poor to Very Good response scale, and, in its online form read:

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

and was responded to on a 5-point Strongly Disagree to Strongly Agree response scale. Second, as seen above, the structure of the items and their response scales changed (in the Faculty of Arts) from an aspect of instruction that was rated on the following scale:

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

to the corresponding aspect of instruction phrased as a positive statement and rated via the Likert-scale:

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

The example using UMI4 above illustrates this change.

In the Faculty of Science, the most global item in the 2006-07 paper inventory read:

*The instructor taught effectively*

and in the 2008-09 online survey read:

*Overall, the instructor was an effective teacher* (this being UMI6),

with both responded to on the 5-point Likert scale given above, running from Strongly Disagree to Strongly Agree.

We do not know at this point, just how much the received scores may have changed solely because of these wording and response-scale changes, above and beyond the changes brought about by online, as opposed to paper, administration. This is a question we are currently attempting to answer with further research.

The results for the two Cohorts 2 in Table 5 provide some insights into what we might expect in the year-to-year fluctuation of mean scores (for identical instructor/course combinations) on the now-final UMIs (vis-à-vis item wording) obtained via the online administration format. The means and differences between means from 2008-09 and those from 2009-10 presented in Table 5 are relatively stable, being based on large enough samples to ensure fairly small standard errors for the means (in the .015 to .018 range for the Arts data and in the .019 to .025 range for the Science data). Year-to-year fluctuations in the UMI means of up to about .035–.040 for the Faculty of Arts and .050 for the Faculty of Science can be expected *by chance* (the difference in magnitudes of these "margins of error" arising from the larger number of courses given in Arts) among year-apart paired samples of identical instructor/course units.

If we consider the mean UMI changes between the two academic years in which the online system was employed (the results for Cohorts 2), most are not statistically significant (nor would we expect them to be). In the Faculty of Arts data, for the six UMIs, the difference for UMI4 does reach statistical significance (using our $p < .01$ criterion); the differences for the other UMIs do not. This change of +.052 scale points for UMI4 from 2008-09 to 2009-10, however, probably lacks any practical significance, representing as it does a standardized effect size of only .11. It may point to a very small improvement in grading practices, but the reader is cautioned not to make too much of this difference.

Similarly, in the Faculty of Science results, five of the six UMI yearly differences (2008-09 to 2009-10) are nonsignificant. The one UMI for which a statistically-significant difference was found, UMI5, recorded a decline of .055 scale points. This change is even harder to explain than that for UMI4 among the Arts results. Given that this –.055 change corresponds to a standard effect size of only –.13, however, we again appear to be dealing with a statistically-, but not practically-, significant phenomenon.

It might be useful to remind the reader that, in the analyses yielding the results reported in Tables 4 and, particularly, 5, samples much smaller than the entire numbers of courses given in each year have been employed. This is because we wanted to examine item mean changes for identical

instructor/course units for greater precision and, with the correlational data, had to use these samples. However, the yearly item mean fluctuations for these subsamples might not precisely mirror those we would obtain for *all* courses given in each faculty each year. Results for these larger yearly samples would, however, reflect effects of what we consider extraneous variables in the analyses and would tell us less about the performance of the items themselves (reflecting factors like personnel changes in the yearly instructor cohort) than do those based on the particular subsamples we have used.

The SEoT Implementation Committee plans to continue to monitor year-to-year UMI scores and examine the shifts that occur over time. Doing so will enable us to detect improvements in teaching at UBC as well as the areas most in need of improvement.