AN EXAMINATION OF THE EFFECTS ON AVERAGE UMI RATINGS OF ONLINE VS. PAPER-AND-PENCIL INVENTORY ADMINISTRATION: THE 2008-09 RESULTS FOR THE FACULTIES OF ARTS AND SCIENCE

27 November 2009

PRESENT ANALYSES

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 nothing more than a recalibration of our student evaluation processes at UBC. Accordingly, 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 below.

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

(a) 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 1

Results by Item over the Entire Faculty (Academic Year 2007-08 to 2008-09, Both Terms)

UMI	Mean 2007-08 (Paper)	Mean 2008-09 (Online)	Difference in Means[a]	SD 2007-08	SD 2008-09	Difference in SDs[a]
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.

(b) 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

Results by Department on the Summary Measure—Average of the Six UMIs—and on UMI 6

		Difference in Means[b] from 2007-08 to 2008-09 for:	
Administrative Unit	**Number of Paired Sections**	**Average of the 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 (by number of sections in administrative unit) Average: (*n* = 706 sections)		**–.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.

(i.) *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.

(ii.) *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.

(c) 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

(a) 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 is presented in its entirety on Page 7.

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).

(b) 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	No. 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 years)	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.