AN INVESTIGATION INTO THE EFFECTS OF INSTRUCTOR GENDER, FIELD OF STUDY, AND STUDENT-RESPONDENT GENDER ON UMI SCORES IN THE 2008-09 SEoT ADMINISTRATION

Abstract

The effects on UMI scores of gender of instructor, gender of student respondent, and field of study were simultaneously examined via a three-way analysis of covariance (ANCOVA), incorporating control for any influences on scores arising from class size and mean course grade. A total sample of 519 UBC instructor/course units from the 2008-09 academic year's offerings was divided into 342 taught by male instructors and 177, by female instructors (roughly replicating the instructor gender proportions at UBC). In addition, these 519 units were divided equally among the Humanities, Social Sciences, and Science, each with 173 instructor/course units. In addition, for each instructor/course unit, mean ratings on each UMI were obtained separately for the male students and female students. With this orthogonal design, seven dependent variables were analyzed—the six UMIs and the average of the six UMIs, taken as an overall aggregated summary measure..

Small instructor-gender effects were found for the averaged UMI measure and UMI 6 (the summative item) in favor of female instructors. However, a consistent instructor-gender × student-respondent gender interaction effect was also found, and this reduced the interpretability of the instructor-gender effects. Analysis of these interactions revealed that, in general, female students tended to rate female instructors significantly more highly than they rated male instructors, but that this effect was not present for male students, who tended to rate male and female instructors relatively equally. In addition, a small, but significant field-of-study effect was found with the averaged UMI scores, with mean scores for the Social Sciences significantly higher than those for Science, but this effect too was compromised by a significant interaction effect involving the student-respondent factor, where it was found that this field-of-study difference was manifested only in the ratings provided by the female student respondents. With two UMIs analyzed separately, the mean of the Humanities/Social Sciences ratings were significantly higher than the Science means, and this effect was not compromised by interaction, although it was small.

Differences were also found between individual UMIs on the basis of a sample with all instructor/course units combined (and student-respondent gender means aggregated). These differences are discussed, and possible implications for teaching improvement are identified.

Overview

The purpose of this study was to provide information on the effects of gender of both Instructor and Student-Respondent, along with those arising from Field-of-Study, on responses to our final set of University Module Items (UMIs). The study was based on Student Evaluation of Teaching (SEoT) results obtained, through online administration of the UMIs, in both terms of the 2008-09 academic year. Questions about whether male and female instructors can be expected to be systematically rated differently, whether male or female student-respondents can be expected to give different ratings, and whether ratings obtained in substantively different academic disciplines can be expected to vary by discipline were addressed. Although there is some (albeit very little) literature relating to these factors, our concern was to examine them in the context of the newly-developed UMIs, now being used by almost all faculties at UBC.

To remind readers of the content of the present UMIs, we list them below on Page 2.

University Module Items (UMIs) in Use at UBC since the 2007-08 Academic Year

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.

These items are responded to on the following 5-point scale:

1 - Strongly Disagree; 2 - Disagree; 3 - Neutral; 4 - Agree; 5 - Strongly Agree.

The Present Experimental Design

Independent Variables

There were three factors in the present study: (a) Gender of Instructor, (b) Gender of Student-Respondent, and (c) Field of Study. The third factor proved somewhat difficult to capture to our full satisfaction because of overlaps between fields. We settled on three levels for this factor: courses in (a) the Humanities, (b) the Social Sciences, and (c) Science (including the Life Sciences). This was after a number of attempts to include some applied faculties. These latter faculties presented some problems in substantially overlapping with the fields included in (a) to (c). We further attempted to break the Science category into what might be termed the "hard Sciences" and Life Sciences, but the number of data points for the analysis was just too small for the latter, and *all* Science departments were, therefore, aggregated into one category in the analysis. Here is the departmental breakdown for each of Categories (a) to (c), which constitute our three levels of the Field of Study factor in this analysis:

(a) *Humanities*: Departments of Art History & Visual Arts, Asian Studies, Central, Eastern & Northern European Studies, Classical, Near Eastern & Religious Studies, English, French, Hispanic & Italian Studies, History, and Philosophy;

(b) *Social Sciences*: Departments of Anthropology, Economics, Geography, Political Science, Psychology, and Sociology;

(c) *Science*: Departments of Chemistry, Computer Science, Earth & Ocean Sciences, Mathematics, Physics, Statistics, Botany, and Microbiology & Immunology.

Experimental Design

The unit of analysis. In the present study the experimental (and, at the same time, observational) unit of analysis was the *instructor/course unit*. By this, we mean that the numbers analyzed were the *means* obtained by Instructor *X* teaching Course *Y* in the 2008-09 academic year at UBC. Such mean ratings were obtained, for each instructor/course combination on each of the six UMIs and on their average. It is thus variation among item (and averaged) means for classes (or instructor/course combinations) that provides the "error" component in the analyses, not that among students rating their instructors. In total we used a sample of 519 instructor/course units.

To elaborate further, we avoided dependencies in the data arising from the same instructor teaching more than one course or multiple sections of the same course by averaging, for each instructor, over all courses taught in the 2008-09 academic year. Thus each data point (unit of analysis) represents a unique instructor—in some cases that instructor's mean scores from one course, and in other cases that instructor's aggregated (over two or more courses) mean scores. Thus, there were actually 778 instructor/course units in the three field of study groups noted above, but after aggregation within instructors, we had 519 unique instructors represented. This means that some of the data points represent one instructor's results from teaching one course, and in others, one instructor's results averaged over two or more courses.

Design variables. There were three analysis of variance (or ANOVA) factors in the design. Both Instructor Gender and Field of Study were between-subjects factors, whereas Student-Respondent Gender was a within-subjects factor. By the latter, we mean that for each instructor/course combination, we had the mean of the male evaluations and the mean of the female evaluations (the fact that the numbers of male and female respondents differed is immaterial in this context). Therefore, for each instructor we had (a) gender, (b) field (of the three above), and two scores for each item (and the average of all 6)—one from male student-respondents and one from female-respondents. This kind of design is referred to as a "Three-Way Between-Within ANOVA Design" (2 × 3 × 2 in this case). It is a very powerful design and enabled us to evaluate: (a) each of the factors separately, (b) all interactions between pairs of factors, and (c) any three-way interaction effect that may be present.

Covariates. In addition, in order to control for (a) Class Size and (b) Mean Course Grade, we obtained measures of these for each instructor/course unit, with the grade variable being the mean grade assigned in the course. These two control variables were added as covariates in the analysis, so that our final design was a three-way between-within ANCOVA (analysis of covariance) design.

Design layout. We then considered how we wanted to frame our hypotheses. With respect to Instructor Gender we had a choice between (a) weighting each gender equally and (b) weighting the genders proportionally to the university-wide breakdown of male/female instructors. Each of these options addresses a slightly different hypothesis. Option (a) examines whether there are instructor-gender differences for equal numbers of male and female instructors. Option (b) examines whether there are instructor-gender differences in a population (of all present and, presumably, future UBC instructors) in which the genders are represented in the unequal proportions found at UBC. We decided on Option (b). Thus, we created what is known as a *proportionally-balanced design* that can be depicted as follows in Table 1 (numbers in the cells are the number of instructor/course units).

Table 1
Layout of the 2 × 3 ×2 Between-Within ANCOVA Design with Numbers of Instructor/Course Sections Indicated in the Cells

		Instructor Gender					
		Male Instructor			*Female Instructor*		
	Field of Study:	*Human's*	*Soc. Sc's*	*Science*	*Human's*	*Soc. Sc's*	*Science*
Student Gender	*Male Respondents*	114	114	114	59	59	59
	Female Respondents	114	114	114	59	59	59

From this layout, it can be seen that we had a total of 342 instructor/course units that had a male instructor, and 177 units that had a female instructor, for a total of 519 data points. This ratio of male to female instructors is 1.93:1, representing 34% female and 66% male instructors, a figure that is very close to the ratio in the UBC instructor population. Each instructor/course unit had two evaluations for each item and the 6-item average—one from male respondents and one from female respondents. It will be noted that this design—with the cell frequencies noted—is an *orthogonal* design, with each effect tested completely independent of all other effects.

Instructor/course units were selected quasi-randomly within the Field of Study factor categories. As an example, we selected the 177 sections of Humanities courses so that the proportions of Art History & Visual Arts, Asian Studies, English, etc. courses in the sample closely mirrored the corresponding proportions in the population of all Humanities courses taught in 2008-09. Thus, if, for example, the population proportion of a particular subject in the Humanities offerings were 20%, we would select 35 sections (approxi-mately 20% of 177) of that subject randomly from the total number of sections of that subject offered in the year. Similarly, the course year (1^{st}, 2^{nd}, …, 4^{th}; *no graduate*) proportions in the sample were in approximate correspondence with those in the full slate of courses offered within the disciplines.

Dependent Variables

The dependent variables were the six UMIs. In addition, we took the average of the six as an overall measure that could be expected to capture the overall perceived quality of the instructor/course unit. As noted earlier, the actual numbers analyzed were the *means*--calculated over the individual ratings provided by the students in the class via the new online administration system--on the six UMIs and their average.

Data Analysis

We first performed a multivariate analysis of covariance, with the six UMIs the multiple dependent variables. For some of the effects, this MANCOVA yielded highly significant results. For these effects, univariate ANCOVAs were conducted, and in some cases these latter analyses were followed up with multiple comparisons and/or analyses of simple main effects.

Results and Discussion of Analyses of the Overall Averaged Dependent Variable, together with Selected Results for Individual UMIs

Testing of ANCOVA Assumptions

Designs like the present one have a number of assumptions that must be met for the results to be precise—*i.e.,* the *p*-values presented with the results are precise and our actual alpha levels are the nominally-correct ones. These assumptions (homogeneity of variance and homogeneity of regression) were tested and found to be tenable in the present analysis. (The usual repeated-measures assumption of sphericity did not apply in this study since there were only two levels of the within-subjects factor.) Therefore, the *p*-values associated with the results that follow are accurate.

Preliminary Multivariate Analysis of Covariance (MANCOVA)

Before we proceeded to univariate tests on the dependent variables of interest, an overall MANCOVA was conducted on the means on UMIs 1 – 6, using the experimental design illustrated in Table 1. Thus, with Class Size and Mean Course Grade covaried, the six UMIs were simultaneously analyzed. Results of this MANCOVA revealed statistically significant multivariate main effects for all three factors:
(a) Instructor Gender, [$F(7, 505) = 7.82$, $p < .00001$]; (b) Field of Study, [$F(14, 1{,}010) = 6.94$, $p < .00001$]; and Student-Respondent Gender, [$F(7, 505) = 3.84$, $p = .0004$].

The multivariate three-way interaction effect was found to be nonsignificant [*F*(14, 1,010) = 1.59, *p* = .0751], as were two multivariate two-way interaction effects: (a) Instructor Gender × Field of Study [*F*(14, 1,010) = 1.10, *p* = .3576] and (b) Field of Study × Student-Respondent Gender [*F*(14, 1,010) = 1.33, *p* = .1853]. However, the remaining two-way multivariate interaction effect, that between Instructor Gender and Student-Respondent Gender, was found to be statistically significant [*F*(7, 505) = 5.33, *p* < .00001]. All multivariate tests were conducted using the likelihood-ratio test (Wilks' Lambda).

The MANCOVA thus suggested that there were significant effects to be found with respect to the individual UMIs and that individual univariate ANCOVAs would provide the necessary more finely-grained results by which to best understand the data. Rather than doing so for each dependent variable in turn, however, which would produce a piecemeal presentation, we instead constructed a summary dependent variable: the average of the six UMIs, and subjected scores on this aggregated measure to an ANCOVA using the same experimental design as used in the MANCOVA and detailed in Table 1. Significant effects found for the averaged UMI variable that were also found with a number of UMIs are noted briefly with respect to these UMIs as well.

ANCOVA Results with Overall Score (Average of the 6 UMIs)

Beginning, then, with this overall dependent variable—which draws from all six UMIs—we present the results of the ANCOVA in Table 2.

Table 2

Results of Analysis of Covariance of the Overall Dependent Variable—Average UMI

Source of Variation	***df***	***MS***	***F***	***p***
Between-Inst./Course Units				
A – Instructor Gender	1	1.679	5.04	.0252
B – Field of Study	2	1.053	3.16	.0433
A × B Interaction – Instructor Gender × Field of Study	2	.078	.24	.7867
Inst/Course units w/in Groups (Error)	511	.333		
Within-Inst./Course Units				
C – Student-Respondent Gender	1	.087	1.61	.2051
A × C – Instructor Gender × Student-Respondent Gender	1	.708	13.12	.0003
B × C – Field of Study × Student-Respondent Gender	2	.167	3.09	.0464
A × B × C Interaction	2	.014	.26	.7712
C × Inst/Course units w/in Groups (Error)	511	.054		

Covariates: Class Size and Mean Course Grade

Main effects

From Table 2, we can see that we have two significant main effects (if we use, as our alpha level, .05), both involving our two between-subjects factors: (a) Instructor Gender and (b) Field of Study. The third main effect, Student-Respondent Gender, was found to be nonsignificant (even though this had been significant in the MANCOVA).

To provide meaning to the statistical results involving the two significant main effects in Table 2, we present some relevant aggregated (over the other two factors), adjusted (for the covariates) mean values below in Table 3.

Table 3

Adjusted Means on the Overall Dependent Variable—Average UMI Score—for Instructor Gender and Field of Study, Aggregated over the Other Factors in the Design

	Effect Tested				
	Instructor Gender		**Field of Study**		
	Male	**Female**	**Humanities**	**Social Sciences**	**Science**
Overall Adjusted Mean	4.011	4.095	4.078	4.089	3.993

It thus appears that, at least with respect to this aggregated dependent variable, ratings for female instructors were, on average, significantly higher than those for male instructors.

With the Field of Study factor, the significant main effect was followed up with multiple comparisons; no difference whatsoever was found between the Humanities and Social Sciences in mean ratings, and a difference that did not rise to statistical significance between the Humanities and Science. Only the difference between the Social Sciences and Science was statistically significant and only with p = .04. For this reason and because the raw scale-point difference between the Social Sciences and Science mean on this dependent variable was small (**.096**) we are not inclined to put much weight on the findings for the Field of Study factor in connection with the overall averaged UMI variable.

We caution the reader to consider obtained results in this study from the perspective of *practical significance* and not merely *statistical significance*. For example, with the Instructor Gender results in Table 3, we have a gender difference between the adjusted means of **.084**, which is—as seen from Table 2—statistically significant (p = .0252). The reader should judge, however, just how much practical importance attaches to this difference (as was the case above with the Social Sciences *vs*. Science means).

Practical significance can be assessed in either the raw scale-point metric (as we have above) or the standardized effect-size metric, which is simply a transformation of the former, or division by an estimate of the standard deviation of the distribution of scores (in this case instructor/course means). This latter index of practical significance has the advantage of being universal, or independent of the magnitudes of the standard deviations. In the present context, however, it may offer little advantage over the raw scale-point difference. We mention the standardized effect size index because for comparisons involving two means, social scientists have become familiar with a system of characterizing indices of practical significance as *small* (standardized effect sizes less than or equal to approximately .20), *medium* (around .50) and *large* (.80 or larger). In this system, both differences noted above (Instructor Gender and Field of Study) represent *small* standardized effect sizes of around .20.

We will return to a brief discussion of these two main effects as they were manifested with UMIs 1–6 in a later section.

Interaction effects

Another reason not to focus too much on the findings for both main effects is the existence of interaction of each of Instructor Gender and Field of Study with Student-Respondent Gender, particularly the former interaction, as can be seen from the *p*-values in Table 2. These statistically significant interaction effects indicate that no unqualified statements about the effects of either factor can be made, and that we must explore how the Student-Respondent Gender factor plays a part in connection with each.

Instructor Gender × Student-Respondent Gender Interaction. This need for further qualification is particularly salient with the Instructor Gender factor where the Instructor Gender × Student-Respondent Gender interaction effect is so highly significant (Table 2). To see this, perusal of the Instructor Gender × Student-Respondent Gender cell means is instructive, as displayed in Table 4.

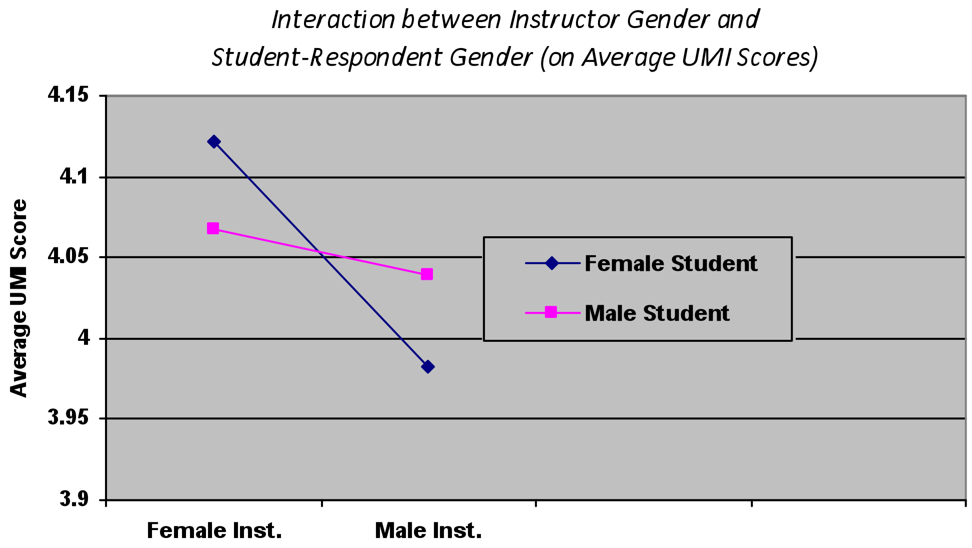
Table 4

Adjusted Cell Means in the Instructor Gender × Student-Respondent Gender Summary Table

		Instructor Gender		
		Female Instructor	*Male Instructor*	*Adjusted Student-Resp. Means:*
Student-Respondent Gender	*Female Student-Respondent*	4.122	3.983	4.0304
	Male Student-Respondent	4.068	4.039	4.0490
	Adjusted Instructor Means:	4.095	4.011	4.0397

These cell means are presented graphically in Figure 1:

Figure 1



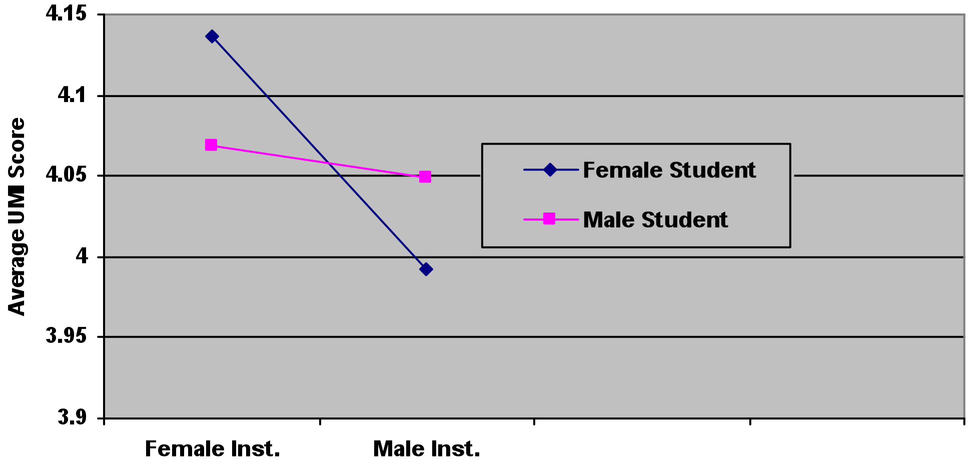
It is clear, from Table 4 and Figure 1, that, although there is an overall difference in favor of female instructors, that difference is coming almost solely from the ratings provided by the female student-respondents. Examining the simple main effects holding Student-Respondent Gender constant, we find that this particular difference (Female Instructors *vs*. Male Instructor as rated by *female* student-respondents) is highly significant [*F*(1, 505) = 12.08, *p* = .0006], whereas the other simple effect (Female Instructors vs. Male Instructor as rated by *male* student-respondents) falls far short of significance [*F*(1, 505) = .44, *p* = .495]. We thus see no evidence whatsoever that male student-respondents tend to rate the instructors differently as a function of instructor gender, whereas there is very strong evidence that female student-respondents do rate instructors differently by gender, with the higher ratings going to female instructors. On average, we see (from Table 4) a difference in ratings for female student respondents of **.139** raw scale points, with an accompanying standardized effect size of .30–.35—a difference that would be regarded as approaching practical significance. For the male student respondents, the corresponding raw scale-point difference was only **.029**, or of no practical importance whatsoever (as well as being far from statistically significant).

[Another observation from Table 4 and Figure 1 is that male and female student-respondents gave very similar ratings when we collapse over Instructor Gender. The means of 4.030 (for the female student-respondents) and 4.049 (male student-respondents) are nowhere near significantly different—as was seen in the row in Table 2 for the Student-Respondent Gender main effect.]

To provide some additional support to the above findings for the overall averaged UMI variable, we present below, in Figure 2, the corresponding results for UMI 6, which states "Overall, the instructor was an effective teacher."

Figure 2

Interaction between Instructor Gender and Student-Respondent Gender (on UMI 6 Scores)



With UMI 6, the two simple main effects are almost identical to those with the averaged UMI variable, with that for female student-respondents highly significant [*F*(1, 505) = 8.64, *p* = .0034, and a raw scale-point difference of .144], and that for male student-respondents resoundingly nonsignificant [*F*(1, 505) = .15, *p* = .6959]. Similar disordinal interaction effects were found for the other UMIs as well.

UMIs 1 – 5. As for the other UMIs, we found that with UMIs 2 and 3, precisely the same pattern emerged as noted above for the overall averaged UMI and for UMI 6—a significant difference in favor of female instructors when rated by female student-respondents, but a resoundingly nonsignificant difference between the instructor genders when rated by male student-respondents. With UMI 4, there were no differences in Instructor Gender ratings when rated by either gender of student-respondent.

We note, in closing this discussion of interaction effects involving the Instructor Gender factor, that this factor did not interact at all with the Field of Study factor. All UMIs exhibited *p*-values ranging from .53 to .89, with the averaged UMI measure exhibiting a *p*-value of .79. This indicates that there were absolutely no differential effects involving the Instructor Gender factor when going from one field of study to another. Further, as noted earlier, the three-way interaction was resoundingly nonsignificant in the analysis of the overall averaged UMI measure (*p* = .77, as seen in Table 2), and similar results were obtained with each UMI in turn.

Field of Study × Student-Respondent Gender Interaction. The reader will recall that the other main effect that was significant was that involving the Field of Study factor (Table 2). However, as with the main effect for Instructor Gender, this effect needs qualification because of the interaction between the Field of Study and Student-Respondent Gender factors. The cell means that help us to see the nature of this interaction, as it occurred with the overall averaged UMI measure, follow in Table 5.

Table 5

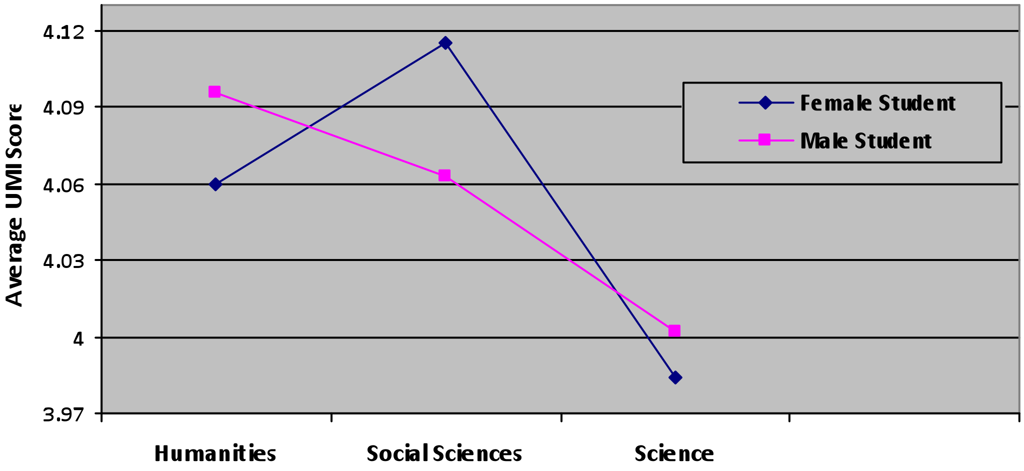
Adjusted Cell Means in the Field of Study × Student-Respondent Gender Summary Table

		Field of Study			
		Humanities	*Social Sciences*	*Science*	*Adjusted Student-Resp. Means:*
Student-Respondent Gender	*Female Student-Respondent*	4.060	4.115	3.984	4.053
	Male Student-Respondent	4.096	4.063	4.002	4.054
	Adjusted Field of Study Means:	4.078	4.089	3.993	4.0535

These cell means are shown graphically in Figure 3.

Figure 3

Interaction between Field of Study and Student-Respondent Gender (Average UMI Scores)



Analyses of the simple main effects involving the Field of Study factor for each student gender yielded a significant result for the female student-respondents [*F*(2, 511) = 3.56, *p* = .0291], but not for male student-respondents [*F*(2, 511) = 1.57, *p* = .2093]. Follow-up pairwise multiple comparisons on the female student-respondent means revealed that only the difference between the means for Social Sciences and Science was statistically significant [*F*(1, 511) = 7.15, *p* = .0077]. The corresponding difference between Social Sciences and Science for the male student-respondents was nonsignificant [*F*(1, 511) = 1.43, *p* = .2322], as was that between the Humanities and Science groups [*F*(1, 511) = 2.98, *p* = .0847].

Analyses of Main Effects with UMIs 1 – 6

When considering the Instructor Gender factor, the most informative interpretation with most dependent variables is provided by the interaction effect between Instructor Gender and Student-Respondent Gender. However, with UMIs 1 (in particular) and 5, the main effect of Instructor Gender is the more potent one. In Table 6, the Instructor Gender means are given for these two UMIs.

Table 6

Adjusted Instructor Gender Cell Means for UMIs 1 and 5

UMI 1: **The instructor made it clear what students were expected to learn.**

Male	*Female*	*Overall (Unweighted) Mean*
4.012	4.153	4.083

UMI 5: **The instructor showed concern for student learning.**

Male	*Female*	*Overall (Unweighted) Mean*
4.139	4.258	4.199

With each UMI, the main effect for Instructor Gender was highly significant. For UMI 1: *F*(1, 511) = 13.74, *p* = .0002; for UMI 5, *F*(1, 511) = 10.69, *p* = .0011. For each of these UMIs, female instructors were more highly rated than male instructors. The standardized effect sizes with respect to these two UMIs average approximately .32 (corresponding to an average raw difference of .131 scale points), indicating effects that are beginning to reach non-negligible proportions.

With respect to the Field of Study factor, the interaction effects with the Student-Respondent Gender factor were largely nonsignificant for the individual UMIs, suggesting that it might be more informative to examine the Field of Study main effects for UMIs 2 and 3, with which highly-significant results were obtained. In Table 7, the Field of Study means appear for these two UMIs.

Table 7

Adjusted Field of Study Cell Means for UMIs 2 and 3

UMI 2: **The instructor communicated the subject matter effectively.**

Humanities	*Social Sciences*	*Science*	*Overall Mean*
4.090	4.105	3.903	4.033

UMI 3: **The instructor helped inspire interest in learning the subject matter.**

Humanities	*Social Sciences*	*Science*	*Overall Mean*
4.015	4.067	3.890	3.991

With UMI 2, the statistical results were $F(2, 511) = 9.38$, $p = .0001$, and with UMI 3, we had $F(2, 511) = 6.32$, $p = .0019$. Follow-up multiple comparisons on these main-effect means revealed that with both UMIs 2 and 3, the differences were significant between each of Humanities and Social Sciences on the one hand and Science on the other. The difference between Humanities and Social Sciences, however, with each UMI was nonsignificant. Thus on these two UMIs, the Humanities and Social Science means were not different from each other, but each was significantly higher than that for Science. With these UMIs, the effect sizes were somewhat larger than we found with the averaged UMI dependent variable. If we take the mean of the Humanities and Social Sciences mean values on UMI 2, for example, we get a value of 4.0975, and the raw scale-point difference between this value and the 3.903 for Science is **.1945**, which corresponds to a standardized effect size of approximately .45, and which would be classified as a medium-sized effect size or one that is not negligible. The parallel analysis with UMI 3 yields a raw scale-point difference of **.151**, or a standardized effect size of approximately .36 between Humanities/Social Sciences, on the one hand, and Science, on the other—again somewhat greater than a small effect size. As noted before, however, the reader is free to regard these differences as worthy or not of further consideration.

Relationships between the Covariates and the Dependent Variables

The covariates used in the ANCOVAs reported above were correlated with the dependent variables. Because of the very large number of correlations possible with this data set, we have had to find more-aggregated summary values to present here. In the interests of economy of presentation, we have aggregated all 519 instructor/course units as the units of analysis in the correlational analyses, thus risking a small degree of between-groups correlation to creep into the reported values. We will comment on this briefly after presentation of these summary correlations, appearing below in Table 8.

Table 8

Correlations between the Covariates and the Dependent Variables
(n = 519 Instructor/Course Units

	Dependent Variable	
Covariate	*UMI 6*	*Average of 6 UMIs*
Class Size	–.23	–.27
Mean Course Grade	.26	.30

Note: All associated p-values < .0001.

The values in Table 8 are quite representative of the individual correlation coefficients we obtained in each of the six Instructor Gender × Field of Study cell. With respect to the Class Size covariate, the average correlation with UMI 6 was –.26, with all of six correlations less than – .24 except for the Male Instructor/Science cell, where the correlation was an anomalous –.02. In general, the Class Size *vs*. UMI 6 correlations were larger in absolute value for the female instructors (average $r = –.35$) than for the male instructors (average $r = –.17$), with this average difference approaching statistical significance (and actually reaching it with an alpha level of .05).

The pattern of correlational results with the Average UMI dependent variable was very similar, with the average across the six Instructor Gender × Field of Study cells equal to –.30, with the mean for female instructors –.39 and for male instructors –.21. We thus might see, as a convenient summary value for the correlation between class size and rated instructor performance with the present data, a correlation on the order of –.25 to – .30. This value makes good sense when we reflect on the variables involved in this correlation.

With respect to the Mean Course Grade covariate, our expectations would likely be a small-to-moderate positive correlation, and the results in Table 8 are consistent with this. The average correlation between Mean Course Grade and UMI 6 scores, over the six cells in the design, was .25, with the mean *r* for female instructors .31 and for male instructors .19. As for the other dependent variable, Average UMI, the correlations with Mean Course Grade averaged .29, with the mean *r* for female instructors .35 and for male instructors .22. As with the other covariate, Class Size, there was one anomalous cell among the six—Male Instructor/Social Sciences—in which the correlations between Mean Course Grade and the two dependent variables were not different from zero. Nonetheless, we might see, as a sort of rounded summary value here for the correlation between Mean Course Grade and rated instructor performance, something on the order of .25 – .30.

One detail that should be noted in the just-preceding results is that the Mean Course Grade variable is a proxy for, but not exactly the same thing as, the grades that the students expect to see in the course. In the present study, a better covariate might have been the average expected (by the students) course grade since that is the perception that could be expected to influence instructor performance ratings. This would have necessitated an additional procedure in the study—soliciting expected grades from the students while the course was in progress—and without that intervention, our best proxy would seem to be the *actual* average course grade. Our assumption here would be that by the time the course evaluations are performed, students have a pretty good idea of the distribution of final course grades.

It is probably worth mentioning that the covariates in this study did not tend to be associated to any significant degree with the three factors in the analyses. This meant that the adjustment to the marginal and cell means arising from the covariates was quite minimal, and the main findings were very similar to those found in a standard analysis of variance performed on the data (without the covariates). Nonetheless, as we can see from the correlational results above in Table 8, the covariates did correlate reasonably substantially with the dependent variables, and the analyses performed in this study were more powerful as a result. Perhaps more conceptually important is the fact that neither covariate—Class Size and Mean Course Grade—was allowed to influence the central findings at all. These extraneous variables (for the present purposes) were held constant, and thus the main findings should be understood as completely independent of, and uninfluenced by, Class Size and Mean Course Grade.

Results from Comparing between-UMI Mean Levels

Finally, it might be of interest to consider the overall UMI means—based on all 519 instructor/course units. These appear in Table 9. To make a reading of Table 9 more meaningful, we again remind readers of the content of the UMIs:

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.

Table 9

*Unadjusted Means and Standard Deviations for the Six UMIs (*n *= 519)*

		Unadjusted Mean	 Std. Dev.
UMI:	1	4.08	.41
	2	4.03	.50
	3	3.99	.49
	4	3.95	.40
	5	4.20	.40
	6	4.06	.52
Averaged UMI:		4.05	.41

We note in passing that the overall university-wide mean over 6,636 instructor/course units from all faculties, including Arts and Science, on UMI 6 for the 2008-09 academic year was 4.12, and the standard deviation was .57. We also note that the means in Table 9 are not adjusted for the effects of the covariates. This is because we felt that they would have more descriptive value this way and could be better compared with corresponding (also unadjusted) values for the university as a whole, perhaps arising in previous and future academic years. In addition, since the comparisons deriving from Table 9 do not involve the experimental factors in this study, improving the inferential properties of the significance tests involving these factors was irrelevant.

The means in Table 9 provide information about which aspects of teaching are being most favorably and least favorably perceived by student raters. The overall mean rating is highest (at 4.20) for UMI 5—"The instructor showed concern for student learning." On the basis of paired-comparison *t*-tests, UMI 5 was found to manifest significantly higher rating means than each of the remaining five UMIs (conservative tests were conducted comparing among the UMI means, with alpha levels of .005). At the other end of the continuum, the lowest overall mean rating (3.95) was found for UMI4—"Overall, evaluation of student learning (through exams, essays, presentations, etc.) was fair." The UMI 4 mean was found, from paired-comparison *t*-tests, to differ significantly from those of all the remaining UMIs except for UMI 3 (which difference approached, but did not quite reach statistical significance). In a way, this is not surprising, in that it is probably the grading (and giving students a grade that reflects what they believe they deserve) that is most salient to students and about which many students would be most critical.

Whether or not this lowest rating indicates the need for more attention being paid to grading practices among instructors as a whole is unclear from these results. It may be, instead, that this aspect of teaching will always be the one most criticized no matter how well it is done. The other somewhat lower-than-average rating, that for UMI 3—"The instructor helped inspire interest in learning the subject matter"—may also be worth noting. The mean rating on UMI 3 was significantly lower than those of all other UMIs except for UMI 4. It is probably the case that actually inspiring students is a higher-order goal that is difficult to achieve for most instructors. It is likely the case that what might be conceptualized, perhaps, as lower-order goals of careful preparation (UMIs 1, 2, and 4) and concern for learning (UMI 5) are easier to achieve and could be seen as occupying a lower stratum in a hierarchy of goals that we might visualize for university instructors.

In the development of university teaching skills, we might be best served by making sure that the lower-order goals are reached first, saving the inspirational aspects of teaching until the easier-to-achieve aspects have been mastered. This is Gary Poole's—and TAG's—domain, however, and we won't speculate further. In any case, we might view the gradient of means in the above table as something of a template for instructor development. It is our hope that the results obtained through the UMIs can be used to facilitate teaching-enhancement initiatives by TAG.

Summary and Conclusions

Design

Three-way analyses of covariance were performed on SEoT UMI data collected during the 2008-09 academic year from instructor/course units in three different fields of study at UBC: the Humanities, the Social Sciences, and Science. The covariates were Class Size and Mean Course Grade. UBC population proportions of female and male instructors were preserved in the sample of 519 instructor/course units. The exact layout of this design can be seen in Table 1. Mean ratings were obtained, for each instructor/course unit, from both female and male student-respondents. The overall analysis process began with multivariate analyses of covariance and then proceeded to univariate analyses when the multivariate results indicated further probing of the data. Although the main focus of the analyses was the aggregated, overall UMI variable (the average of the six UMIs), some selected analyses of the individual UMIs were performed when the preceding analyses suggested the need for more finely-grained examination.

Overall Performance Levels

Before summarizing the findings, we might note that the sample-wide level of rated instruction would have to be considered high. Further, we have seen above that this is reflected to an even greater degree when we consider the university-wide results. If we focus on just UMI 6, which is concerned with students' overall impressions of the quality of instruction, we see averages of 4.06 (this sample) and 4.12 (university as a whole). These averages reflect good perceived teaching at this university and, incidentally, are very similar to the corresponding UMI 6 averages that were obtained through the previous pencil-and-paper administration mode.

Sample Representativeness

In addition, the similarity of the UMI 6 mean for both groups of instructor/course units (present sample and larger university-wide aggregation of which the present sample is a part), along with an even greater similarity in their standard deviations (.52 *vs.* .57) suggests that the present sample is quite representative of the larger set of all instructor/course units found in the 2008-09 offerings.

Noteworthy Effects Found

In the analyses of the overall averaged UMI mean scores, we found two main effects: (a) for Instructor Gender and (b) for Field of Study. These main effects, however, were found to be complicated conceptually by the interactions between each and the Student-Respondent Gender factor. The statistical results appear in Table 2. We note here that any *means* discussed earlier and in the sequel are to be understood as *adjusted* (by the covariates) means. As noted earlier, the question of the *practical* significance of these main-effect differences must be considered by the reader.

The most highly (statistically) significant finding in the present study was the Instructor Gender × Student-Respondent Gender interaction effect. This can be seen in Table 2 for the averaged UMI dependent variable and also in the results for UMIs 2, 3, and 6. In these cases, the female instructor mean was significantly higher than the male instructor mean when the ratings were those of female student-respondents, but the corresponding difference between the instructor genders was nonsignificant when the ratings were those of male student-respondents. Aspects of this effect can be seen in Table 4 and Figures 1 and 2.

In other cases, though (UMIs 1 and 5), both female and male student-respondents rated female instructors more highly, on average, than they rated male instructors. These main-effect results can be found in Table 6. In all of UMIs 1, 2, and 5, and the overall averaged UMI measure, this significant Instructor Gender main effect was found. Thus, we might summarize all of this by noting that, in general, we may say that female instructors were more highly rated than male instructors, but in several cases this resulted from the ratings provided by female student-respondents only.

With respect to the Field of Study factor, when the overall averaged UMI dependent variable was analyzed, there was a significant difference between the means for Social Sciences and Science, in favor of the former, but only on the basis of ratings provided by female student-respondents. There were no significant differences among the three fields of study from ratings provided by male student-respondents. Thus, although the overall main effect for Field of Study was significant for this averaged dependent variable, this effect must be understood in terms of the Field of Study × Student-Respondent Gender interaction, as detailed above in this paragraph. The specifics of this analysis can be found in Table 5 and Figure 3.

When considering UMIs 2 and 3, however, Field of Study was found not to interact with Student-Respondent Gender, and the Field of Study factor instead yielded a highly-significant main effect. The nature of this effect was that ratings in the Humanities and Social Sciences did not differ from each other, but that each differed significantly from the ratings found in Science, with the Humanities/Social Sciences ratings higher. The specifics of these results can be found in Table 7. Here the differences were approaching practical-significance levels.

Relationships with the Two Covariates

The two covariates, Class Size and Mean Course Grade were largely unrelated to the three independent variables, but were moderately correlated with the dependent variables. Class size was found to be negatively correlated with mean ratings on UMI 6 and for the averaged UMI dependent variable. These Class Size *vs.* Dependent variable correlations were in the –.20 to –.30 range. Positive, and slightly higher, correlations were found between Mean Course Grade and the dependent variables (in the .25 to .30 range).

Differences among Mean Levels on the Six UMIs

Among the six UMIs, UMI 5 manifested the highest mean in this sample and UMI 4, the lowest. The gradient of the UMI means in Table 9 may have useful implications for teaching improvement, and this possibility is discussed in the text following the results in Table 9.

___