Course Evaluation Project Pilot Test — Data Analysis Report

Statistical Consulting and Collaborative Research Unit
(Martin Lysy and Feiyu Zhu)
and The Course Evaluation Project Team, Phase 2
(Sonya Buffone, with support from David DeVidi, Kofi Campbell, Andrea Chappell, Donna Ellis, Matthew Gerrits, Jasmin Habib, and Clarence Woudsma)
University of Waterloo

March 27, 2020

Summary. In late November 2018 the draft Student Course Perception Survey (SCP) was piloted across campus. In total, 41,737 SCP surveys from 2,196 courses were submitted by students from across the six Faculties and two of the Affiliated Institutions (Renison and Conrad Grebel). This research report provides a detailed summary of the findings from the analysis of the SCP pilot test responses. The report is organized into 11 sections: Sections 1-5 provide an overview of the Course Evaluation Project and outline the research goals of the SCP pilot test. Section 6 defines the levels of measurement used throughout the analyses, while Section 7 provides some key descriptive statistics from the pilot test. In Section 8 we focus on differences in scores between male and female instructors as accounted for by several explanatory variables at a time. Section 9 presents findings from regression analyses that explore each core item individually at both the student and course level. We also conducted a qualitative factor analysis to determine potential theoretical "groupings" of the nine SCP items. The analysis provides evidence to support combining the core items to create two composite measures. In Section 10, we turn our attention to an analysis of our proposed composite measures. Section 11 contains some summary remarks. Overall, our analyses suggest that the instrument correlated as expected with explanatory variables (e.g., attendance, class size, expected grade). Much of the attention in the analysis is devoted to the question of how SCP scores in courses differ between male and female instructors and the extent to which these scores are associated with a variety of other factors. In general, findings reveal gender associations are very small. For a few important cases, evidence is found of more substantial differences, though the evidence in these cases is weak due to the small number of instructors involved. We recommend further investigation of these cases in the future, and emphasize the need for caution when interpreting results for those cases if the SCP goes into use on campus.

Contents

1 Introduction & Background **2**

2 Overview of the Student Course Perception Survey Pilot Test **3**

3 Research Aims **3**

3.1 Association between SCP Scores and Various Predictors 4

3.2 Composite Metrics 5

3.3 Evidence to Inform Educational Toolkits 5

4 Teaching Effectiveness at the University of Waterloo 5

5 Uncertainty Quantification 7
5.1 Sampling Error and the Population of Interest 7
5.2 Margin of Error 8
5.3 Other Possible Modelling Choices 8

6 Levels of Analysis 9

7 Demographic Statistics 10
7.1 Faculties 10
7.2 Student Gender 11
7.3 Instructor Gender 12
7.4 Non-Response and No Basis 16

8 Differences by Instructor Gender 18

9 Regression Analysis 27
9.1 Student-Level Analysis 27
9.2 Course-Level Analysis 33

10 Composite Metrics Analysis 37

11 Summary Remarks 40

A Course-Level Qualitative Factor Analysis 47
A.1 Correlation Matrix of Response Items 47
A.2 Factor Analysis Model 47
A.3 Number of Factors 49
A.4 Factor Structure 49
A.5 Factor Analysis Results 55

B Relationships Between Explanatory Variables and Composite Measure Scores 57

C Bivariate Regression Plots by Gender 57

D Student-Level Regressions 58

E Course-Level Regressions 67

F SCP Pilot Test Survey Items Fall 2018 72

1 Introduction & Background

The University of Waterloo has been exploring the potential for a new campus-wide course evaluation model for some time. In 2014, the Associate Vice-President, Academic established the Course Evaluation Project Team (CEPT1) to update course evaluations to align with current institutional teaching and learning priorities and best practices.

In November 2016, CEPT1 released a report recommending that the University adopt a cascaded course evaluation instrument to measure student course perceptions (SCPs). As part of this cascaded model it was recommended that all Faculties include a common set of institutional core items that focus on three important dimensions of effective teaching, as identified in the literature: Course Design, Course Delivery and the Learning Experience[1]. CEPT1 also drafted sample questions for each of the three dimensions of teaching effectiveness. The campus community was consulted for feedback on the recommendations in the draft report released in Fall 2016. Further review of the literature, extensive project team discussions, and the results of the Fall 2016 consultation process led to a revised list of recommendations released by CEPT1 in a final report in April 2017[2].

Senate approved the report and its recommendations in September 2017. The key recommendation was that the University proceed with a cascaded model in all Faculties. To reach this goal, it was recommended that a "Phase 2" (CEPT2) of the project be carried out to review, if necessary modify, and pilot test the draft SCP instrument developed by CEPT1.

In early 2018, the new CEPT2 committee was struck, and a project coordinator/research specialist was hired to support the project. In July 2018, CEPT2 held a series of focus groups with undergraduate students from each of the six Faculties to help inform its decision about what modifications, if any, to make to the draft instrument[3]. Analysis of the focus group data[4], review of the literature, and project team discussions led to minor revisions of two items in the draft instrument developed by CEPT1. To provide important Waterloo specific data and gain a better understanding of the survey's performance, CEPT2 piloted the proposed SCP survey in November 2018. This research report provides a detailed overview of the pilot test for the SCP survey and a summary of the findings from an analysis of pilot test results.

2 Overview of the Student Course Perception Survey Pilot Test

In late November 2018, the draft SCP survey was piloted across campus (see Appendix F which outlines the questions used in the SCP pilot test survey)[5]. The pilot test ran parallel to the regularly scheduled end-of-the-term course evaluations in all classes scheduled for the Fall 2018 term, both face-to-face and online, for which official surveys were running on the Evaluate system. In total 41,737 pilot test surveys from 2,196 courses were submitted by students from across the six Faculties and two affiliated institutions (Renison and Conrad Grebel). Given that students completed 72,652 traditional end-of-term evaluations, the calculated response rate was relatively high (58%). This is a strong response given that it was made clear to students that the results of the Pilot Test were only to be used to investigate the new instrument and that completing it was extra work after completing the official survey.

3 Research Aims

The pilot test was designed to address three key research aims:

[1] https://uwaterloo.ca/associate-vice-president-academic/sites/ca.associate-vice-president-academic/files/uploads/files/final_ceptdraftreportnov7.pdf

[2] https://uwaterloo.ca/associate-vice-president-academic/sites/ca.associate-vice-president-academic/files/uploads/files/ceptdraftreportfinalapril27.pdf

[3] https://uwaterloo.ca/waterloo-course-evaluations/focus-group-methodology

[4] https://uwaterloo.ca/waterloo-course-evaluations/focus-group-analysis-summary-key-themes

[5] https://uwaterloo.ca/associate-vice-president-academic/frequently-asked-questions

1. Determine the strength of association between SCP scores and predictive variables (e.g., gender, class size, etc.).
2. Group SCP items into composite scores measuring each of the three dimensions of learning experience identified by CEPT1: Course Design, Course Delivery, and Learning Experience.
3. Develop a statistical toolkit for the SCP instrument.

3.1 Association between SCP Scores and Various Predictors

First, we sought to analyze the extent to which certain variables as identified in the literature (gender, class size, etc.) correlate with SCP scores. (Of course, causality cannot be inferred based on the results of this pilot test.)

A large body of research examines how student ratings are influenced by factors outside of the control of instructors or unrelated (or at least not obviously related) to teaching effectiveness (e.g., class size, instructor gender, grade expectations, etc.). Bias, with respect to student evaluations, is best understood as existing when "student, teacher, or course characteristics affect the evaluations made, either positively or negatively, but is unrelated to any criteria of good teaching" [Marsh (2007), p.498)[6].

In general, the research literature on this topic is complex, with limited consensus on how different factors influence SCP scores[7]. Despite a substantial (and empirical) body of academic research to support the validity and reliability of student evaluations of teaching (Benton and Cashin, 2012, 2014; Beleche et al., 2012; Centra, 2009; Feldman, 1993; Hativa, 2014; Marsh, 2007; Theall and Franklin, 2001; **?**; Willits and Brennan, 2017; Winer et al., 2012), in recent years, claims of bias have received significant attention in both popular press (e.g., *University Affairs*, *The Chronicle of Higher Education*, etc.) and the academic literature (see for example, (Boring et al., 2016; MacNell et al., 2015; Stark and Freishtat, 2014). As such, one key aim of the pilot test was to examine the extent to which certain factors discussed in the literature are associated with student course perception surveys at Waterloo.

It is important to highlight that the performance of any institutional tool must be understood within the context in which it is utilized. Empirical research at other U15 institutions has found that some factors (e.g., gender) identified in the literature as biasing evaluation scores do not appear to have a substantial effect on course evaluation scores in their local context, while other factors are found to be of more significance (e.g., class size)[8]. The purpose of the pilot test was to explore the extent to which key factors identified in the literature correlate with student course perception scores at the University of Waterloo. Specifically, we considered the following predictors of SCP scores:

1. **Student-Level Variables:** Gender, expected grade (self-reported), perception of the course workload, course attendance (self-reported).

2. **Instructor-Level Variables:** Instructor gender, instructor rank.

3. **Course-Level Variables:** Class size, class type (i.e., elective or required), attendance type (i.e., online or in-class), Faculty of course offering.

[6] Arguably, by this definition, as Benton and Cashin (2014) explain, the "correlations between student ratings and class size, or between student ratings and student interest in the course, are not biases because students in small classes and students who are interested in the subject matter actually do tend to learn more and, hence, give their teachers higher ratings" (p.295).

[7] For an informative discussion refer to this blog post: http://cte.rice.edu/blogarchive/2015/07/09/studentevaluations.

[8] https://mcgill.ca/mercury/files/mercury/course_evaluation_results_interpretation_guidelines.pdf; https://teaching.utoronto.ca/wp-content/uploads/2018/09/Validation-Study_CTSI-September-2018.pdf.

To explore the relations between these variables and SCP scores we examined univariate, bivariate and multivariate statistics, including regression models at the student and the course levels.

3.2 Composite Metrics

Since one of the goals of the pilot test was to assess whether the SCP items could be grouped into composite measures, we performed a statistical factor analysis on the SCP instrument. This allowed us to determine whether/how strongly the items included in the SCP survey "grouped" into the three different theoretical constructs (i.e., Course Design, Course Delivery and Learning Experience) as predicted by CEPT1. Factor analysis is used to "describe the underlying structure that explains a set of variables" (Mertler and Reinhart, 2005, p 249). This type of analysis provides an opportunity to explore the extent of shared variance among the items included on the SCP survey. In other words, a factor analysis can help assess the extent to which the various survey questions measure the same or different underlying constructs.

3.3 Evidence to Inform Educational Toolkits

The results of the analysis of the pilot test results will be used to develop recommendations and toolkits for implementation of the SCP instrument. One of the overarching aims of the analysis was to use the findings to inform toolkits for users of the new student course perception tool (instructors, academic administrators, and students). The toolkits are designed to inform users about how to interpret and understand results from the new student course perception survey.

4 Teaching Effectiveness at the University of Waterloo

Student ratings of instruction (SRIs) reflect students' perceptions of how well they have been taught and how much they were engaged with the course material. As Hativa (2014) explains, "SRIs correlate with student perceptions of effective instruction and with the conceptual structure/main behaviors of effective instruction" (p. 46).

In designing any course evaluation survey, careful consideration must be given to the institutional context in which the instrument will be used (Ory and Ryan, 2001; Gravestock and Gregor-Greenleaf, 2008; Theall and Franklin, 2001; Willits and Brennan, 2017). The validity and utility of any evaluation tool depends heavily on the development of questions that reflect institutional goals and practices of teaching (Gravestock and Gregor-Greenleaf, 2008; Ory and Ryan, 2001; Theall and Franklin, 2001).

Discussions about effective teaching practices and priorities at Waterloo are not new, as evidenced in several institutional reports released in recent years. For instance, in 2011 the Deep Learning Report highlighted that, at Waterloo, effective teaching is best understood as that which facilitates deep learning. The report draws on Chickering and Gamson (1987)'s seminal work, which describes the key characteristics needed for good teaching and learning, and Bain (2004)'s longitudinal study, to define effective teaching practices and priorities. Drawing on this holistic working definition of effective teaching, CEPT1 designed the student course perception survey to include items that measure the following three dimensions: Course Design, Course Delivery, and Learning Experience.

Most recently, five evidence-based principles of effective teaching, which stem from the research literature, were proposed in the University of Waterloo's Undergraduate Learning White Paper Report[9]. The principles enumerated in the report are that, at Waterloo, effective teaching:

1. Uses alignment in design principles.
2. Fosters motivation.
3. Embodies inclusivity.
4. Encourages deep learning.
5. Enables lifelong learning.

The report further notes that "These five fundamental principles need to exist in an environment where both the institution's instructors and senior administrators demonstrate a commitment to effective teaching."

1. **Alignment** in design occurs when outcomes that are focused on learning are made explicit for learners in courses and programs, the assessments of learning match the outcomes, and the incorporated activities prepare learners for the assessments (Biggs & Tang, 2007).

2. **Motivation** occurs when learning experiences, inside and outside the classroom, are relevant and of value to learners, provide them with choice, and feel achievable yet appropriately challenging (Svinicki, 2004).

3. **Inclusivity** occurs when learning environments and experiences engage learners with differences respectfully and are designed to enable all to learn (Ouellett, 2005).

4. **Deep Learning** results from experiences that encourage learners to make connections, apply knowledge in new contexts, engage in learning activities and analytical thinking on their own and with others, and retain their learning (Christensen Hughes and Mighty, 2010).

5. **Lifelong Learning** occurs from experiences that teach students to think about their thinking, become self-aware as learners, take responsibility for their learning, and self-assess their learning (McGuire, 2015).

The new student course perception survey is intended as a tool to help measure these teaching principles and priorities by measuring student perceptions of how well they are achieved. The items included on the new student course perception survey align with the three overarching dimensions underlying the student learning experience (i.e., course design, course delivery and learning experience), as well as four of the five principles outlined above (alignment, motivation, inclusivity, deep learning). The list of items developed by CEPT1 was intended to cluster into three groups, each group corresponding to a dimension as listed above[10]. In its preliminary focus group investigations, evidence was discovered that in general the items in the proposed instrument were interpreted by students in ways that matched those anticipated by CEPT1 on the basis of their research.

There were a couple notable differences between students understanding of some items and the anticipated interpretation of those items. For example, one question CEPT1 included as part of the course delivery

[9] The report is available to members of the UWaterloo community with their WatIAm credentials at this site:https://uwaterloo.ca/strategic-plan/bridge-to-2020/issue-papers/undergraduate-learning. While this report has not been formally endorsed (e.g., by Senate), it is worth highlighting that it was produced by a working group that included faculty, administrators, staff and students, and was made widely available to the campus community as part of the consultation process for the 2020-2025 Stratetic Plan. It echoes the results of other consultations about effective teaching carried out at Waterloo.

[10] For discussion and a summary table, see https://uwaterloo.ca/associate-vice-president-academic/sites/ca.associate-vice-president-academic/files/uploads/files/background_report-_compare_sample_questions_with_current_faculty_questions.pdf

set asked about workload demands in the course. However, focus groups with students across the six Faculties in Spring 2018 revealed that this question did not seem to be regarded by students as an indicator of quality of teaching. In general, for the purposes of the pilot test, this question has been used instead as a control variable, to assess the extent to which this item might be associated with scores on the other SCP items. Moreover, the pilot test gave some reason for concern about the two "overall" items ("Overall, I learned a great deal from this instructor", and "Overall, the quality of my learning experience in this course was excellent"). The focus groups initially revealed that these global items failed to connect with the themes identified as ones that determined whether the students regarded the instruction for the course as worthy of a high or low evaluation. It is notable that these overall items are the only two Likert Scale questions that were intended by CEPT1 to measure learning experience (along with two open-ended questions)[11].

The items in the pilot tested instrument were essentially the Likert scale questions recommended by CEPT1, with some minor wording adjustments to a couple of items in response to the results of the focus group research carried out in Spring Term 2018.

5 Uncertainty Quantification

5.1 Sampling Error and the Population of Interest

For the analysis, CEPT2 worked with the Statistical Consulting and Research Collaboration unit in the Mathematics Faculty. The preference of our collaborating partners from this unit was to formulate the research questions in our analysis in ways that do not implicitly depend on some or other statistical model. The starting point for the analysis, then, is that the ideal data set for the pilot test would be one that included all students at UW and its Affiliated Institutions completing every question on the SCP survey for every course in which they were enrolled during the Fall 2018 term. The research questions were then formulated in ways that, had we had this ideal data set, could have been answered precisely, i.e., with no statistical uncertainty.

Working in this way, the statistical uncertainty in our results has one source, namely that we are working with only a sample of the ideal data set. In the report, standard errors are calculated as though the sample is randomly selected. Of course, in the present case, the sample is not randomly selected. We must therefore be cautious about conclusions made given that we have no way to know how well the survey results represent or reflect the entire student population. This limitation, of course, also holds true when we consider the sample of students who complete regularly scheduled course evaluations in any given term. Results from the CEPT2 focus groups with students at UW, and a large body of research, have shown that certain factors (e.g., high student engagement, a highly positive or negative course experience, the perception that the instructor cares and will use the feedback, a sense of responsibility to help future students, etc.) all increase the likelihood that students will complete a course evaluation in any given course (Adams and Umbach, 2012; Chapman and Joines, 2017; Crews and Curtis, 2011; Lewis, 2001; McGowan and Osguthorpe, 2011; Tucker et al., 2008).

In total, 41,737 surveys were submitted by students from across the six Faculties and two of the AFIW. This participation resulted in a relatively high response rate (58%) when compared to the total number of submissions received for the regular end-of-the-term evaluations (72,652) for the Fall 2018 term. This

[11] A summary of these issues can be found here: https://uwaterloo.ca/waterloo-course-evaluations/focus-group-analysis-summary-key-themes#SummaryFindings2-3.

participation is quite substantial considering that this was only a pilot test and thus it was extra work for students to complete it. Our official response rates by course are given in Figure 9. We see response rates are between 28% and 39% for on-campus courses, and somewhat lower for online courses. This response rate is in line with those found at the University of Toronto on their Course Evaluation Framework (response rates range from 32-50% depending on class size) and the response rates found for other online course evaluation frameworks (Goos and Salomons, 2017), as well as surveys of student engagement (NSSE, 2016), and online survey research in general (Cook et al., 2000; Shih and Fan, 2008, 2009)[12].

We can have some confidence that our pilot sample is reflective of the sample of students who completed the regular end-of-the-term evaluations in Fall 2018 (and so, presumably, reflects a typical sample for other terms) for a couple of reasons. First, as highlighted above, the response rate for the pilot test is relatively high. Second, the pilot test was run parallel to and, in conjunction with, the regular end-of-the-term course evaluations[13].

5.2 Margin of Error

According to James et al. (2015), the interpretive validity of a measure is the extent to which our intrepretations of a test score actually reflect the reality (or 'truth') of that score in a specific assessment situation (see also Osterlind (2010)). With respect to student evaluations of teaching, it is important to consider three potential threats to the interpretive validity of our score: 1) the class size; 2) the response rate; and 3) the sample variability (or standard deviation score). The interpretive validity of our score is threatened when: 1) a class size is small (especially for classes of less than 25 students); 2) response rates are low; and 3) sample variability is high James et al. (2015). When these three conditions are present, the quality of the evaluation results may be undermined because the margin of error will be high. The margin of error tells us the amount of sampling error in our test results. A large margin of error undermines our confidence that the test results (drawn from a sample of the class; so those who actually respond to the SCP survey) are a true reflection of the population (the entire class).

5.3 Other Possible Modelling Choices

An important limitation to proceeding without modelling assumptions is that the "population of interest" for the pilot study consists only of the students and instructors in Fall 2018. Without further assumptions, we cannot make statistical statements about correlations between the SCP items and explanatory variables in future terms with different students and instructors. There are, of course, other choices that might be made for the formulation of the research questions. Decisions about which are appropriate depend on subtle questions about precisely what theoretical construct one takes the instrument to be measuring. It is, of course, an interesting question whether other legitimate choices might have significantly affected the results. Decisions about where to invest the limited resources available for this project within the Statistical Consulting and Collaborative Research Unit have meant that we could not pursue those questions in detail.

That said, the Unit did do some preliminary work on an approach that would have, among other things, permitted statistical statements to be made about a different cohort of students and instructors having

[12] Centre for Teaching Support & Innovation, (2018), *University of Toronto's Cascaded Course Evaluation Framework: Validation Study of the Institutional Composite Mean (ICM)*. Toronto, ON, University of Toronto.

[13] To be clear, the surveys "spoke" to one another. When students completed the regular evaluation they were encouraged to complete the pilot test, and vice versa if they chose to complete the pilot test first. This was the case for every course students were enrolled in.

similar characteristics to those of Fall 2018. While completion of this work and a full accounting of it must wait for a future opportunity, enough has been done to say with some confidence that this approach does not substantially change the results reported below: only error bars would change (to become larger), but generally the difference would be fairly small.

6 Levels of Analysis

Three levels of analysis are considered here:

1. **Student Level.** Each unit at the student level consists of a student/course pairing. This means that each student counts for as many courses as they are enrolled in.

2. **Course Level.** Each unit at the course level consists of an instructor/course pair (an instructor is counted for each course they taught; for example, if an instructor taught three courses they would be counted for three scores, and each section of a multi-section course results in an instructor/course pair). A useful way to understand the difference between Student-Level and Course Level analysis is illustrated in the following example: If there were only two courses taught at UW in Fall 2018, one with 10 students and one with 1000, then the Student-Level average on a given response item gives each of the 1010 student-course pairs an equal weight, whereas the Course-Level average gives each course equal contribution, i.e., each student in the smaller class contributes 100x times more than each student in the larger class.

3. **Instructor-Level.** Each unit corresponds to one instructor. This is the level at which we examined such things as the instructor gender distribution by appointment type (e.g., tenured, sessional, etc.), but it is the wrong level for most purposes in this analysis (for instance, to examine the instructor gender distribution by class size, as an instructor can teach multiple classes, which makes Course-Level more appropriate). The level of analysis adopted was informed by logistical and practical decisions about how to most appropriately answer the research question posed.

7 Demographic Statistics

7.1 Faculties

Figure 1 displays the breakdown of SCP survey responses by Faculty, at both the student-level and course-level (course-level responses are the numbers in brackets). We can see that the Faculty of Engineering had the highest number of students participate in the pilot test (9,997 in 337 courses), followed by Mathematics (9,265/360) and Arts (8,183/644).

The Graduate_Studies course corresponds to a single course offered by all Faculties, and was removed from the analysis when Faculty was in play as a variable. The Engineering/Interdisciplinary and Environment/Interdisciplinary courses were merged with Engineering and Environment, respectively, and the Science/Engineering courses with Science.

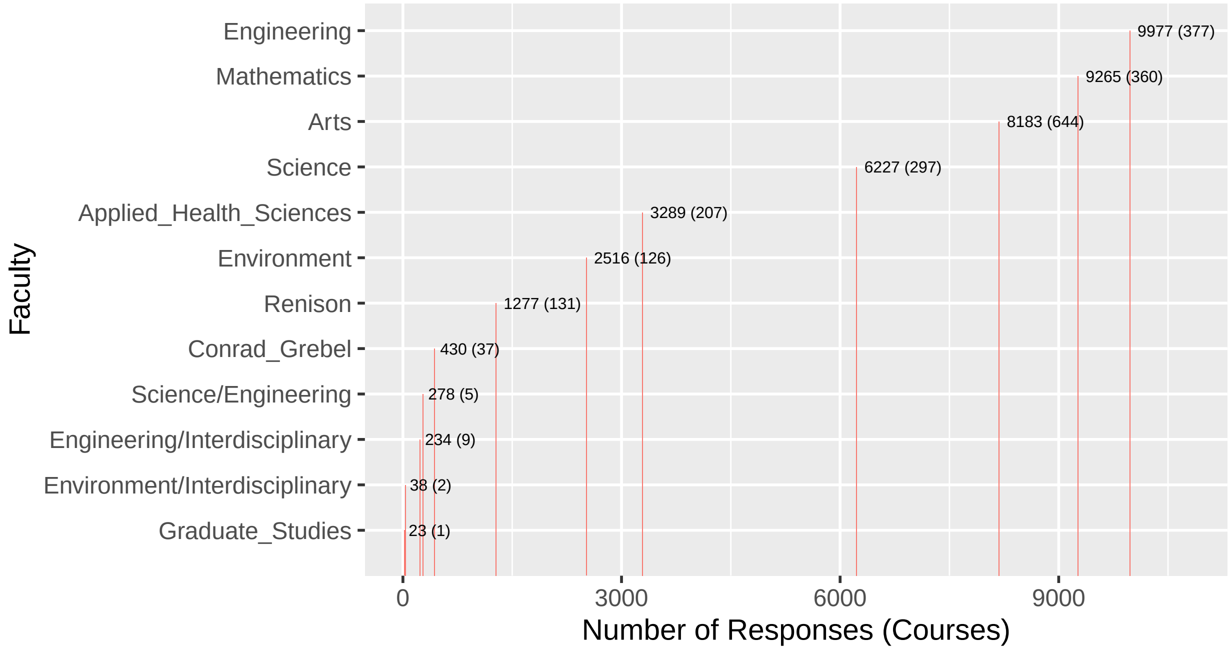


Figure 1: Number of students and courses (in brackets) by Faculty.

7.2 Student Gender

Figure 2 displays students' self-reported gender across all surveys. We can see there is a relatively equal distribution of males and females in the sample, with 48% of the sample identifying as female and 45% identifying as male.

Due to the anonymity of course evaluations, we do not have students' unique ID numbers. Therefore, there is no way to determine how many times each student is counted in the data. As is the case with regular end-of-the-term course evaluations, students are asked to complete a survey for each course they are enrolled in. Students who completed the pilot test four times (in four different courses) would therefore be counted four times in the pilot test results.

In order to protect the confidentiality of the small number of respondents who selected gender categories other than male and female, for the remainder of this analysis we combined the following categories into one category we have labelled `Identified_Other`: `Non-binary`; `Agender`; `Not_Listed`; `Genderqueer`; `Transgender_Male`; `Transgender_Female`. The number of responses for the `Identified_Other` category was rather small and thus if included in some parts of the analysis would have produced results with large standard errors, and would also potentially undermine confidentiality of survey participants. Therefore, when considering gender as a variable in parts of the analysis, we used binary coding (male/female) and excluded the responses for those who identified otherwise or preferred not to respond. Those excluded responses were included in the analysis when the risk of compromising the anonymity of participants was minimal, e.g., in the regression analyses.

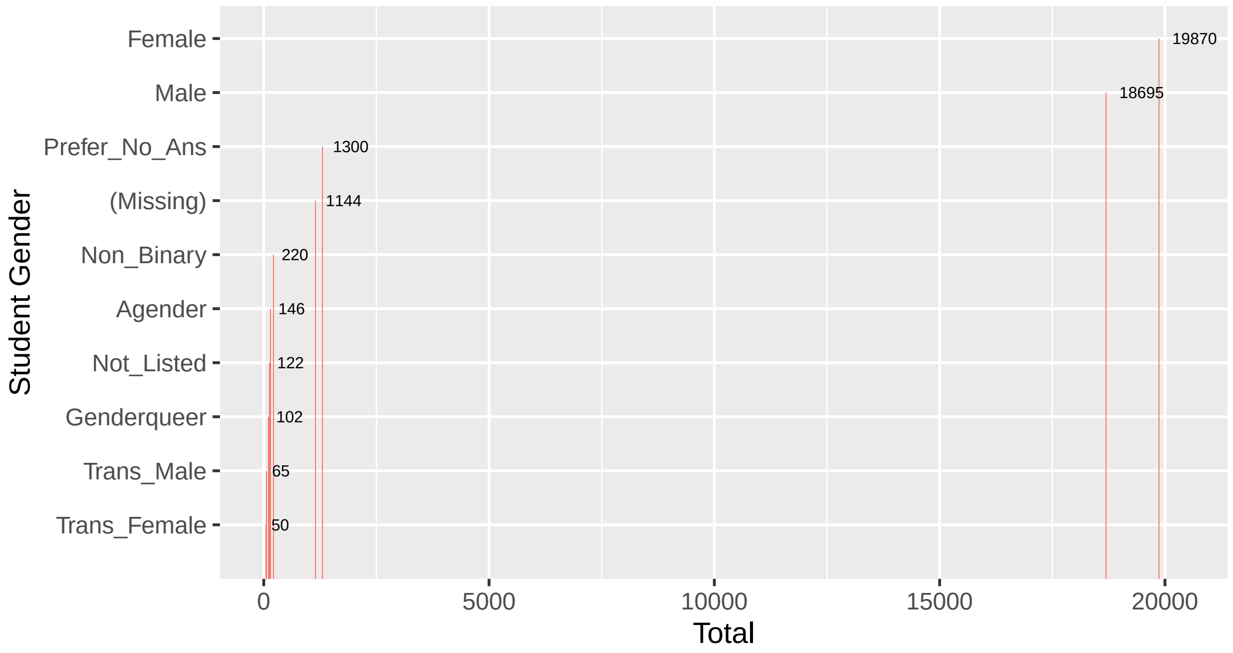


Figure 2: Number of students by gender.

7.3 Instructor Gender

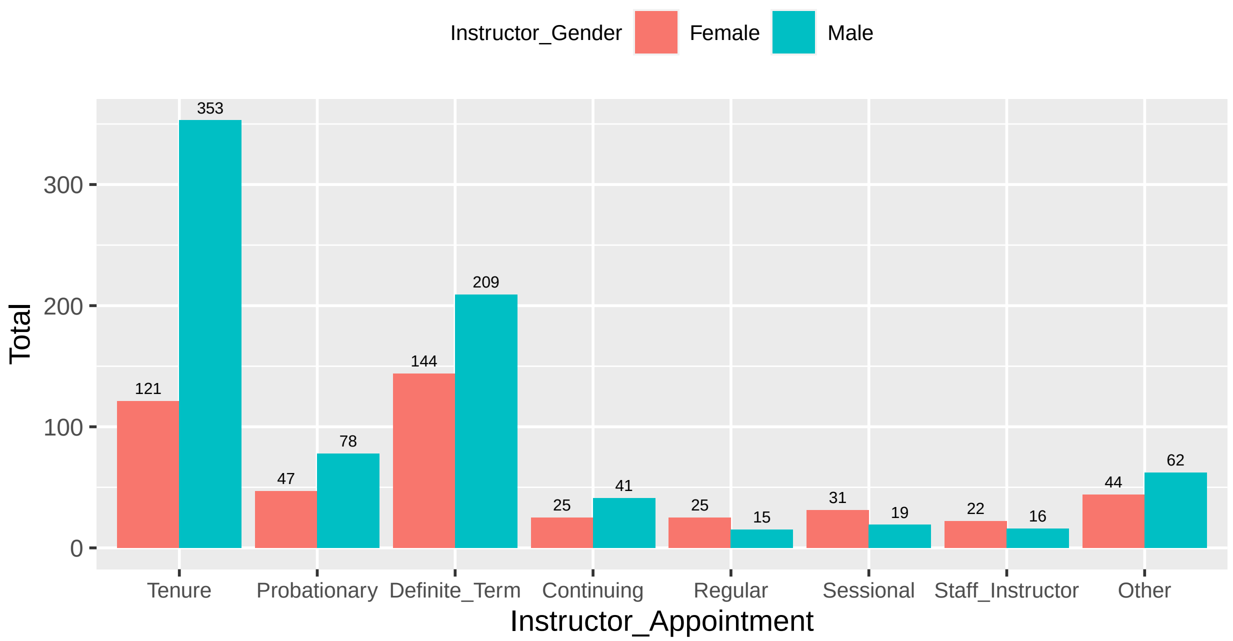


Figure 3: Number of instructors by gender and appointment type.

Figure 3 displays the breakdown of instructor gender (according to HR records) by appointment type in the courses included in the pilot test. Male instructors outnumber female instructors in most appointment types, with the most pronounced difference for tenured instructors, where we see more than twice the number of male instructors (n = 353) with `Tenure` status compared to female instructors with that status (n = 121). Male instructors also outnumber female instructors in the `Probationary` and `Definite_Term` categories. Female instructors are more likely than male instructors to show up in the the following appointment categories: `Regular` (25), `Sessional` (31) and `Staff` (22). The `Regular_Faculty` presumably includes tenured, probationary, continuing and contract faculty members, but they were not further specified in the HR data obtained. The `Other` category includes a wide variety of different appointment types, including some postdocs, some graduate students, and some research faculty.

Figure 4 shows the breakdown of female and male instructors for the courses involved in the pilot test by Faculty. The Faculty of Applied Health Sciences had an equal ratio of male to female instructors teaching during the Fall 2018 term. In Science, Math, and Engineering, we see nearly three times more males teaching courses. In the Faculty of Arts, we can also see that male instructors outnumber female instructors, but the difference is far less than we see in some other Faculties. At Renison, we see nearly twice as many females as males teaching courses (or, more precisely, twice as many female- as male-taught courses).

Figure 5 shows the fraction (percent) of instructors by gender teaching different size classes. For the fall term, female instructors were slightly more likely to teach courses with 1-50 students (64% compared to 54% for males), whereas male instructors were more likely to teach larger classes with 51-200 students (46% versus 36% for females). It is worth noting that only about 11% of the sample of female instructors taught courses with 101-200 students (compared to 18% for males) and a mere 3% of both female and instructors taught courses with 200+ students (which of course resulted in a greater number of very large courses with male instructors due to the larger number of male instructors).

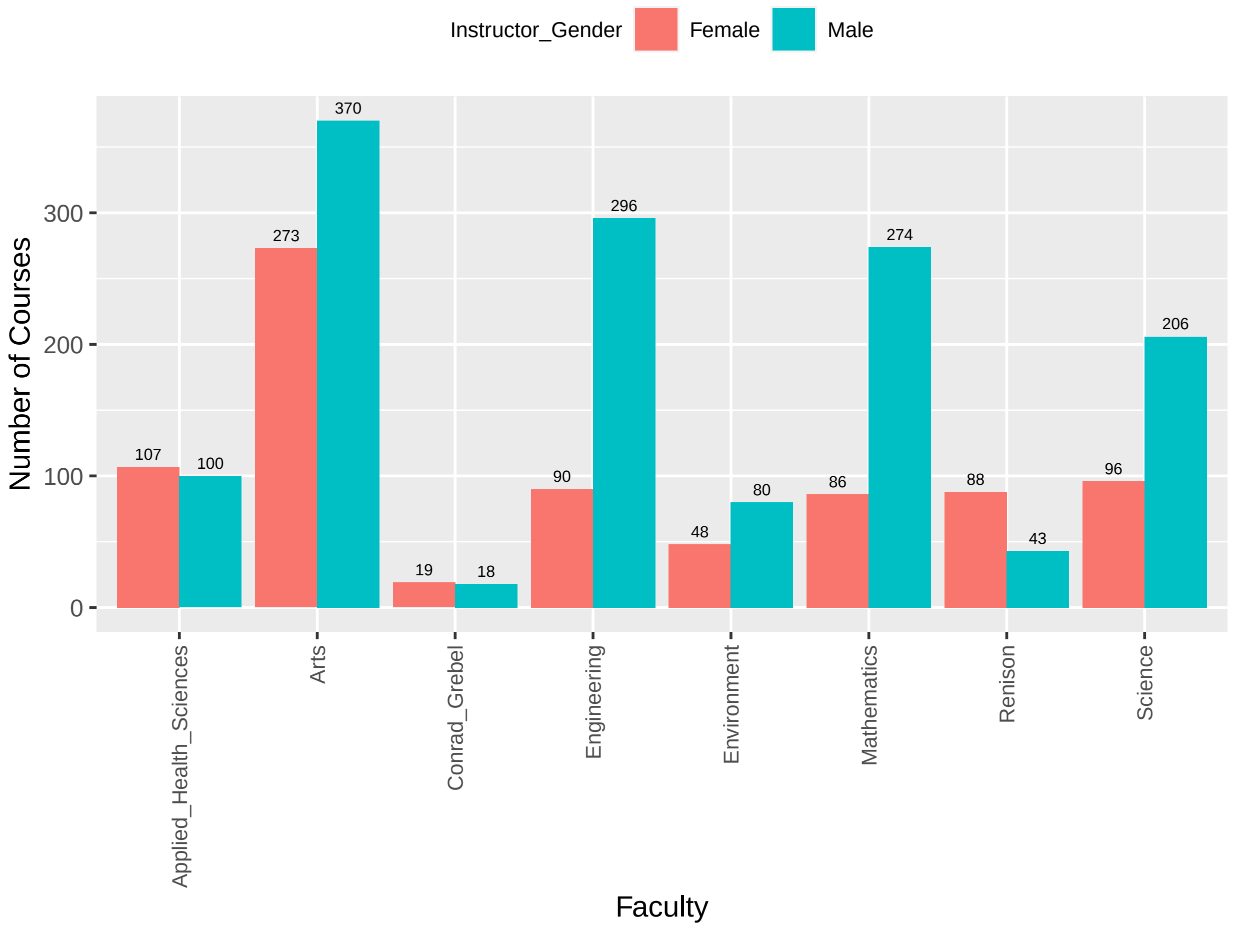


Figure 4: Number of courses taught by gender and faculty.

Figure 6 displays students' expected grade (self-reported) by instructor gender. The graph shows that among survey respondents there is virtually no difference in students' grade expectations regardless of whether the instructor is male or female. We can also see that most students, about three-quarters, of those surveyed expect to receive a grade between 70% and 90%.

Figure 7 shows students' self-reported class attendance by instructor gender. Similarly to the previous figure, it is clear that there are minimal differences in self-reported attendance rates, regardless of the instructors' gender. It is important to note that well over three quarters (83%) of those surveyed indicated that they `Almost_Always` attend class (for both male and female instructors). There has been some concern raised in the literature that conducting course evaluations online increases the likelihood that students who do not attend class will complete course evaluations. However, for the pilot survey, for both male and female instructors, considerably less than 10% of the respondents reported attending class `Almost_Never`, `Less_Half`, or `Half` of the time.

Figure 8 shows students' self-reported rating of the course workload by instructor gender. The bar graph shows that students perceive the workload to be the same for both male and female instructors, with over half of the sample rating the course workload as `Average` and about a quarter of the sample rating the course workload as `High` for both male and female instructors. On the other hand, less than 10% of those surveyed (for male or female instructors) felt the workload in any given course was `Very_Low` or `Low`.

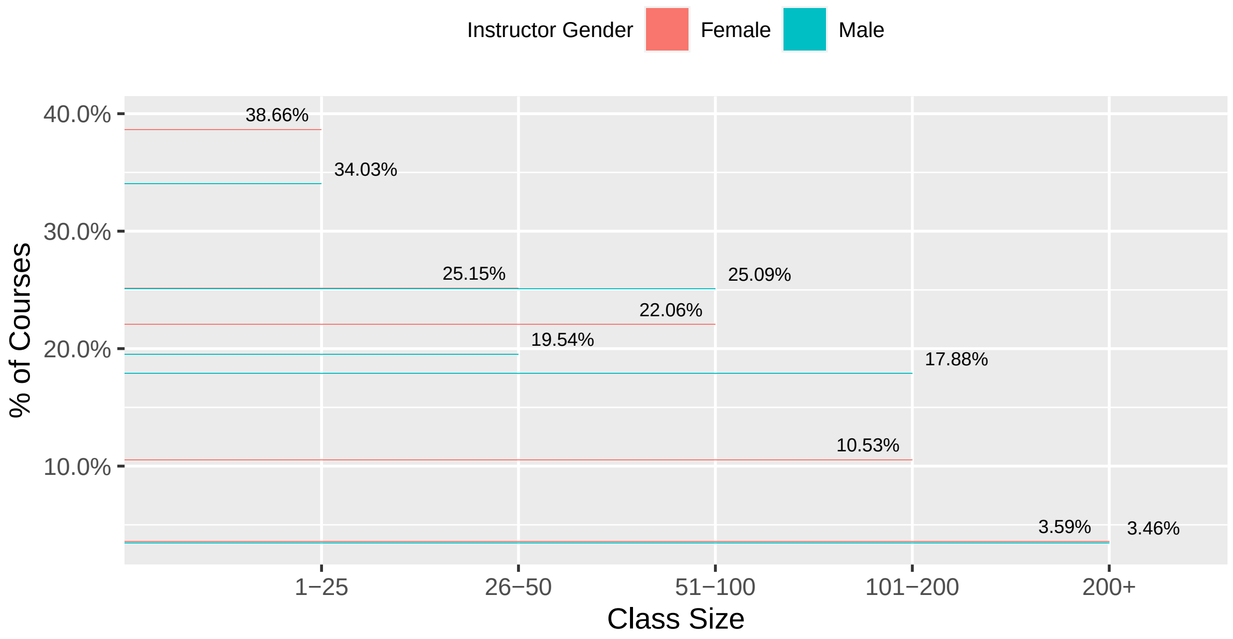


Figure 5: Fraction of courses by Class Size within each Instructor Gender.

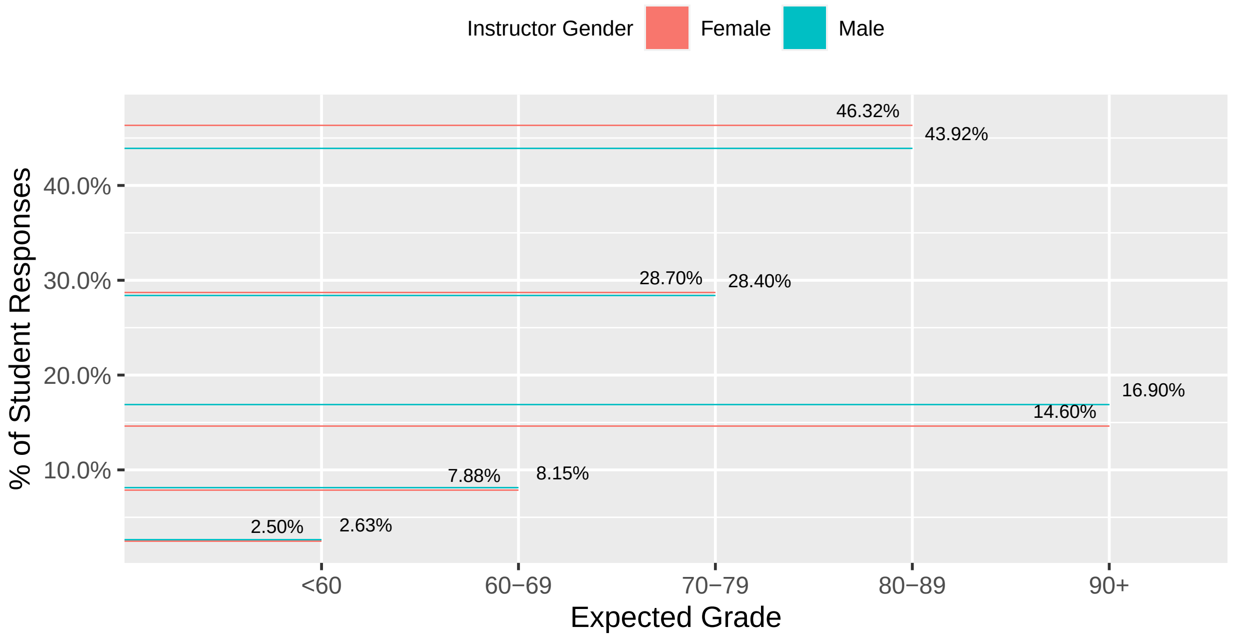


Figure 6: Proportion of students' Expected Grade response by Instructor Gender.

Figure 9 displays the percentage of students enrolled in a course who completed a pilot test evaluation in each course by class size and attendance type (i.e., in-class/online). (As noted in the introductory sections of this report, the response rates for the SCP were about 60% of the rates for official surveys, but we have no information for how the response rates for the two types of surveys compare by class size.)

It is clear that SCP evaluations are completed by a higher percentage of students taking in-class courses than those taking online courses, across all class sizes. It seems that this difference in response rates

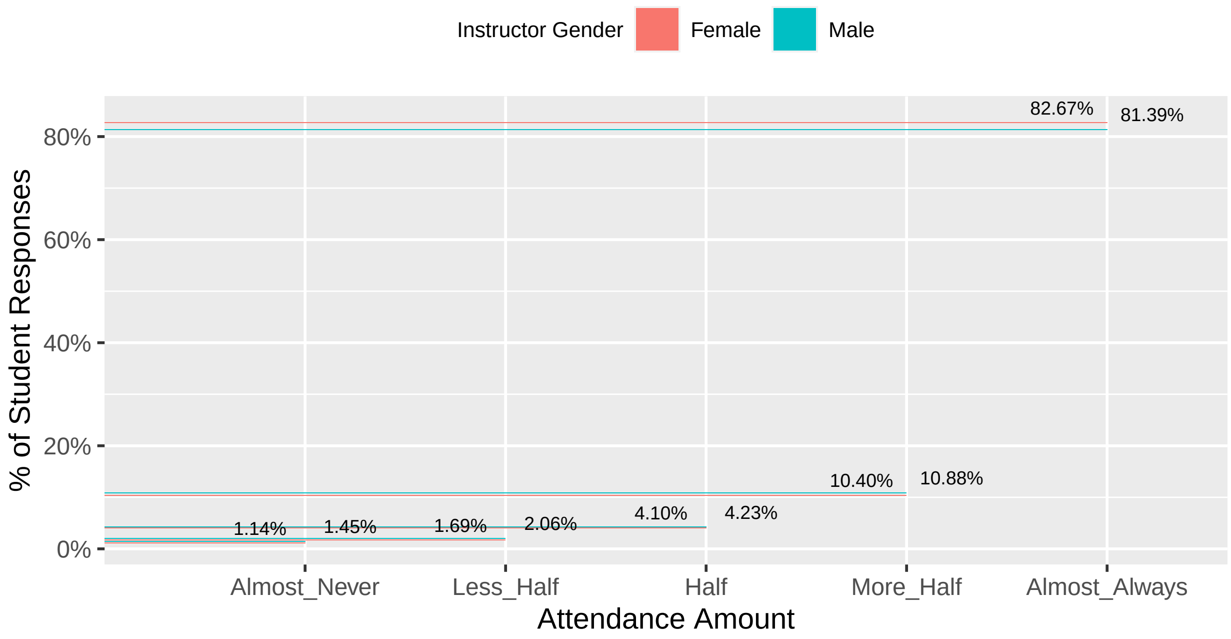


Figure 7: Proportion of students' Attendance Amount response by Instructor Gender.

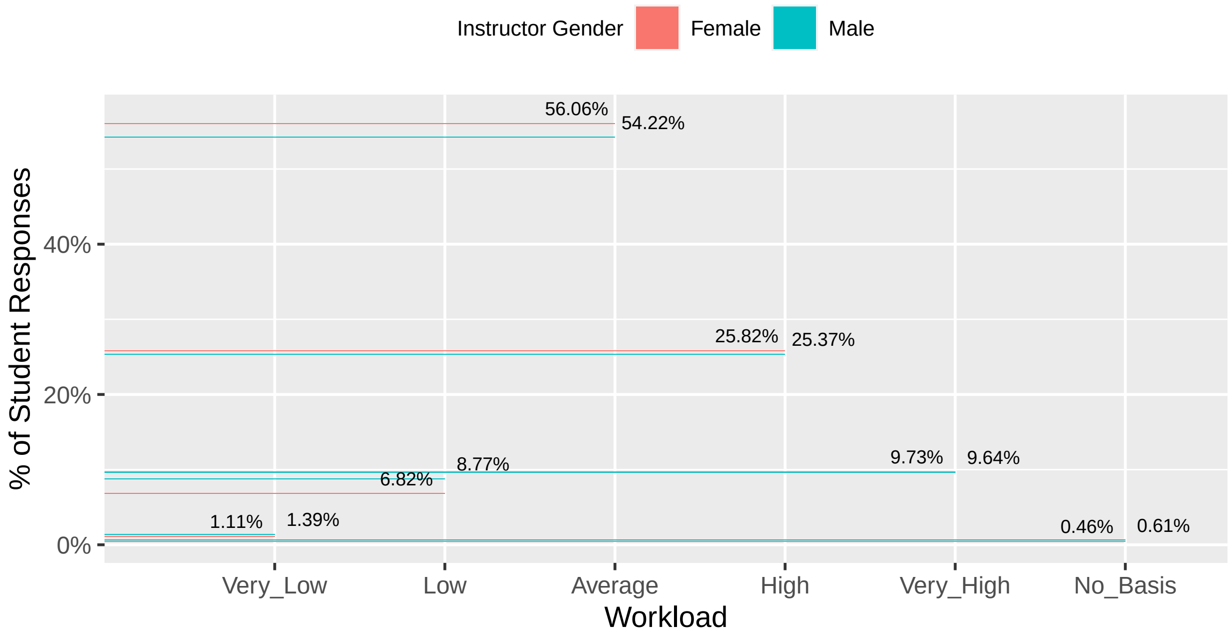


Figure 8: Proportion of students' Workload response by Instructor Gender.

increases as class size gets larger, particularly for classes with 100+ students. For example, in courses with 101-200 students, 34% of the students who took an in-class course completed the SCP pilot test, compared to only 15% of students who took a course that size online.

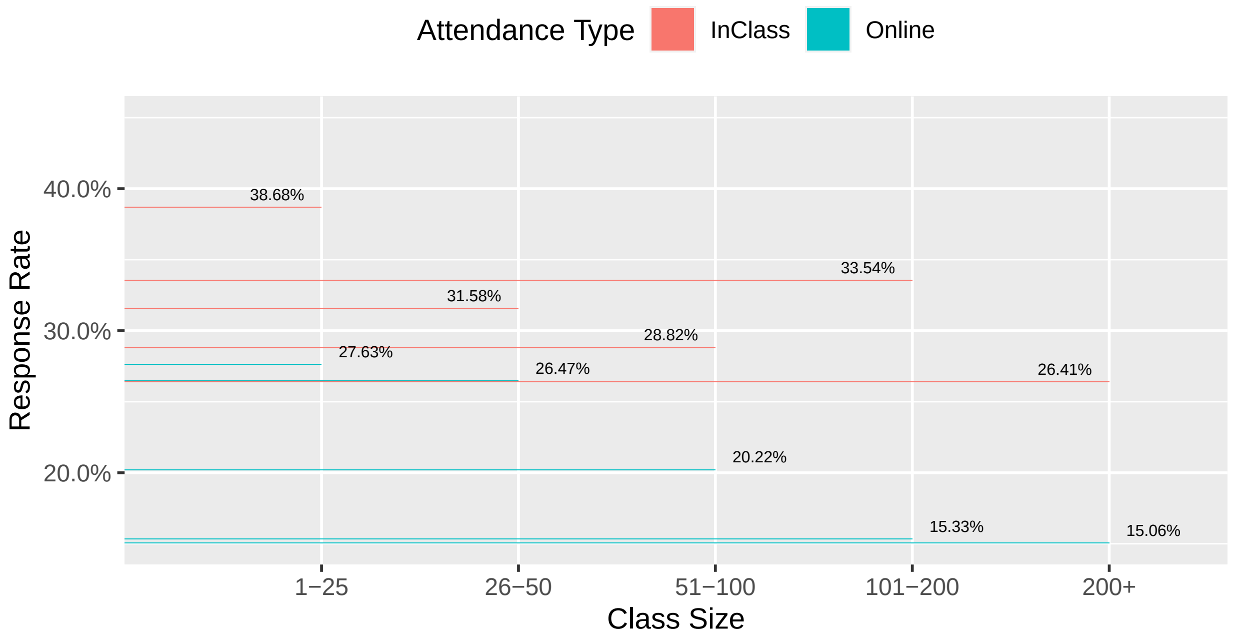


Figure 9: Survey response rate by class size and attendance type.

7.4 Non-Response and No Basis

For the SCP pilot test, the CEPT2 made a strategic decision to include a new category (in addition to the 1-5 Likert scale categories) for each core survey item. This new category was labelled: 'have no basis for rating' on the course questionnaires. The decision to include this category was informed by both the literature and our findings from the focus groups with students at UW. In the absence of such a `No_Basis` option, the '3' on a five-point Likert scale is often used as a catchall response for perceptions such as 'don't know,' 'not applicable.' 'no opinion,' etc. Students may also elect to leave a question blank if they do not feel they can assess it, which creates missing data problems. We wanted to explore the use of this category as it would give us some indication as to whether or not the questions were in fact measuring things that resonated with students' learning experiences. Specifically, if the 'have no basis for rating' category was selected with high frequency it could indicate that the core questions are measuring things that fail to resonate with the student learning experience. The same logic applies when missing data is high.

Figure 10 displays the proportion of students who selected the new category `No_Basis` and the proportion of missing data by class size. Overall, less than 10% of the respondents selected 'have no basis for rating' even once, regardless of the course size. In classes of 200+ this option was selected one or more times by only 7% of students. The proportion of missing data is also quite low, with less than 6% of cases with any missing data across all course sizes.

```
Note: Using an external vector in selections is ambiguous.
i Use `all_of(resp_names)` instead of `resp_names` to silence this message.
i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
This message is displayed once per session.
```

Table 1 displays the frequency (%) distribution of all nine questions in the pilot test. It is clear that `Agree` and `Strongly_Agree` are the most commonly selected categories on the 5-point scale with between 75% and 85% of responses for most items. There has been some concern in the literature that students who are

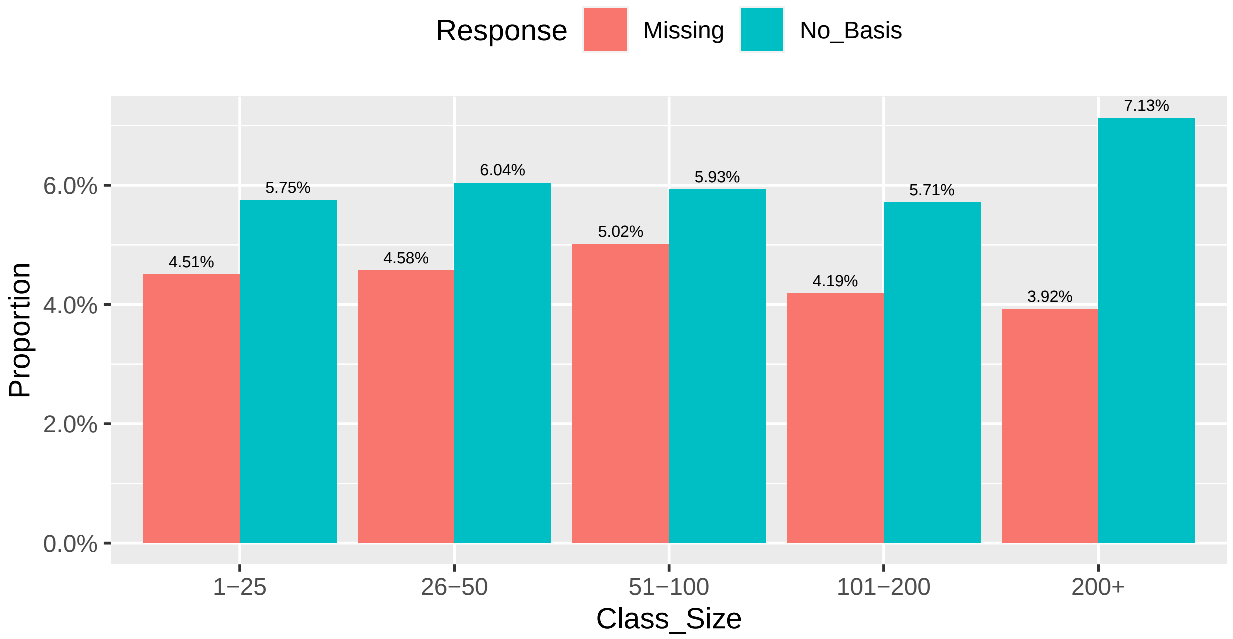


Figure 10: Fraction of student responses with one or more "Missing" or "No-Basis" response by class size.

Table 1: Proportion of responses per SCP item (%).

	Identified_LO	LO_Assessed	Course_Activities	Return_Grades	Concepts_Conveyed	Learning_Environment	Stimulated_Interest	Amount_Learned	Learning_Experience
Strongly_Disagree	2.0	1.8	3.2	3.0	3.1	2.5	5.08	3.65	3.97
Disagree	3.7	4.4	8.5	6.6	5.5	4.1	8.79	6.57	8.59
Neither	8.4	10.3	11.5	9.2	10.1	11.9	16.57	12.70	15.50
Agree	43.1	47.4	41.3	42.0	41.5	38.1	32.87	37.50	37.90
Strongly_Agree	39.8	32.3	32.3	36.1	37.7	41.0	35.01	38.03	32.80
No_Basis	1.4	2.1	1.9	1.9	1.1	1.3	0.77	0.65	0.41
(Missing)	1.5	1.7	1.2	1.2	1.0	1.1	0.91	0.91	0.85
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.00	100.00	100.00

“out to get” instructors are more likely to complete course evaluations (e.g., as retaliation for poor grades). However, as was revealed through our focus groups, though students report that they are more likely to submit a course evaluation when they have a strong opinion (positive or negative) about the course or instructor, this does not result in high percentages of low scores. This situation is also evidenced in the literature and is why we tend to see skewed distributions (with scores clustering at the top end of the scale), with students assigning scores of four and five with the most frequency.

8 Differences by Instructor Gender

The Figures and Tables in this section display summaries of SCP scores by the nine core items and the key explanatory variables in this study: faculty, instructor gender, student gender, class size, expected grade, and perceived workload. Numerical averages for the nine core items were calculated by coding the five ordinal response categories (ranging from "strongly disagree" to "strongly agree") from 1 to 5.

Whenever possible, numerical averages were calculated at the course level (i.e., each course average contributes equally to the overall average). The exception to this rule was when the explanatory variable is at the student level (e.g., expected grade).

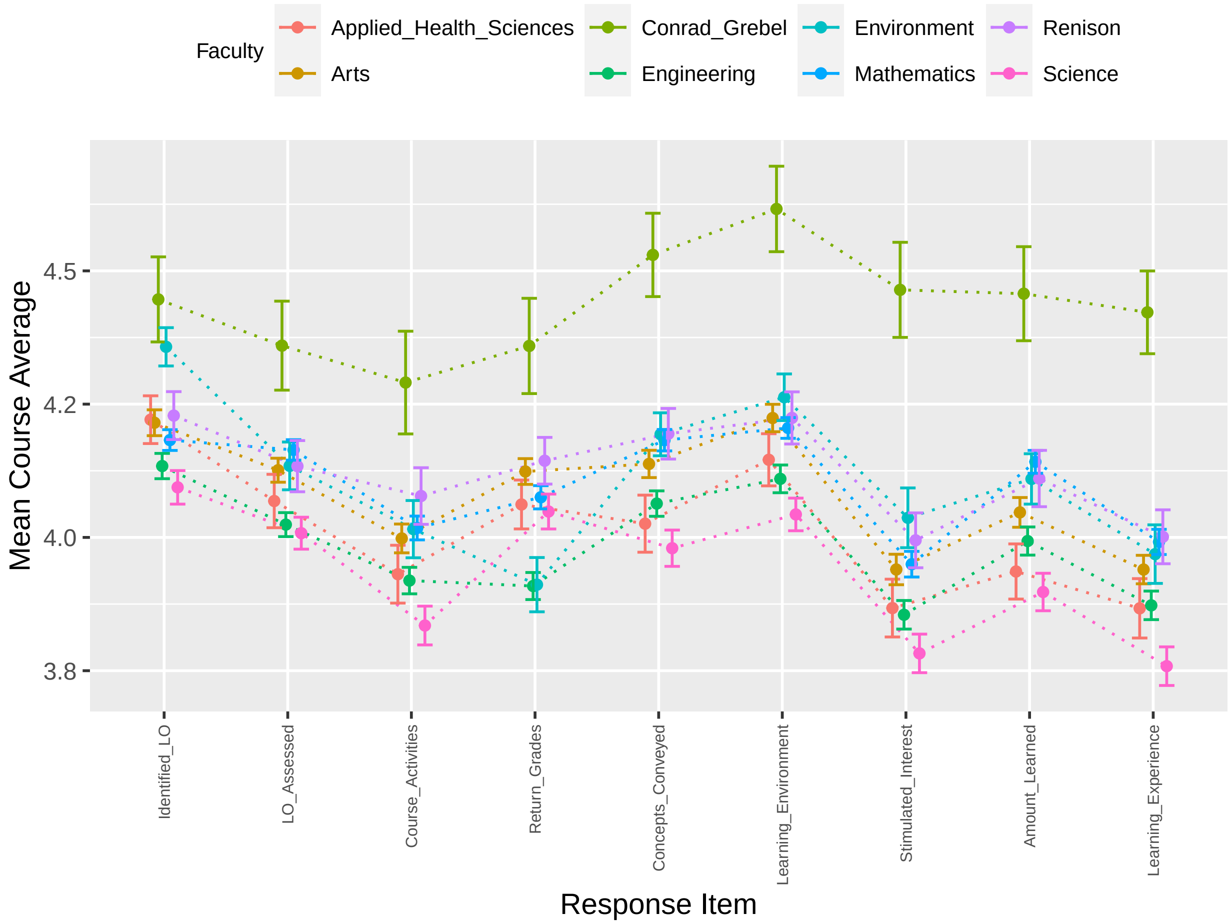


Figure 11: Mean course average per Response Item by Faculty.

Figure 11 displays the mean course average across all response items for each Faculty. Also displayed are error bars, i.e., confidence intervals of ± 2 standard errors. These error bars reflect the potential change in the reported values which might occur if a different random sample of Fall 2018 students in each class had filled out the SCP surveys instead.

It is clear that most average scores fall somewhere between 3.8 and 4.2. Conrad Grebel has the highest average scores by item, between 4.2 and 4.6. Renison, Math, Environment and Arts follow Conrad Grebel with scores near 4.2 across most SCP items. Science has scores clustering between 3.8 and 4.0. Overall, there seems to be at least some consistency in how the nine items are rated (i.e., they seem to follow the same 'up/down' trend). Three items, `Identified_LO`, `Concepts_Conveyed`, and `Learning_Environment` have higher average scores (4.2 or more; this is evidenced by the peaks in the graph), while we see a dip in the overall average for Course_Activities, Stimulated_Interest and Learning_Experience (4 or less for some Faculties). It is important to note, however, that the differences in average scores per item are not large, with a difference between questions of no more than about 0.3 points on a five-point scale. For most purposes, differences in scores by Faculty of instruction probably do not matter much, as performance reviews happen within a Department and tenure decisions at the Faculty level. However, they are important to know about for instructors who teach courses in Faculties to which their home department does not belong.

The remaining Figures in this section focus on differences in average SCP scores between male and female instructors. The difference is statistically significant at the 95% level when the error bars do not cover zero. It is however, important to note that given that we are working with a rather large sample size, most differences will be statistically significant. Therefore, it is also important to attend to the magnitude of the differences in scores.

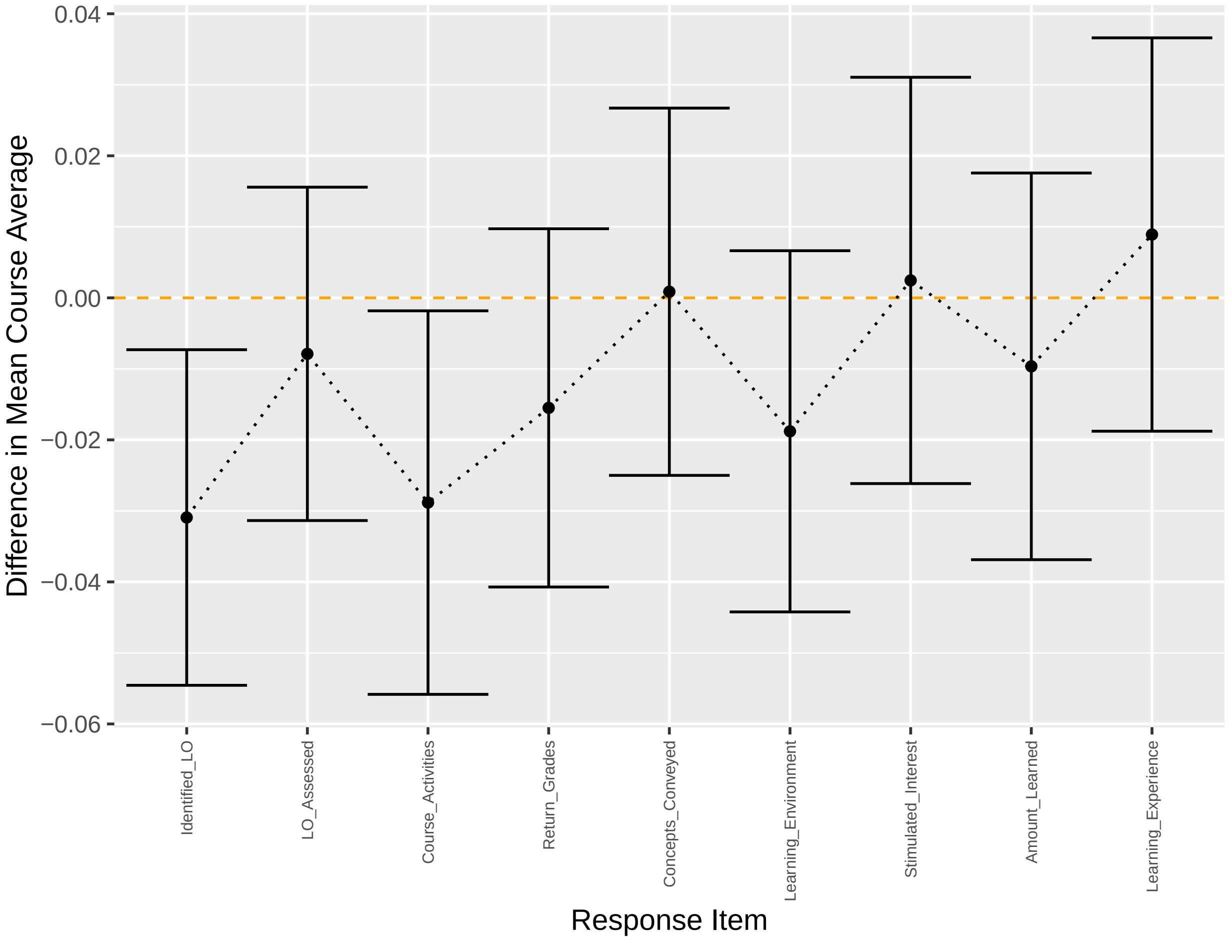


Figure 12: Difference in mean course average between male and female instructors per Response Item.

Figure 12 shows the average difference in scores (at the course level) for male and female instructors, across all SCP items. The following formula was used to calculate the difference between male and female instructors:

$$\text{difference} = \text{mean(male course average)} - \text{mean(female course average)}.$$

This means that any negative value in the graph is associated with female instructors scoring higher than male instructors for that item. Conversely, a positive value signifies males scoring higher than females for a given item, while a value of 0 reflects no difference in scores for male or female instructors.

Differences are relatively small (no more than 0.03 for any SCP item), and not statistically significant except for `Identified_LO` and `Course_Activities`. Even then, it is important to hightlight that minuscule differences become statistically significant if sample sizes are large enough. Therefore, it is important not to over-interpret (e.g., the statistically significant difference of -0.03 in `Identified_LO` as conclusive evidence that female instructors are perceived as more effective than males on this metric). In fact, many Universities have judged that it is ill-advised to report scores beyond one-decimal point. Indeed, if we report scores to only one-decimal point the gender differences in Figure 12 would no longer be detectable.

To give a clearer picture of the difference between average scores for male and female instructors, **Table 2** additionally displays the average score and standard error for each instructor gender. The largest difference in scores between male and female instructors is 0.031 found for the item `Identified_LO`. On this item, female instructors score higher with an average score of 4.21 compared to an average score of 4.18 for male instructors. The second highest difference in scores is found for the `Course_Actvities` item, for which female instructors score an average of 3.99 points, compared to 3.96 for male instructors, a difference of 0.029 points on the five-point scale. In general, the largest difference in scores is quite marginal, at most 0.03 points on the five-point scale. As noted above, it is worth reiterating that it can be very misleading to attribute too much importance to differences of less than one decimal point.

Table 2: Table of average scores for male and female instructors with their differences.

Response	Male Avg.	Female Avg.	Difference	Male S.E.	Female S.E.	S.E.
Identified_LO	4.18	4.21	-0.031	0.0070	0.0095	0.012
LO_Assessed	4.09	4.10	-0.008	0.0069	0.0095	0.012
Course_Activities	3.96	3.99	-0.029	0.0081	0.0108	0.013
Return_Grades	4.05	4.06	-0.015	0.0071	0.0104	0.013
Concepts_Conveyed	4.11	4.11	0.001	0.0073	0.0107	0.013
Learning_Environment	4.17	4.19	-0.019	0.0073	0.0104	0.013
Stimulated_Interest	3.92	3.91	0.002	0.0081	0.0118	0.014
Amount_Learned	4.03	4.04	-0.010	0.0079	0.0111	0.014
Learning_Experience	3.92	3.91	0.009	0.0080	0.0113	0.014

Of course, the table just considered presents evidence about an average of averages, and so there is potential for it to mask important differences. For instance, if it is typical that instructors in small enrolment courses receive higher scores than those teaching larger courses, and if proportionally more female instructors teach low enrolment courses, important differences by instructor gender might be undetectable. Since our view was that the campus community would be very interested in questions concerning differences by instructor gender, we considered how instructor gender interacted with a number of other variables.

Figure 13 shows the average difference in scores for male and female instructors while accounting for class size. Again, we see that the absolute difference in average scores across all items is typically less than 0.1. The notable exception to this is in large classes (200+ students) in which male instructors obtain larger scores by 0.15-0.35 across all SCP items. Female instructors also appear to fare slightly worse in courses with 51-100 students (difference in score about 0.1 across most SCP items). On the other hand, in smaller

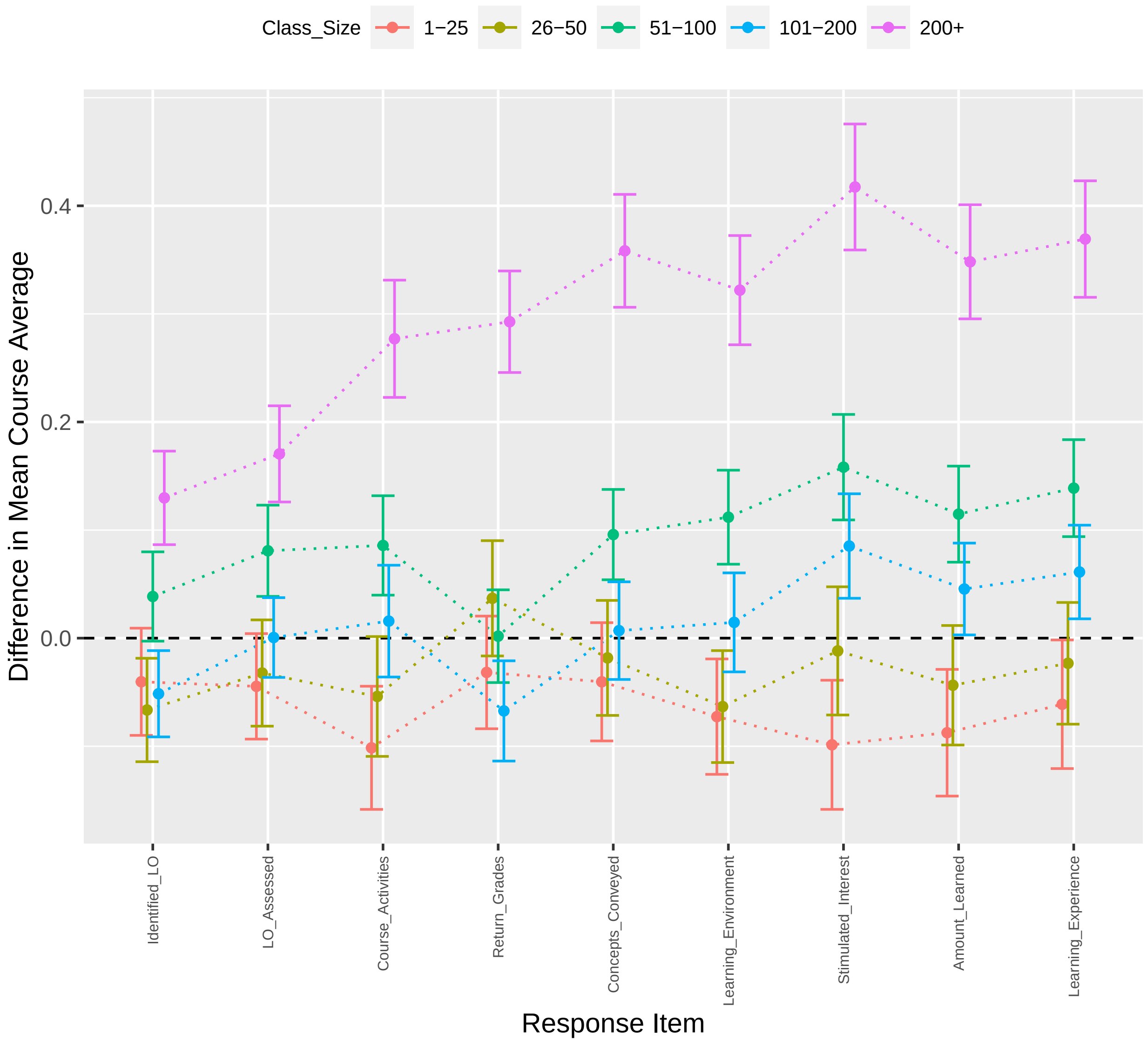


Figure 13: Difference in mean course average between male and female instructors per Response Item by Class Size.

classes with 1-25 and 26-50 students we see males actually fare slightly worse than female instructors (though by less than 0.1 points lower). We determined that the difference in scores between male and female instructors across class-sizes warranted further investigation to better understand this relationship.

The following Figures break down gender differences by class size and several other predictors: student gender, expected grade, workload, faculty, and instructor appointment type.

Figure 14 displays the average difference in scores between male and female instructors accounting for class size as well as student gender. We only report results for students identifying as Female or Male, as the number of students identifying as other gender categories is far too small to draw meaningful conclusions. Student gender does not appear to have much influence. We still see that female instructors receive lower than average scores in large class sizes (200+), but both male and female students tend to assign similar scores. For example, in courses with enrollments of 200+, for both male and female students, female instructors score 0.4 points lower on `Stimulated_Interest`, about 0.3 points lower on `Amount_Learned`, and about 0.4 points lower on `Learning_Experience`. For all other class sizes, the

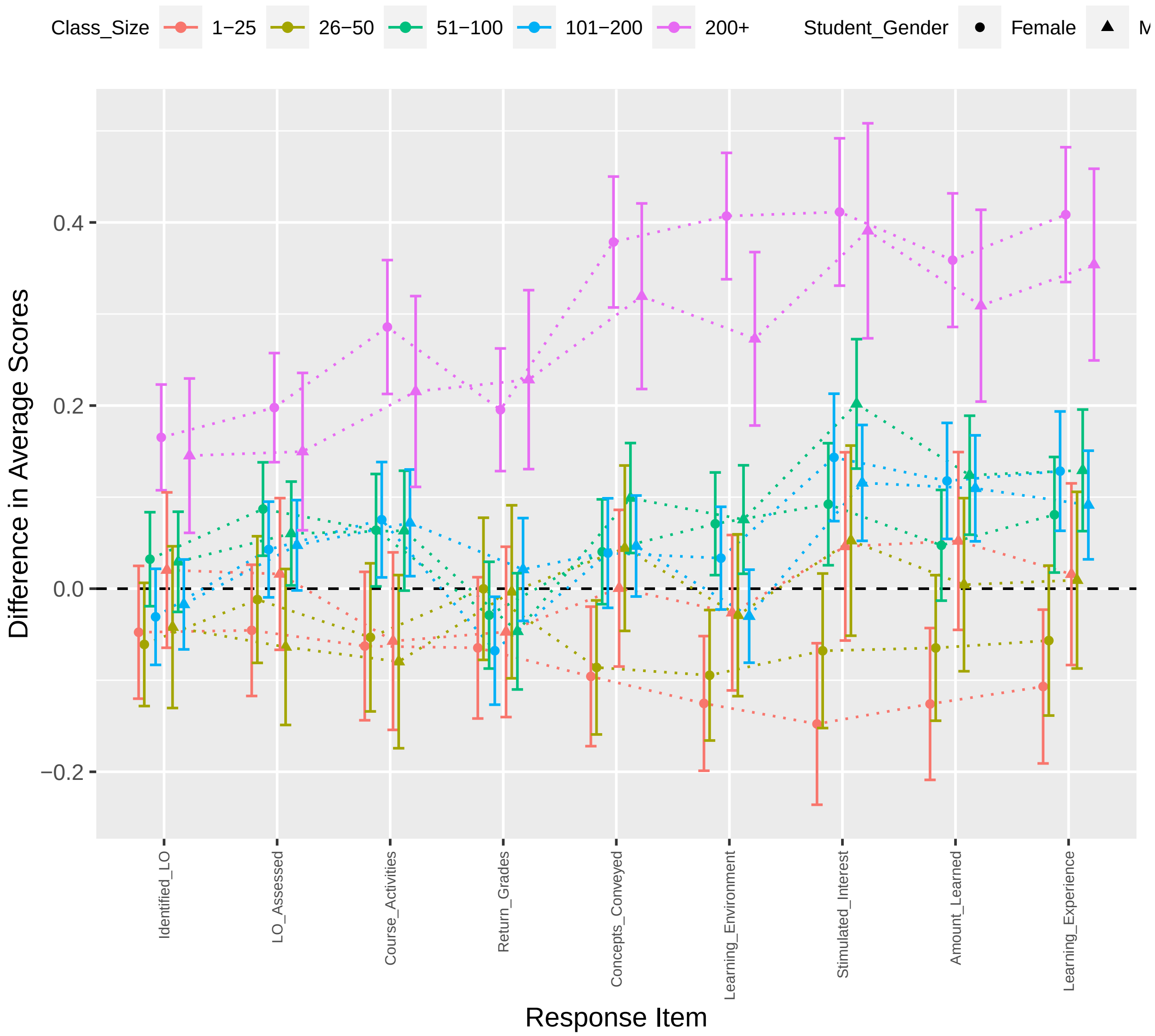


Figure 14: Difference in average score between male and female instructors per Response Item by Student Gender and Class Size.

scores assigned by both male and female students tend to cluster around the zero line, with differences in average scores of < 0.1 points, and usually substantially less. In other words, for both male and female students and male and female instructors and in every class size, the error bars across the items cross the zero-point and are almost completely parallel to one another.

Figure 15 shows the average difference in scores between male and female instructors, accounting for class size and students' expected grade (self-reported).In general, the difference between male and female instructor SCP scores decreases as the students expected grade increases. Almost none of the differences are statistically significant (as is evidenced by the fact that the error bars cover zero), except for the classes of 200+ students.

Figure 16 shows the average difference in scores between male and female instructors, accounting for class size and course workload (student self-reported). Once again, the significant differences are almost all in the 200+ class sizes. It appears that instructor gender differences are smaller for students who perceive the

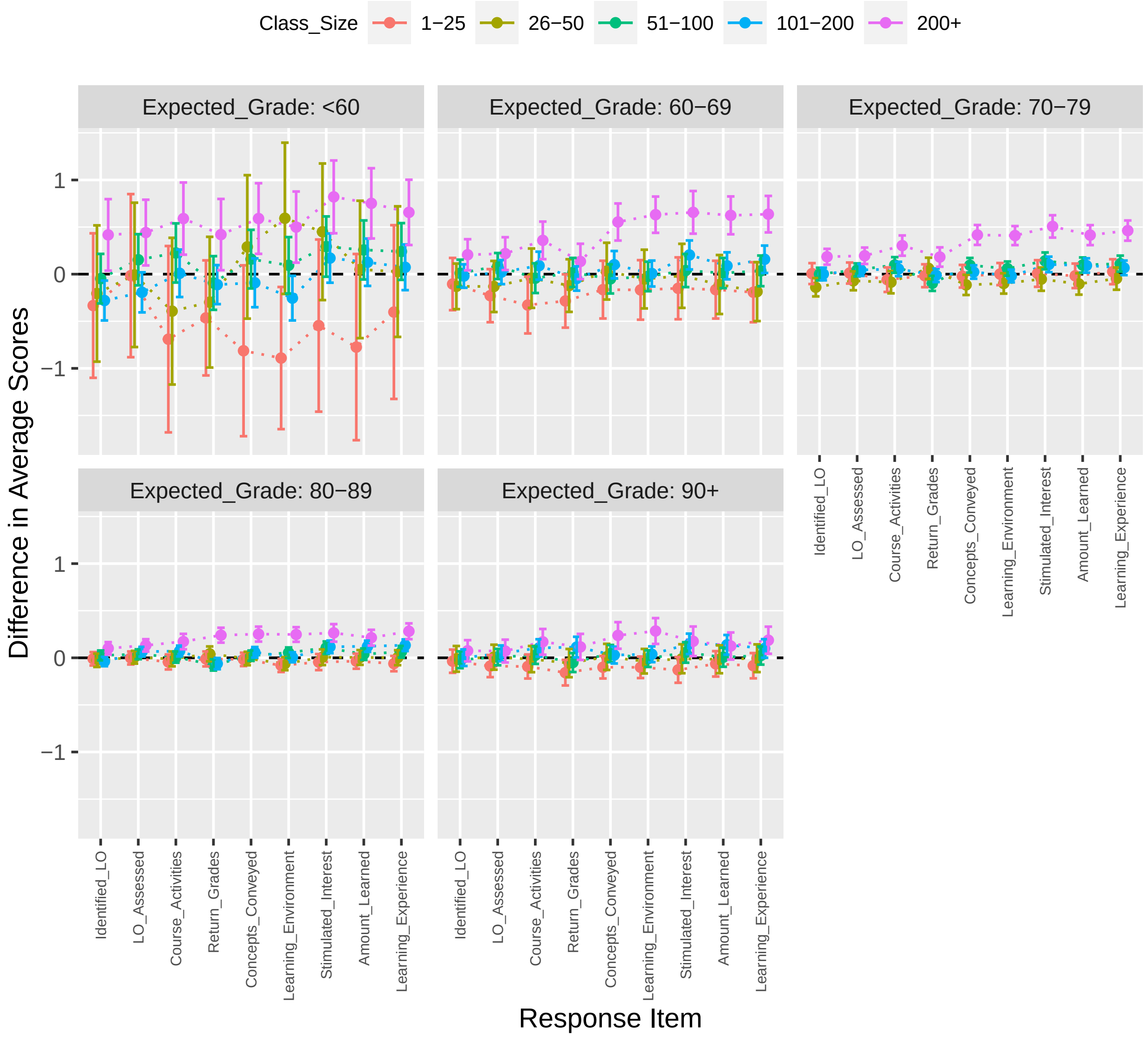


Figure 15: Difference in average score between male and female instructors per Response Item by Expected Grade and Class Size.

workload as `Average` or `High`, as opposed to `Very_Low/Low` or `Very_High`. However, the large error bars for either of the workload extremes are due to only 10% of students choosing either of these categories, as opposed to the middle categories `Average` and `High`.

Figure 17 displays the average difference in scores between male and female instructors, accounting for class size and appointment type. At first glance, these graphs indicate that instructor appointment status may be the most salient variable when combined with instructor gender when considering the difference in scores in large classes. We notice that the graphs for `Tenure`, `Definite_Term`, `Continuing`, `Regular` and `Other` instructors show a clustering of scores around 0 across all class sizes and survey items. This result indicates that for these appointment types there is little difference in the average score for male and female instructors. However, when we look at the `Probationary` (middle graph, top of page) appointment type, we can see that female instructors score just over one-point less than male instructors in courses with 200+ students. Similarly, when we look at `Sessional` appointment status (far right graph, middle row) and

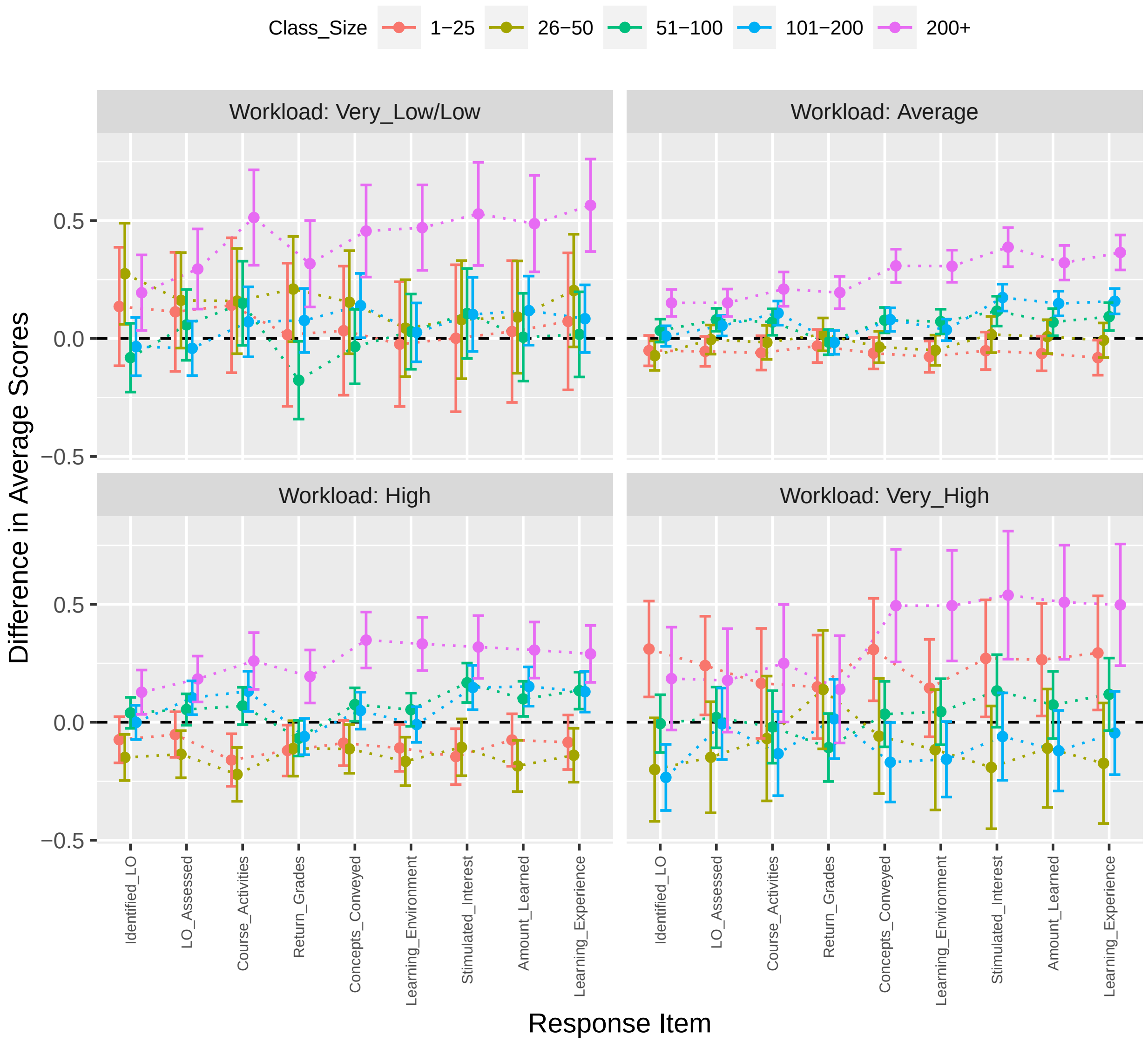


Figure 16: Difference in average score between male and female instructors per Response Item by Workload and Class Size.

Staff appointment (bottom left graph, final row), we can see that female instructors teaching courses with 101-200 students receive about 1-point less than male instructors with the same appointment type, who are teaching courses of the same size. Finally, male Staff instructors tend to fare worse in course sizes of 26-50, scoring about 1-point less than female instructors with the same appointment type.

A one-point difference in the average score on an SCP survey item is rather substantial. In a practical sense, it could mean the difference between a score of 3 versus 4. It is important to consider why we see such disparity in scores for female instructors whose rank is probationary, sessional or staff, while on the other hand we see that male staff instructors fare worse (about one-point) than female instructors in small classes. There are many possible explanations for these findings. At various times, it has been suggested in the research literature that lower scores by appointment type could result from rank representing an approximate proxy for instructor age (McPherson and Jewell, 2007; Zabaleta, 2007), or rank (Isely and Singh, 2005; McPherson and Jewell, 2007). Alternatively, differences in scores across instructor ranks

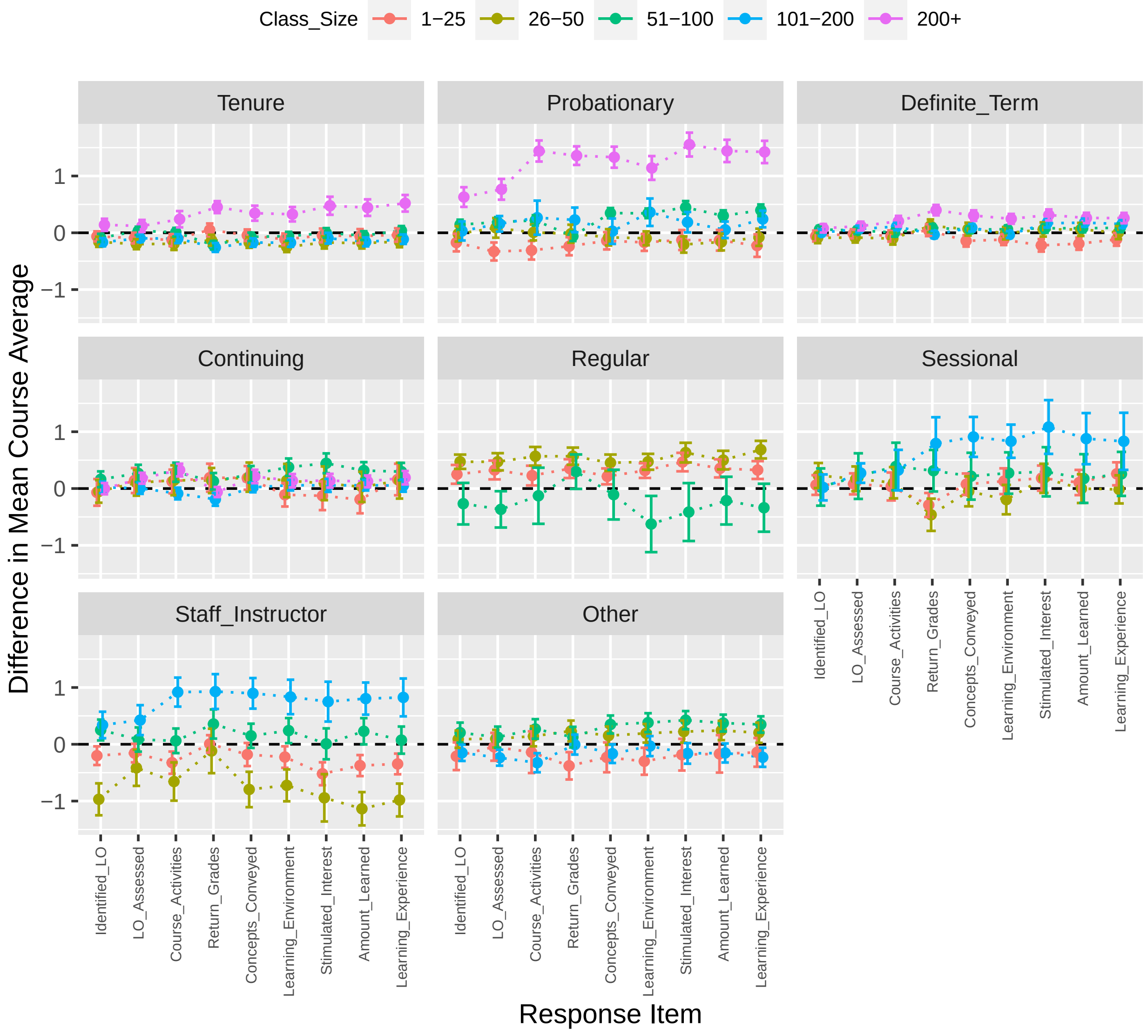


Figure 17: Difference in mean course average between male and female instructors per Response Item by Appointment Type.

could result from students' perceptions of power or status (**?**; of course, it is not clear that most students possess a clear understanding of the rank and status hierarchy in a university).

Given the potential implications of these findings for some of the most vulnerable female faculty members (e.g., probationary, staff etc), we wanted to investigate these findings further. In order to rule out spurious findings, it is important to carefully consider the sample size and how it might explain the differences we are seeing in scores between male and female instructors. As a result, we ran some descriptive statistics to examine the actual raw numbers of instructors by rank. Figure 18 displays this data.

Figure 17 is concerning because it suggests that probationary female instructors teaching course sizes of 200+ receive SCP scores up to one full point lower than male instructors of the same rank, teaching the same class size. However, Figure 18 indicates we need to be very cautious about assigning too much weight to this finding. The sample sizes are below the threshhold where they can be reported precisely while still preserving anonymity, so we report only that there were fewer than five courses

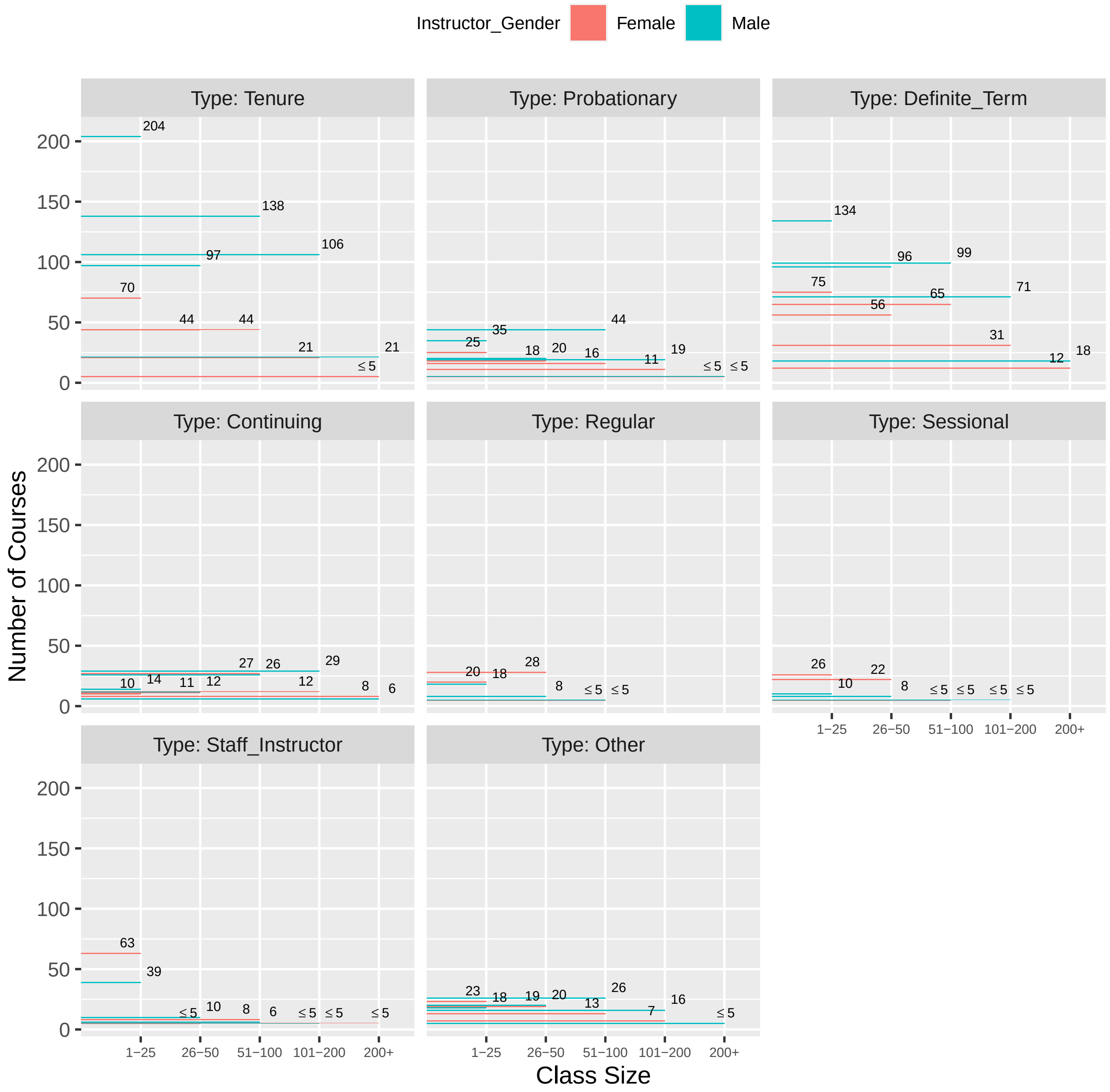


Figure 18: Number of courses taught by female and male instructors by class size for each appointment type.

with enrollments of 200+ and probationary male instructors, and fewer than five with probationary female instructors. Similarly, if we look at the `Sessional` and `Staff` graphs, we see that there were five or fewer female instructors and male instructors with those appointment types teaching courses with 101-200 students. Therefore, apparent differences between instructor gender for particular combinations of appointment type and class size must be approached with caution and should be further studied when we have more data on the new instrument. In other words, these differences should not be over-interpreted (there simply is not enough data to do so), but we strongly suggest that this remains an area of importance for future study.

Looking at Figure 18, one may wonder why e.g., the very small number of instructors with `Probationary` appointments in classes of 200 students or more does not translate to large error bars on the corresponding estimates of instructor gender differences in Figure 17. This is because the only source of uncertainty the error bars account for is due to the randomness in the sample of Fall 2018 students who submitted an SCP survey. The error bars take the specific set of instructors for Fall 2018 to be fixed. Therefore, if there was only one female instructor with a probationary appointment teaching a class of 200+ students in Fall 2018, and every student in that class filled out an SCP survey, then in the present statistical framework we know everything about female instructors with a probationary appointment teaching a class of 200+ students *in Fall 2018*, and the error bars would have a width of zero. Clearly such a framework does not allow us to generalize results to a different set of female instructors with these characteristics in a different term. This is a limitation of the present statistical approach. As noted in Section 5, the postulation of additional modelling assumptions which would admit such generalizations is the subject of further inquiry.

Figure 19 shows the difference in mean course average between male and female instructors across each response item, accounting for class size and Faculty.

The most salient differences are for 200+ class sizes in Arts and Environment, and in 101-200 class sizes at Conrad Grebel. However, Figure 20 reveals that the number of instructors teaching courses in these categories is very small, and so once again we caution against drawing firm conclusions about gender differences in large class sizes in different Faculties. The next largest set of differences (0.3-0.5) are in 26-50 class sizes at Renison and in Environment, and in 51-100 class sizes in Environment. These differences are more robust, insfoar as they are based on at least 10 instructor-course pairs for each gender.

9 Regression Analysis

9.1 Student-Level Analysis

We ran a multiple regression model at the student level to examine the effect of our predictor variables on each of the items in the SCP survey. Specifically, we looked at student, instructor, and course variables including the effect of the following:

- Student-Level Variables: Student_Gender, Workload, Attendance_Amount, Expected_Grade, Course_Type;
- Instructor-Level Variables: Instructor_Gender;
- Course-Level Variables: Faculty, Class_Size, Attendance_Type.

The following regression model was fit to the student-level SCP responses:

$$\begin{aligned}\texttt{SCP\_score} \sim &\texttt{Instructor\_Gender} + \texttt{Student\_Gender} + \\ &\texttt{Workload} + \texttt{Attendance\_Amount} + \texttt{Expected\_Grade} + \texttt{Course\_Type} + \\ &\texttt{Faculty} + \texttt{Class\_Size} + \texttt{Attendance\_Type}\end{aligned} \tag{1}$$

For the most part, the variable names in the model (1) should be self-evident as they are defined in terms of the questions asked. This includes the following coding scheme:

- SCP_score consisted of each of the following 9 SCP items (dependent variables):

```
[1] "Identified_LO"       "LO_Assessed"          "Course_Activities"
[4] "Return_Grades"       "Concepts_Conveyed"    "Learning_Environment"
```

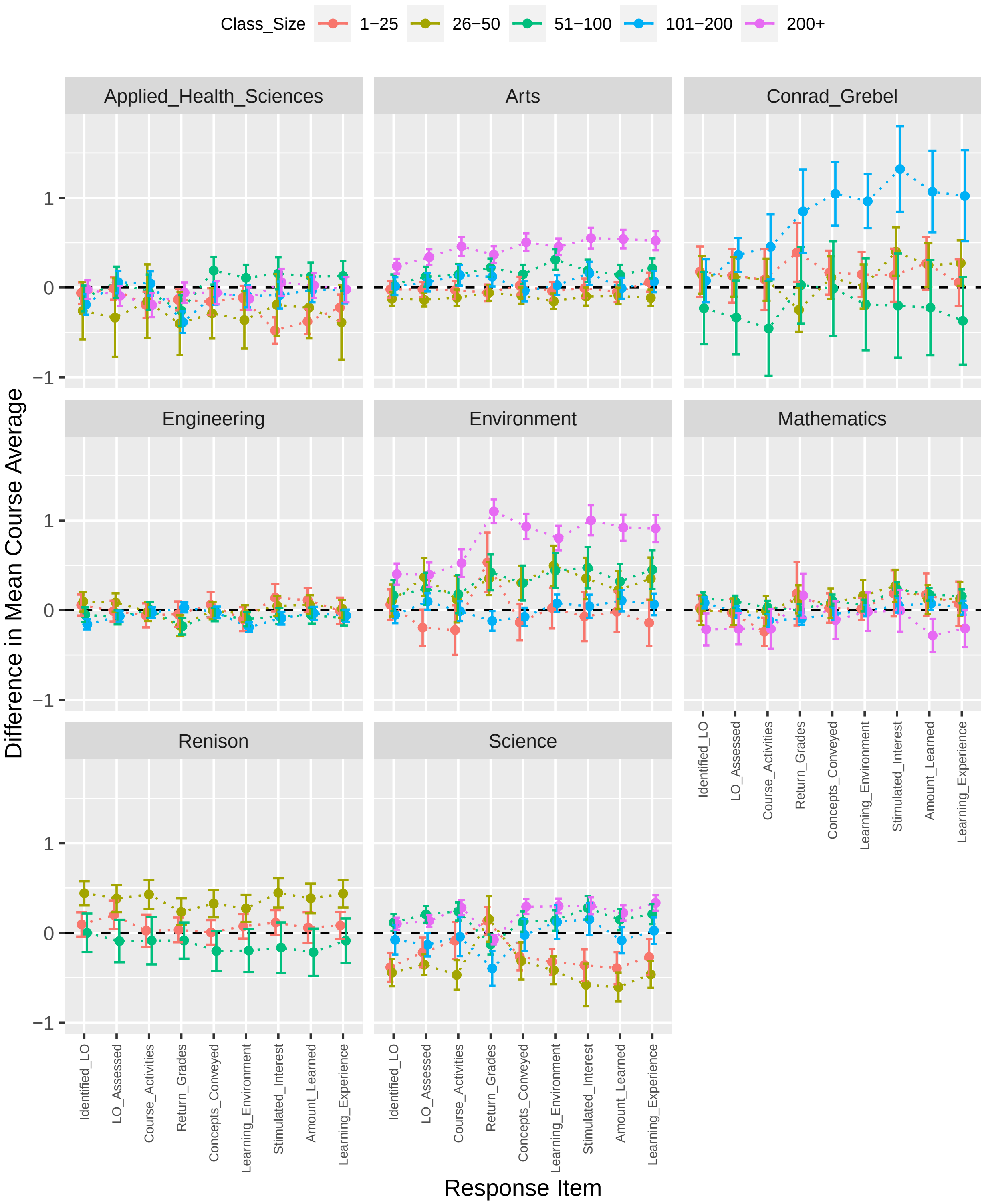


Figure 19: Difference in mean course average between male and female instructors per Response Item by Class Size with facet by Faculty.

```
[7] "Stimulated_Interest"  "Amount_Learned"       "Learning_Experience"
```

Each SCP item was coded on a scale of 1-5 (Likert scale: 'Strongly Disagree' =1 to 'Strongly Agree'=5).

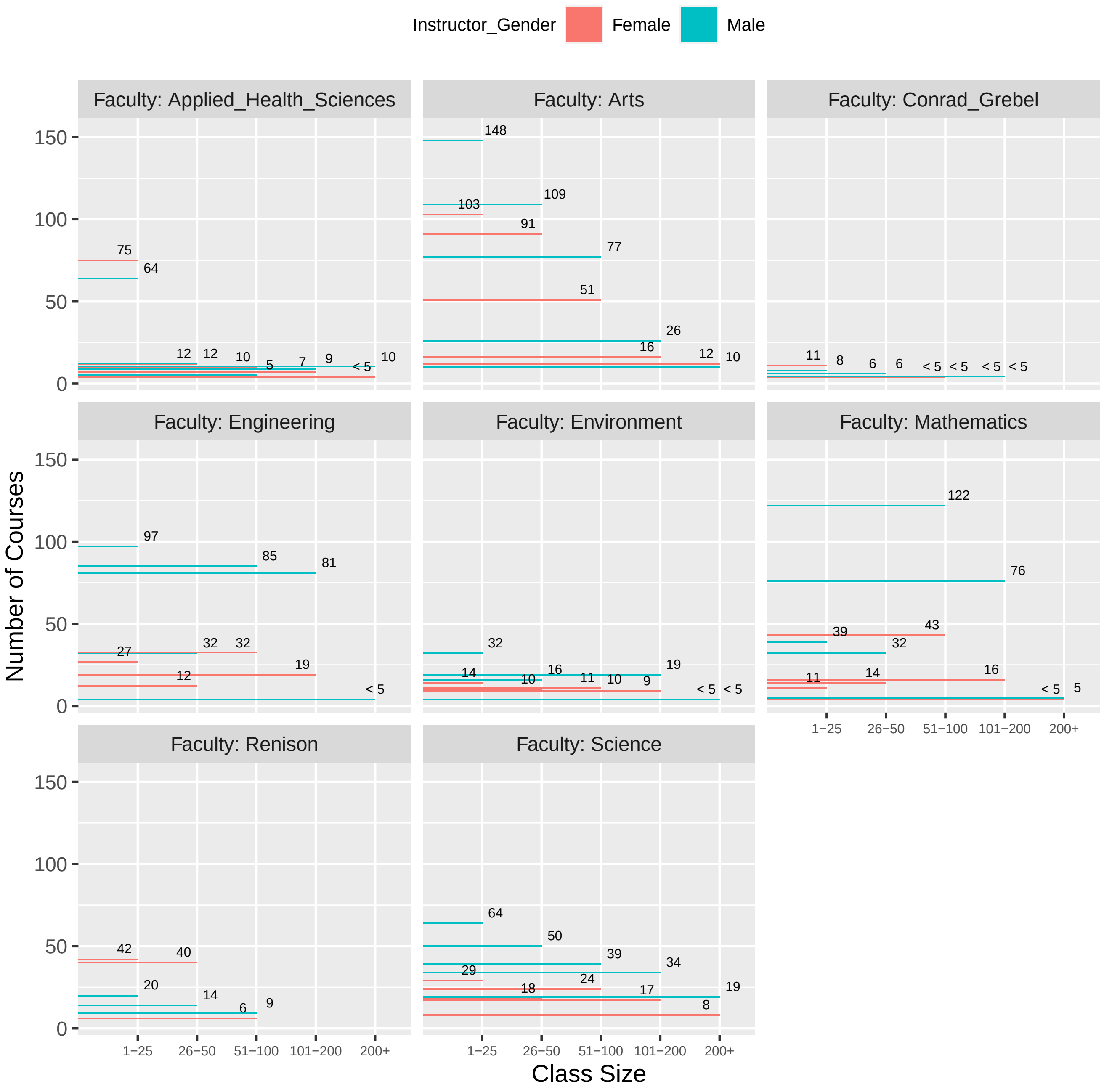


Figure 20: Number of courses taught by female and male instructors by class size within each faculty.

- `Instructor_Gender` was treated as categorical with two gender categories: `Female` and `Male`. The reference category is `Male`.
- `Student_Gender` was divided into the following categories: `Female`, `Male`, `Prefer_No_Ans`, and a combined category `Identified_Other` encompassing students self-identifying as `Non_Binary`, `Agender`, `Not_Listed`, `Genderqueer`, `Trans_Male`, or `Trans_Female`. This last category was created to combine gender categories with extremely small sample sizes. From Figure 2, the number of student-course pairs in each gender category is `Female`: 19870, `Male`: 18695, `Prefer_No_Ans`: 1300, `Identified_Other`: 705.

- Workload was treated as a categorical variable coded as follows:

  ```
  [1] "Very_Low"  "Low"        "Average"   "High"      "Very_High"
  ```

  The reference category is Average.

- Attendance_Amount is coded as follows:

  ```
  [1] "Almost_Never"  "Less_Half"     "Half"          "More_Half"
  [5] "Almost_Always"
  ```

  The reference category is Half.

- Expected_Grade is coded as follows:

  ```
  [1] "<60"   "60-69" "70-79" "80-89" "90+"
  ```

  The reference category is 70-79.

- Course_Type was treated as categorical with two categories: Required and Elective where Required is the reference category.

- Faculty was collapsed as previously described, and treated as categorical. The reference category is Applied_Health.

- Class_Size was collapsed as we previously described, and treated as categorical. The reference category is class size of 51-100 students.

- Attendance_Type was also categorical with In_Class as the reference category and Online as the second category.

A summary of the regression results for each of the nine dependent variables are presented in Table 3. The complete regression output is provided in Appendix D.

Table 3: Coefficients of separate linear models fit to each response variable at the student level. Dots correspond to entries with p-value larger than the pre-specified cutoff level of 0.01. The parentheses indicate entries with p-value between 0.01 and 0.001. Bold values correspond to the 25 percent of entries with the largest absolute magnitude (e.g., $|-10| > |1|$).

	Identified_LO	**LO_Assessed**	**Course_Activities**	**Return_Grades**	**Concepts_Conveyed**	**Learning_Environment**	**Stimulated_Interest**	**Amount_Learned**	**Learning_Experience**
(Intercept)	**3.97**	**3.75**	**3.5**	**3.64**	**3.58**	**3.68**	**3.24**	**3.39**	**3.25**
Instructor_Gender									
Female	.	-.04	-.05	.	-.06	-.04	-.12	-.09	-.09
Student_Gender									
Female	-.04	.	-.04	.	-.04	.	-.06	(-.03)	-.05
Identified_Other	.	.	.	.	.	(-.12)	.	.	.
Prefer_No_Ans	-.19	-.16	-.23	-.2	-.21	-.24	-.21	-.21	-.25
Workload									
Very_Low	**-.42**	**-.48**	**-.46**	**-.34**	**-.38**	**-.36**	**-.54**	**-.7**	**-.58**
Low	-.12	-.15	-.19	-.08	-.13	-.08	-.21	-.27	-.2
High	.	.	-.06	-.07	.	(-.04)	.	.07	.
Very_High	-.14	-.21	**-.34**	-.25	-.23	-.24	-.18	-.14	**-.31**
Attendance_Amount									
Almost_Never	**-.43**	-.26	**-.33**	(-.16)	**-.65**	**-.47**	**-.59**	**-.64**	**-.53**
Less_Half	-.17	(-.12)	.	.	-.29	-.18	-.27	-.28	-.22
More_Half	.12	.1	.15	.16	.2	.25	.29	**.31**	.27
Almost_Always	.22	.19	.25	.2	**.36**	**.36**	**.5**	**.51**	**.45**
Expected_Grade									
<60	**-.41**	**-.45**	**-.81**	-.24	**-.78**	**-.62**	**-.75**	**-.79**	**-.9**
60-69	-.16	-.23	**-.35**	-.08	-.3	-.27	**-.35**	**-.31**	**-.37**
80-89	.17	.22	**.33**	.15	.25	.21	.28	.22	.29
90+	.24	**.34**	**.51**	.22	**.37**	.3	**.42**	**.33**	**.46**
Course_Type									
Elective	.04	.	.	.05	.08	.05	.16	.1	.1
Faculty									
Arts	.	.12	.18	.18	.17	.13	.15	.17	.19
Conrad_Grebel	.23	**.31**	**.43**	**.38**	**.46**	**.4**	**.56**	**.5**	**.54**
Engineering	.	.07	.14	.08	.1	.	(.07)	.09	.15
Environment	(.07)	.08	(.09)	(-.09)	.14	(.08)	.14	(.09)	.15
Mathematics	.	.21	.24	.21	.28	.2	.24	**.31**	**.35**
Renison	.	.12	.27	.23	.22	.14	.22	.25	.28
Science	.	.1	.16	**.31**	.14	.08	.17	.15	.19
Class_Size									
1-25	.	.06	.	.13	.06	.11	.	.	.
26-50	.	.	.	.09	.	(.05)	.	.	.
101-200	.	.	.	.06	.	.	.	.	.
200+	.	.	.	.08	-.06	-.09	-.09	.	.
Attendance_Type									
Online	.	-.1	-.14	-.17	-.27	-.27	-.23	-.28	-.21
R_Squared	.05	.07	.11	.04	.12	.1	.11	.12	.13

The following example illustrates how to interpret the numbers in **Table** 3: the intercept value of 3.97 for `Identified_LO` corresponds to the estimated student-level average in the reference category (Applied Health Science course, Male Instructor, Male Student, Average workload, etc.). The value of 0.22 for `Almost_Always` attendance type means that, if we consider cases for which `Student_Gender`, `Instructor_Gender`, `Workload`, `Expected_Grade`, `Course_Type`, `Faculty`, `Class_Size`, and `Attendance_Type` are the same, the average difference in `Identified_LO` scores between those who `Almost_Always` attend and those who attend `Half_The_Time` is 0.22. Thus, the Instructor_Gender row in Table 3 indicates that, for students in any given category of gender, workload, attendance amount, etc., the difference in how they rate female instructors vs male instructors is between $-0.04 \sim -0.12$, depending on the item.

Considering student gender, assuming everything else is equal, the average difference between Female and Male students' SCP scores is fairly small (about 0.06). This is notable, because some research literature suggests that student gender is an important explanatory factor (see (Boring et al., 2016; MacNell et al., 2015; Potvin et al., 2009; Stark and Freishtat, 2014; Superson, 2002). It is worth mentioning that our findings are similar to those found at other U15 schools, including the University of Toronto and McGill University (see also (Willits and Brennan, 2017). Interestingly, students who prefer not to state their gender provide somewhat lower SCP scores than Male students ($0.16 \sim 0.25$ units on the $1 \sim 5$ scale).

Perceived workload appears to account for a greater difference in average scores for several SCP items. Interestingly, the trend is non monotonic. Students reporting `Very_High` workload give scores $0.14 \sim 0.34$ lower than those reporting `Average` workload, which is similar to the difference for a `Low` workload compared to `Average`. `Very_Low` workloads are associated with the biggest differences, with scores $0.7 \sim 0.34$ lower than for an `Average` workload.

Expected grade (self-reported) has many of the largest values in the table. All else equal, students expecting a grade of 90+ give scores $0.22 \sim 0.51$ higher than students expecting a grade of 70-79, and $0.45 \sim 0.81$ higher than students expecting grades <60. It is worth bearing in mind, though, that only 2.5% of respondents expect grades less than 60, and 60% of those sampled expect grades of 80 or higher.

The findings for perceived workload and expected grade are interesting given the not infrequent appearance of opinion pieces in various venues suggesting that course evaluations have led to a grade inflation ''crisis,'' as instructors give higher marks and demand less of their students in pursuit of higher evaluation scores. A number of researchers have examined the relationship between students' expected grades and their rating of instructors (Braskamp and Ory, 1994; Centra, 2003; Clayson et al., 2006; Feldman, 1976); (Isely and Singh, 2005; Marsh and Dunkin, 1997; Marsh and Roche, 2000; McPherson and Jewell, 2007; Willits and Brennan, 2017; Wines and Lau, 2006; Zabaleta, 2007). Most research finds positive, though often low to modest correlations.

While we can say that our results suggest that instructors who reduce workloads in search of higher SCP scores might be pursuing a bad strategy, the grade inflation question is more complicated. Our results do find a relationship between higher expected grades and higher SCP scores. We do not have the appropriate data to come to any conclusions, but we do note that there is debate in the literature about the nature of any causal relationships between grades and SCP scores. Among the suggestions are (Benton and Cashin, 2014; Feldman, 2007):

- The validity hypothesis: students who learn more receive higher grades and assign higher ratings.
- The leniency hypothesis: instructors can "buy" high ratings if they assign high grades to students.

- Student characteristics: highly motivated/interested students lead to greater learning, which leads to both higher grades and ratings.

For Attendance_Amount, students with higher self-reported attendance gave higher SCP scores, with students reporting they Almost_Never attended giving much lower scores (0.16 ∼ 0.65) than those attending Half_The_Time and 0.36 ∼ 0.9 less that those attending Almost_Always. It is worth recalling from Figure 7 that just above 1% of those completing the pilot test reported attending class Almost_Never, and only about 3% were in either the Almost_Never and the Less_Half categories.

The values for Course_Type reveal that compared to required courses, elective courses have a slightly higher average across all but two SCP items (and a look at the detailed tables in Appendix D which report the values with p values above the cut-off show that this is true for all items), but the differences are marginal, with values ranging from 0.04 ∼ 0.16.

The table values for Class_Size are relatively small (0.05 ∼ 0.13). There is, however, a tendency for student ratings to be higher in smaller classes than larger ones, which aligns with existing research (Algozzine et al., 2004; Feldman, 2007) and findings in large-scale analyses completed at McGill University and the University of Toronto. Both institutions found that, compared to smaller classes, larger classes were rated less favorably. The differences we observed were smaller than those observed at either of those institutions, and it is worth nothing that at both UofT and McGill class size was found to be more strongly associated with course evaluation scores than instructor gender. We return to the question of class size when considering the course-level analysis below.

The reference category for Attendence_Type is In_Class (versus Online). The table values reveal that online courses receive lower average scores across almost all SCP items, with differences ranging from 0.1 ∼ 0.28, depending on the item.

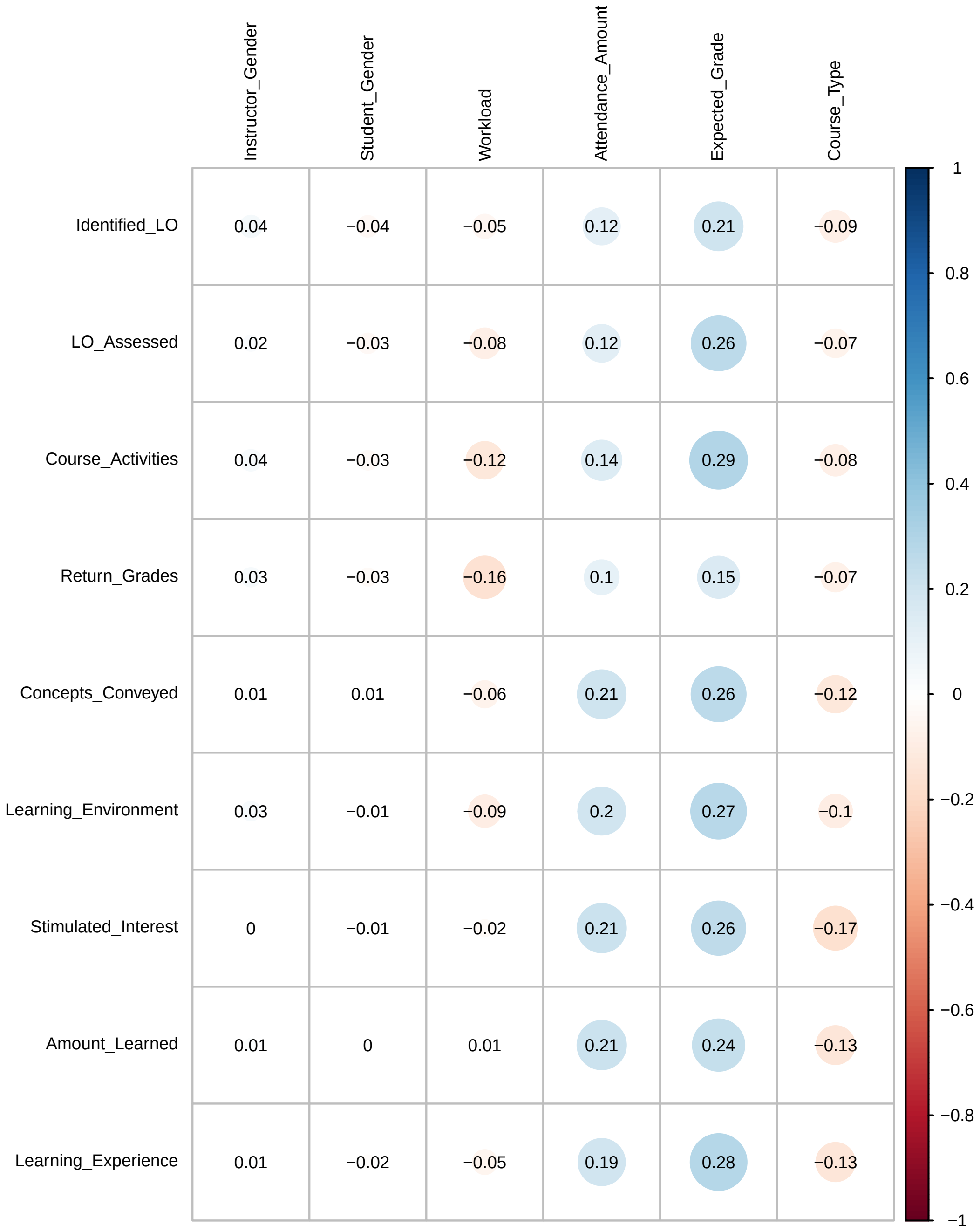
In Appendix D we provide tables with the complete regression models at the student-level for each of the nine SCP items individually. These models include the estimate for each explanatory variable (including those for which the p-value was larger than the cutoff of 0.01, and thus were not included in the summary table), standard error, statistic (z-score) and the p-value for each of the predictors included in our model.

9.2 Course-Level Analysis

While the information at the student level of analysis is important and useful, many will likely feel that information about the results at the course level are more pertinent. It is, after all, average results per course that are reported to instructors, and that instructors report to administrators for purposes of performance review. We therefore also carried out a variety of investigations at the course level. The regression model employed here is the same as that of model (1) in Section 9.1, with the following changes to student-level variables which needed to be aggregated at the course level:

- SCP_Score: the average SCP score for each course.
- Student_Gender: the percentage of female students in each course.
- Workload: the percentage of students in each course who rated the workload as High or Very High.
- Attendance_Amount: the percentage of students who reported attending class more than half of the time in each course (categories More_Half and Almost_Always).

- `Expected_Grade`: average expected scores of each course, where < 60 was coded as `55`, 90+ was coded as `95` with all the other categories being represented by the midpoint of the category range (i.e., 70-79 becomes 75).
- `Course_Type`: the percentage of students in the course who indicated that the course was `Required`.
- The other instructor-level and class-level variables (Instructor_Gender, Faculty, Class_Size) do not require averaging, as they are consistent within a single course.



	Instructor_Gender	Student_Gender	Workload	Attendance_Amount	Expected_Grade	Course_Type
Identified_LO	0.04	−0.04	−0.05	0.12	0.21	−0.09
LO_Assessed	0.02	−0.03	−0.08	0.12	0.26	−0.07
Course_Activities	0.04	−0.03	−0.12	0.14	0.29	−0.08
Return_Grades	0.03	−0.03	−0.16	0.1	0.15	−0.07
Concepts_Conveyed	0.01	0.01	−0.06	0.21	0.26	−0.12
Learning_Environment	0.03	−0.01	−0.09	0.2	0.27	−0.1
Stimulated_Interest	0	−0.01	−0.02	0.21	0.26	−0.17
Amount_Learned	0.01	0	0.01	0.21	0.24	−0.13
Learning_Experience	0.01	−0.02	−0.05	0.19	0.28	−0.13

Figure 21: Correlation matrix between SCP response items and predictor variables

Figure 21 displays a correlation matrix of all nine SCP items on the pilot test and the predictor variables included in our model (`Instructor_Gender`, `Student_Gender`, `Workload`, `Attendance_Amount`, Expected_Grade, Course_Type). The correlation matrix reveals that for the most part the SCP items included on the pilot test are not correlated with the model predictors. In general, for Instructor_Gender, Student_Gender, Workload, Course_Type, we see virtually no association, with correlations between items ranging between -0.01 and 0.17. For the items Attendance_Amount and Expected_Grade, we see weak to moderate correlations at most, with values ranging between 0.1 and 0.29. If we compare the correlation values in this table to the correlation matrix of response items in Appendix A, we can see that there are strong correlations among the SCP pilot test items, but not with the predictor variables included in the pilot test.

Next, we performed a weighted linear regression according to the class size of each class/course. The regression output for each of the nine SCP items included in the pilot test are provided in Table 4. The complete output for each item is provided in Appendix E.

Table 4: Coefficients of separate linear models fit to each response variable at the course level. Dots correspond to entries with p-value larger than 0.01. Parentheses correspond to entries with p-value between 0.001 and 0.01. Bold values correspond to the 25% of displayed entries with the largest magnitude.

	Identified_LO	**LO_Assessed**	**Course_Activities**	**Return_Grades**	**Concepts_Conveyed**	**Learning_Environment**	**Stimulated_Interest**	**Amount_Learned**	**Learning_Experience**
(Intercept)	**4.16**	**4.02**	**3.85**	**3.92**	**4.21**	**4.29**	**4.1**	**4.19**	**3.94**
Instructor_Gender	.	.	.	.	(-.07)	(-.06)	-.12	(-.08)	-.1
Student_Gender	(-.13)	-.15	(-.16)	.	.	.	.	.	.
Workload	.	-.12	-.23	**-.35**	(-.13)	-.21	.	.	.
Attendance_Amount	**.78**	**.69**	**.95**	**.76**	**1.29**	**1.14**	**1.52**	**1.46**	**1.4**
Expected_Grade	.01	.02	.03	.01	.02	.02	.02	.02	.03
Course_Type	.	.	.	.	-.12	.	-.26	-.17	-.16
Faculty									
Arts	.	.17	.22	.22	.18	.15	.	(.16)	.2
Conrad_Grebel	(.24)	**.33**	**.45**	.	**.43**	**.37**	**.5**	**.45**	**.5**
Engineering	.	.15	.23	.	.18	.	.	.	.23
Environment	.19	.22	.26	.	**.3**	.25	**.3**	.25	**.32**
Mathematics	.	.27	**.3**	.26	**.29**	.23	.23	**.31**	**.36**
Renison	.	.	.26	(.26)	(.24)	.	.	(.24)	**(.28)**
Science	.	.15	.21	.27	.18	(.13)	.19	.18	.23
Class_Size									
1-25	.	.	.	(.14)	.	.	.	.	.
26-50	.	.	.	.	.	.	.	.	.
101-200	.	.	.	.	.	.	.	.	.
200+	.	.	.	(.12)	.	(-.09)	.	.	.
Attendance_Type	.	.	.	.	-.19	-.2	(-.17)	-.2	.
R_Squared	.12	.16	.19	.09	.19	.18	.19	.18	.2

As was the case with the student-level regression model, the regression analysis at the course level reveals that instructor gender has a small influence on average scores for some SCP items. The values for Instructor_Gender hover around −0.06 to −0.12 for different items which means that, all else being equal, male instructors receive a course average score at most about 0.1 points higher for some survey items compared to female instructors.

The coefficients for student gender reveal that holding all other factors constant, courses with 1% more female students had LO_Assessed averages -0.15/100 units lower. So for example, for courses differing by 20% in the number of female students but otherwise the same, the average score is predicted to differ by -0.15/100 x 20 = -0.03.

The coefficients for Workload reveal that holding all other factors constant, an increase of one percent of students who rate the course workload as High or Very_High in a course will lead to 0.16 × 1% points higher course average score on Amount_Learned and a decrease of 0.3 × 1% points in the average score for Return_Grades.

With respect to Attendence_Amount, holding all else constant, a one percentage increase in the number of students who report attending class more than half of the time results in the highest coefficients on this table. One might think that this could result in substantial differences in scores between courses. To use an example similar to the one just considered, if in one course 50% of students who completed a survey reported attending more than half the time, while in another it was 70%, since a one percent increase in the number of students who report attending class more than half of the time results in a 1.52 × 1% point increase in the average score for Stimulated_Interest (p=0.00), this translates to a difference of 0.34 points for that item. However, it is worth remembering that we saw in Figure 7 that 93% of students who completed the survey reported attending class more than half of the time, so scenarios like the one imagined are highly unlikely.

The coefficients for Course_Type reveal that, holding all else constant, an increase of one percent in the proportion of students taking the course as Required results in a slight decrease in the average score for some of the SCP items. For example, a one percentage increase in the number of students taking the course Required results in an average score that is 0.26 × 1% points less for Stimulated_Interest (p= 0.00), and a score that is 0.17 × 1% (p=0.00) points less for Amount_Learned. Thus a course that is required for 100% of the students responding to the survey vs. one where the course is required for none will expect a score 0.26 less for Stimulated_Interest and 0.17 less for Amount_Learned.

With respect to Class_Size, we see that most values fail to meet the significance cut-off of 0.01. We do see that students in classes with enrollments of 200+ rate the Learning_Environment 0.09 (p=0.00) points lower compared to courses with 51-100 students.

The coefficients for Attendance_Type reveal that compared to in-class courses, online courses receive lower average scores on some SCP items. For example, online courses receive an average score that is about 0.2 (p=0.00) points less for Learning_Environment and 0.09 (p=0.00) points less on Stimulated_Interest, compared to in-class courses.

In Appendix E we include tables with the complete regression models at the course level for each of the nine SCP items individually. These models provide the estimate, standard error, statistic (z-score) and the p-value for each of the predictors included in our model.

In the next section of the report, we turn our attention to a detailed analysis of three potential composite measures, derived from a qualitative factor analysis of the the items included in our SCP pilot test survey.

10 Composite Metrics Analysis

One key research aim of the pilot test was to evaluate the extent to which SCP items could be grouped into the three theoretical learning constructs identified by CEPT1, namely: Course Design, Course Delivery, and Learning Experience. The SCP survey questions were created with the intention of measuring the three learning constructs identified by CEPT1 as follows:

1. Course Design: Identified_LO, LO_Assessed, Course_Activities.
2. Course Delivery: Stimulated_Interest, Return_Grades, Concepts_Conveyed, Learning_Environment.
3. Learning Experience: Amount_Learned, Learning_Experience.

We compare this theoretical grouping to a data-driven grouping which clusters SCP items according to how closely they correlate with each other (the correlation matrix is given in Figure 29). The exact statistical procedure is a Confirmatory Factor Analysis, the details of which are in Appendix A. The qualitative results of this analysis indicate a grouping of the SCP items somewhat different from that theorized above.

1. Course Design: Identified_LO, LO_Assessed, Course_Activities.
2. Implementation: Amount_Learned, Learning_Experience, Learning_Environment, Concepts_Conveyed, Stimulated_Interest.
3. Return Grades: Return_Grades.

Indeed, while the Factor Analysis confirmed the predicted grouping for the Course Design construct, the items anticipated to theoretically group under Course Delivery and Learning Experience appear to be strongly correlated and thus form a single construct that we have termed Implementation. This is with the exception of the single item Return_Grades, which did not seem to be correlated with the other SCP items. Since this question was intended to gather student perceptions of an important component of course delivery, namely whether the instructor provided timely feedback, we recommend that the question be reworded and then re-tested to examine whether it will then cluster more closely with the other Implementation items.

We consider composite metrics for each student consisting of equal weighting of the SCP items in each of the groupings above. In the following Figures, we repeat the analyses for Instructor Gender differences from Section 8, but with the composite metrics above. We also consider a fourth composite metric, termed Implementation_Simple in the tables below: It is Implementation without the "overall" items Amount_Learned ("Overall I learned a great deal from this instructor") and Learning_Experience ("Overall, the quality of my learning experience in this course was excellent"). CEPT2 was interested in the question of whether this composite measure ameliorated or exacerbated, for instance, differences by instructor gender because this information could help inform a decision on whether it made sense to exclude the "overall" questions from the instrument.

There is little consensus in the literature concerning the use of "overall" or "global" ratings of instruction. Some researchers support the use of overall ratings (see Abrami et al., 2007). On the other hand, one of the principal critiques of such measures is that they fail to adequately represent the multidimensionality of teaching (Marsh, 2007). Moreover, some research suggests that such measures are especially prone to bias (Benton and Cashin, 2014; Cohen, 1980; Theall and Franklin, 2001). As noted above, the focus groups we conducted with students across the six Faculties at UW also revealed that the global items included in our SCP pilot test survey did not align with any of the key themes identified by students. These considerations, and the fact that a smaller set of core questions would leave more room for Faculty-level and Program-level questions in the eventual cascaded instrument, suggested that further investigation was warranted.

As we can see from Figures 22-28, there is very little difference in the composite `Implementation` and `Implementation_Simple` metrics, so the value added by including the “overall” items `Learning_Experience` and `Amount_Learned` is open to question.

Perhaps predictably, the Figures that follow are close in appearance to Figures 11-17. This is what was to be hoped if the goal was to find composite measures that do not exacerbate, for instance, differences by instructor gender. Since the information included in these Figures is very similar to that which we have already described, the commentary that follows will be very minimal.

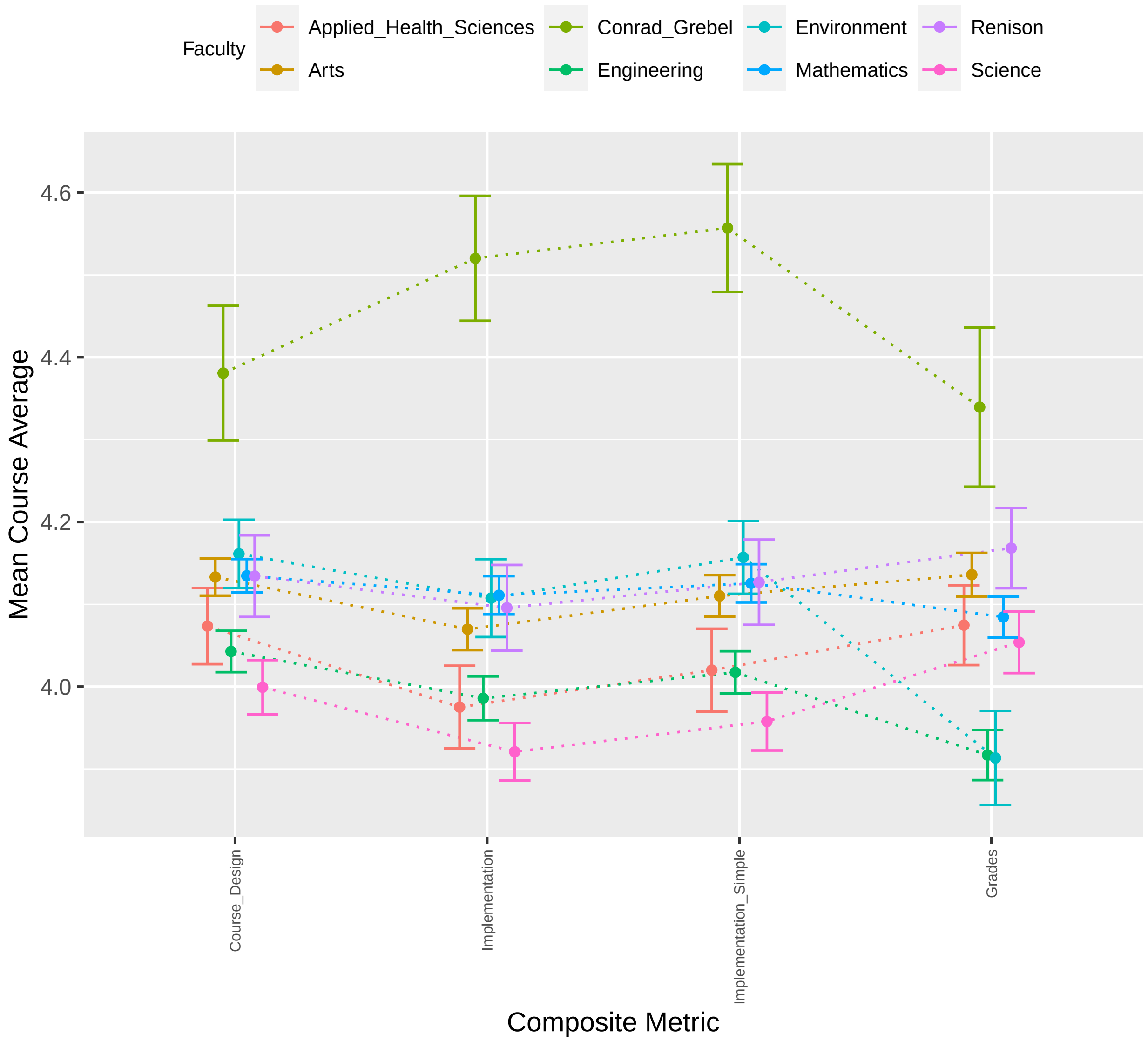


Figure 22: Mean course average per Full Composite Metric by Faculty.

Figure 23 displays the difference in average scores between male and female instructors for each composite measure. We can see the difference in scores in general is quite marginal, up to 0.05 points at most. These findings suggest that averaging the items to create each composite measure does not appear to compound any differences in scores between male and female instructors.

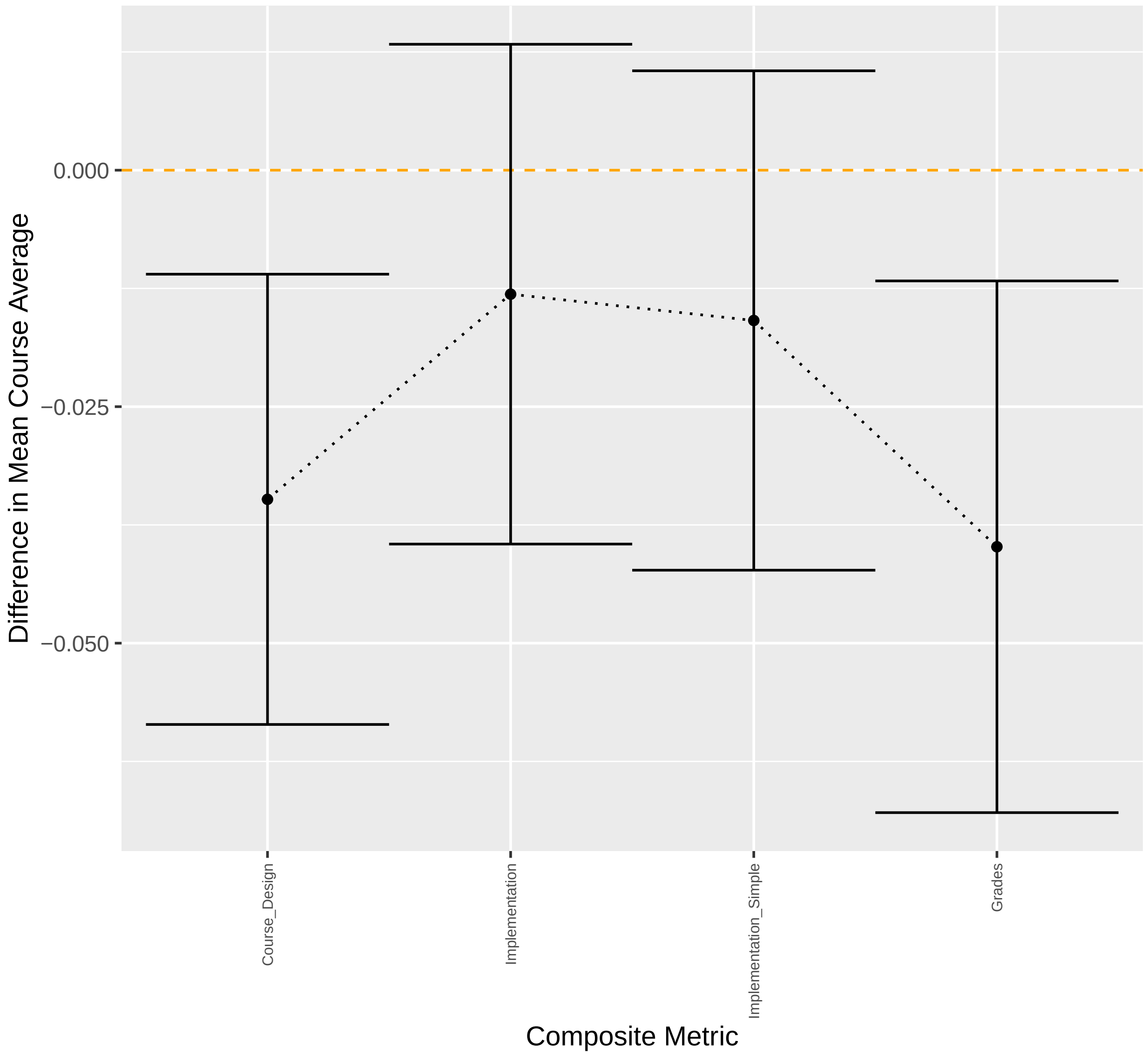


Figure 23: Difference in mean course average between male and female instructors per Full Composite Metric.

Figure 24 displays the difference in the mean composite scores between male and female instructors accounting for class size. In class sizes of 200+, we can see that females score an average of about 0.2 points less for `Course_Design`, 0.4 points less for `Implementation` and about 0.3 points less for `Grades`. As noted earlier, some caution is warranted with respect to these results, given the small number of very large courses in the pilot test.

Figure 28 displays the difference in the mean composite scores between male and female instructors accounting for class size and appointment type. The results, unsurprisingly, mirror those found when the analysis considered individual items rather than composite scores. Once again, the results are interesting. And once again, they must be treated with caution because of the very small numbers of courses on which the most striking results depend.

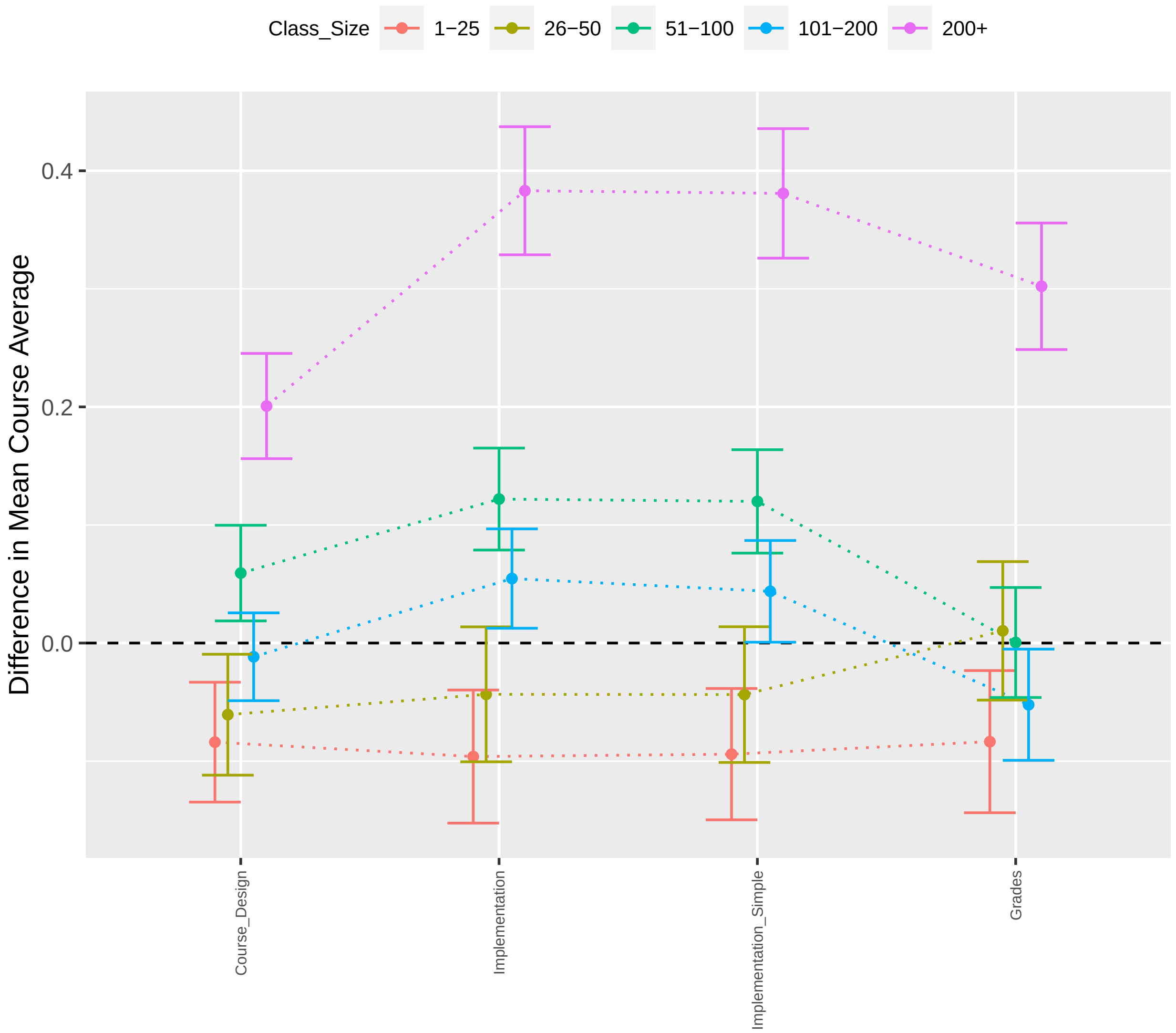


Figure 24: Difference in mean course average between male and female instructors per Full Composite Metric by Class Size.

11 Summary Remarks

The guiding question for this analysis is not whether non-instructional factors are associated with teaching evaluation scores but, rather, to what extent such factors are associated with scores on student course perception surveys (SCPs) at the University of Waterloo. The pilot test results outlined in this report shed some light on this. The literature on student ratings of teaching has looked at the potential influence of a number of factors on ratings. Even acknowledging that some of these factors influence student ratings, it is important to recognize that there is little consistency (and often not much attention to) the effect size for those factors. This analysis offers at least a preliminary look at the extent to which SCP scores, for the instrument tested in the pilot test, are associated with a variety of different factors often discussed in the literature. The large sample size in the pilot test allows a fairly detailed analysis, and so some confidence in some of the conclusions. On the other hand, it was a cross-sectional (one-time) survey, and some suggestive results, especially at the course level of analysis, suffer from very small sample sizes.

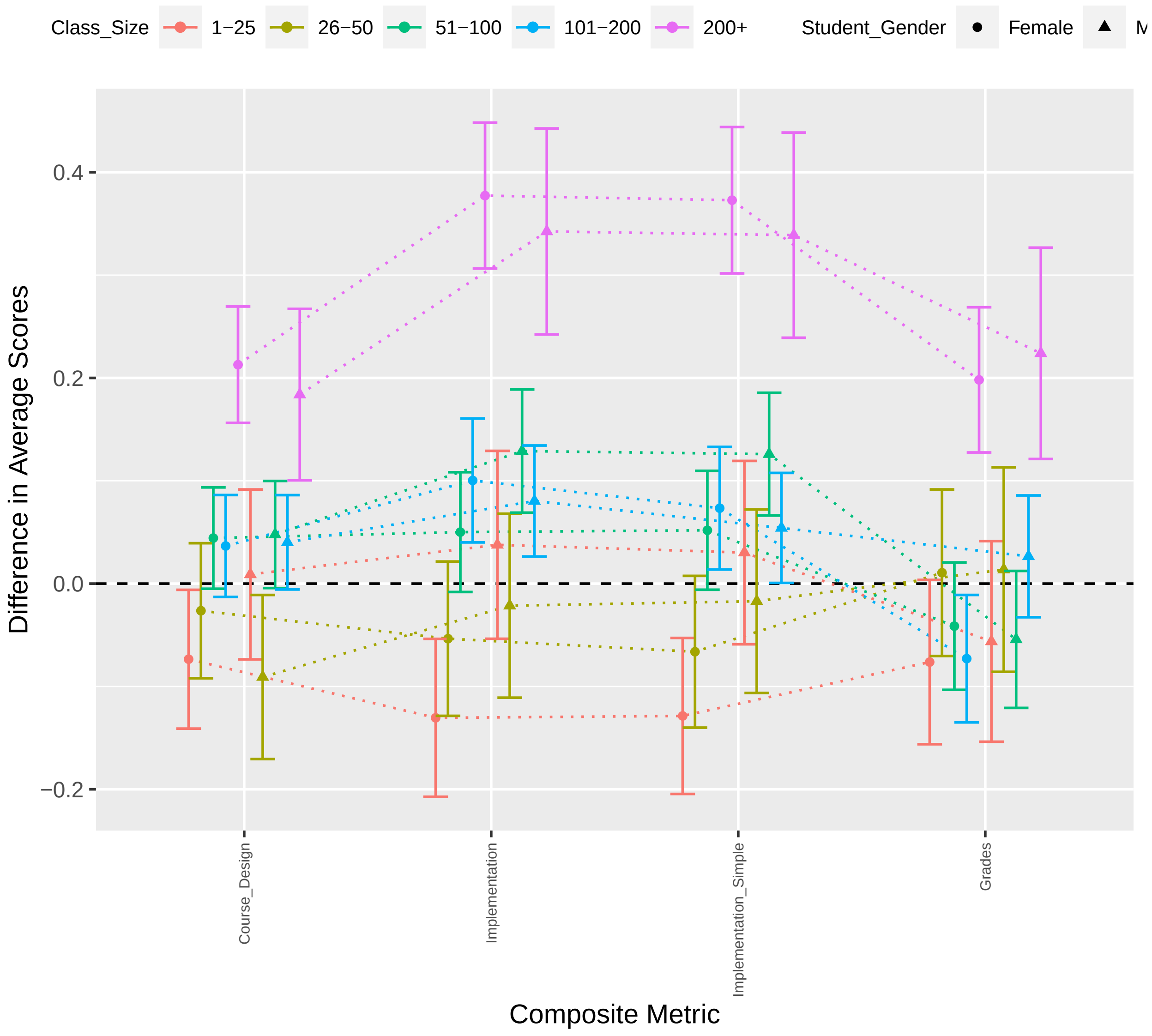


Figure 25: Difference in average score between male and female instructors per Full Composite Metric by Student Gender and Class Size.

Among the notable findings in the report, we think it worth highlighting these.

1. **The categories `Agree` and `Strongly Agree` are the most commonly selected by students on the SCP, with more than three-quarters of the sample selecting these response options across all nine survey items.**

2. **Students who attend class more often tend to give higher scores.**

3. **More than 80% of students who filled out the SCP reported attending class `Almost_Always`.**

   These results should help assuage some concerns sometimes voiced by instructors. One sometimes hears the suggestion that online rather than strictly in-class evaluations provide an opportunity for students who do not come to class to skew results downward. We see in the pilot test that, since very few non-attenders complete the survey, this need not be a significant concern.

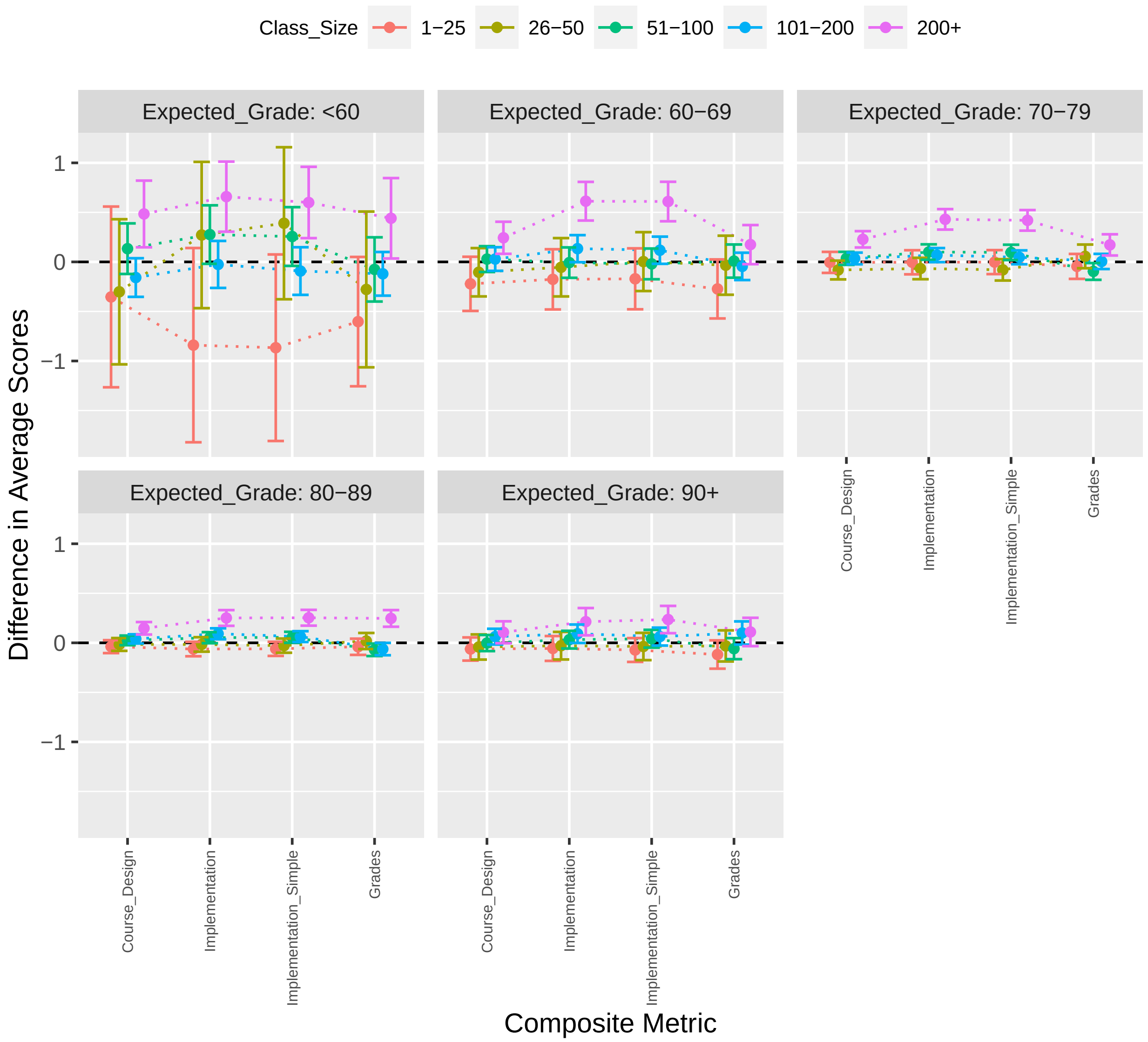


Figure 26: Difference in average score between male and female instructors per Full Composite Metric by Expected Grade and Class Size.

4. **About 85% of student respondents expected to receive a grade of at least 70% in the course, and about 60% expected a grade of 80% or higher.**

5. **Our regression analyses showed that students who expect higher grades tend to give higher scores.**

   There has been some concern in the literature that students use course evaluations as a means of retaliation (e.g., for poor grades). And, indeed, the student level regression suggests that a student who expects a grade of below 60% will on average assign a score between 0.5 and 1.3 points lower than a student expecting a mark higher than 90%, depending on the survey item. On the other hand, the high percentage of those completing the survey who expect marks of at least 80%, coupled with the fact that only about 2.5% of survey respondents expected a grade of less than 60%, should go some way to assuaging the concern of retaliation. It is also worth noting that if an entire class expected to receive ‘low grades’ this may be indicative of poor teaching practices, or more spefically,

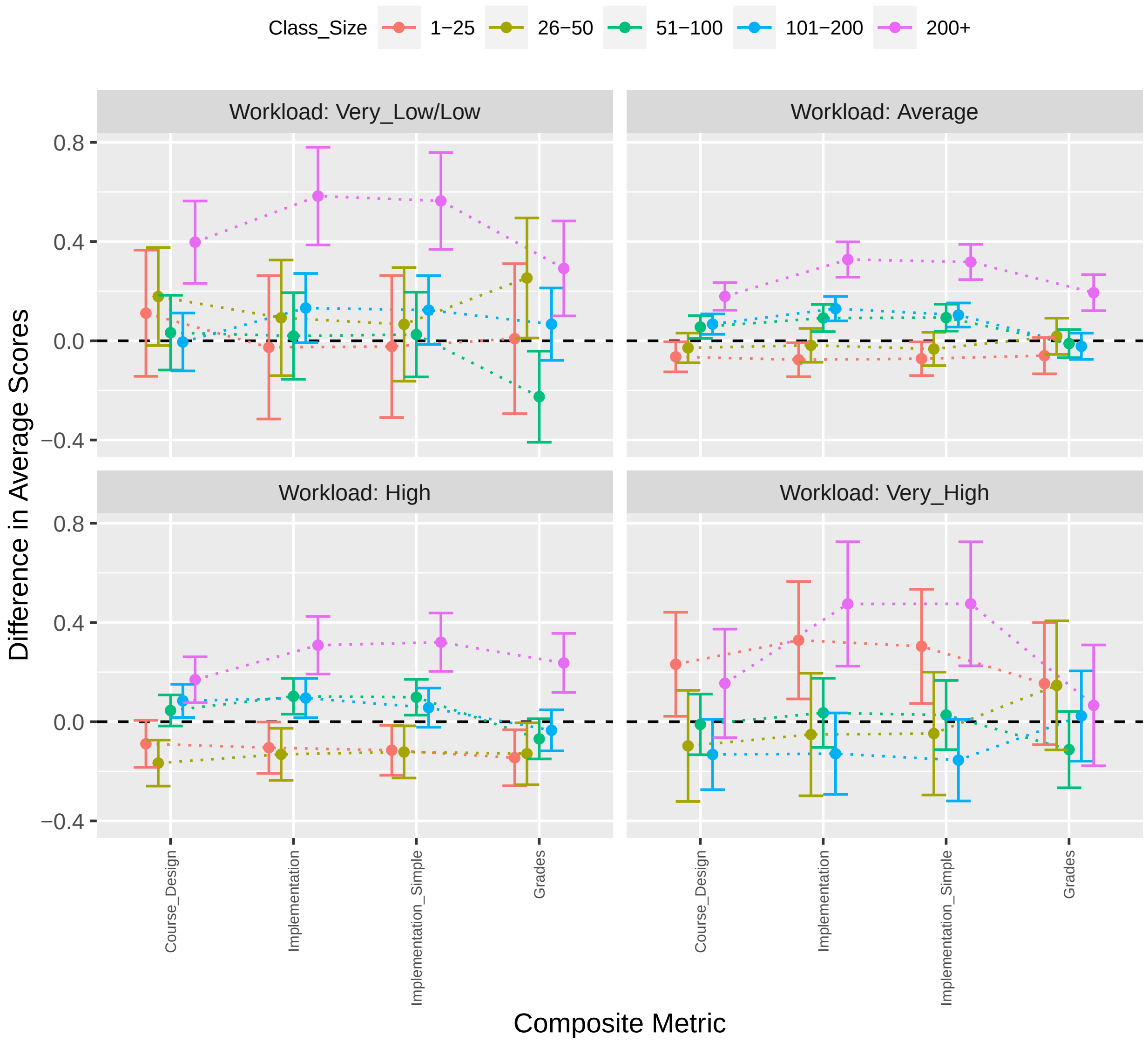


Figure 27: Difference in average score between male and female instructors per Full Composite Metric by Workload and Class Size.

poor assessment practices.

How the results relate to the concern sometimes heard that reliance on student surveys in the evaluation of teaching leads instructors to give higher grades and so fuels grade inflation is a more complex matter. The relationship between expected (or final) grades and scores on course perception surveys has frequently been examined in the literature (Braskamp and Ory, 1994; Centra, 2003; Feldman, 1976);(Marsh, 2007; Marsh and Dunkin, 1997; Marsh and Roche, 2000; Willits and Brennan, 2017). Some studies find a weak relationship here [Beleche et al. (2012);Benton.etal16, while others argue that SET scores are negatively associated with performance in subsequent courses (**?**Carrell and West, 2010). Findings from our SCP survey revealed that higher expected grades are associated with better SCP scores. However, it is worth recalling that the relationship between higher grades and higher scores on student surveys is a contested matter, and that several hypotheses have been offered to explain it.

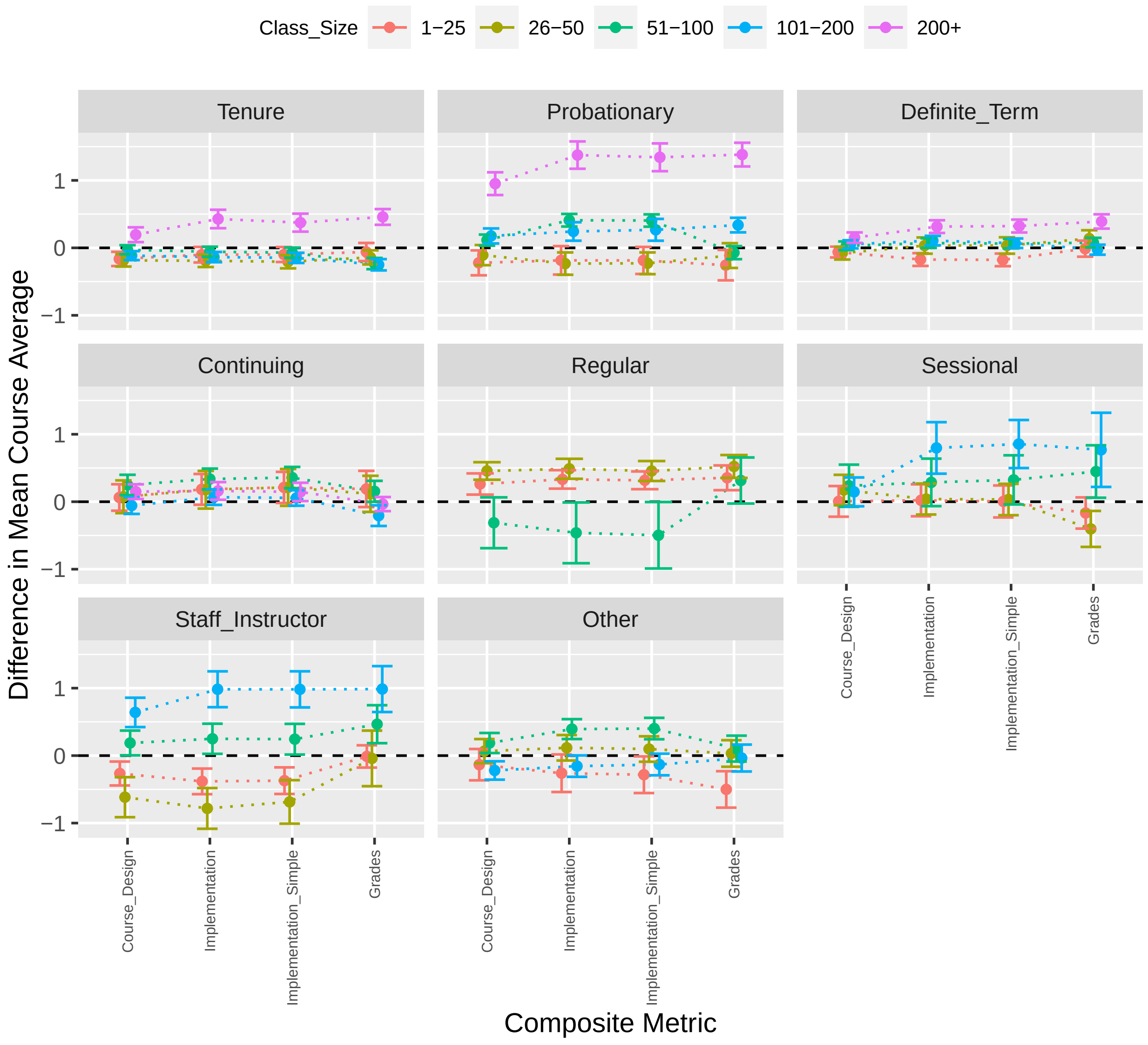


Figure 28: Difference in mean course average between male and female instructors per Full Composite Metric by Appointment Type.

6. **Courses perceived to have very low workloads are associated with much lower SCP scores, and those with low or very high workloads are associated with lower SCP scores.**

   This is an interesting result since it runs contrary to the suggestion, sometimes heard in concert with the claim that reliance on SCPs contributes to grade inflation, that SCPs promote a general ''dumbing down" of courses. Courses that are perceived to be easy, and especially those perceived to be very easy, at least in terms of workload, are associated with lower scores than those perceived to be of average or somewhat above average difficulty.

7. **Larger classes are rated lower than smaller classes.**

   This is something we would have predicted. However, the relationship between class size and SCP scores in the pilot test was smaller than the relationship reported at other Canadian research intensive universities. It will be worth continuing to investigate the size of this relationship if the tested instrument does indeed become the core of Waterloo's new cascaded SCP process.

8. **Online courses receive lower ratings than in-class courses, and have somewhat lower response rates.**

   This result has long been suspected. The pilot test gives us some idea of the size of the differences. Moreover, the presumed differences have sometimes been attributed to existing SCP tools on campus including questions that are manifestly inappropriate for online courses ("the instructor is available to meet" is a popular example). The differences with the draft instrument are harder to attribute to questions that assume in-person delivery of instruction.

9. **Some associations between instructor gender and SCP scores were observed, though generally they were very small.**

   Some of the results considering multiple variables suggested the possiblity of more significant relationships (e.g., differences in scores between *probationary* male and female instructors *of very large enrolment courses*). Unfortunately, in these cases the sample sizes were so small that firm conclusions are very fraught. These relationships should certainly be further studied if the tested instrument is adopted; and in the meanwhile the need to treat scores for these cases with caution should be flagged for the campus community.

   Instructor gender has received extensive attention in the literature (Basow, 2000; Boring et al., 2016; Centra, 2009; Feldman, 1993; MacNell et al., 2015; **?**; Stark and Freishtat, 2014; Willits and Brennan, 2017). The impression of the CEPT was also that this was the area of potential bias in SCP results of most general concern to the campus community. As such, more of our efforts of analysis were devoted to this matter than to any other. At a general level, the results from the pilot test, at the course level, showed that the differences between average scores for male and female instructors was marginal, at most 0.03 points on the five-point scale. These findings are similar to those found at other U15 institutions like the University of Toronto and McGill University, where differences in scores between male and female instructors on their evaluation instruments were no more than 0.1 points on the 5-point scale. As noted in the body of the report, many universities do not report course averages to more than one decimal place in order to avoid over-interpretation of such marginal differences.

   We also considered differences when instructor gender was combined with other predictors, including: student gender, expected grade, workload, faculty, and instructor appointment type, and for a few combinations the relationship to SCP scores when gender was combined with multiple predictors. Sample sizes were too small for firm conclusions, but these analyses reveal some important areas for continued monitoring and investigation.

10. **A confirmatory, qualitative factor analysis provides support for creating two composite measures which we have termed `Course_Design` and `Implementation`.**

    Three of the four SCP items originally intended to measure `Course_Design` (`Identified_LO`, `LO_Assessed`, and `Course_Activities`) were indeed found to group into the same underlying construct. The factor analysis revealed that the SCP items originally intended to fall into the `Course_Delivery` construct (`Stimulated_Interest`, `Return_Grades`, `Concepts_Conveyed`, `Learning_Environment`) also group together, with the exception of `Return_Grades` which turned out not to group with any of the other items. Since what that question was intended to indicate the provision of timely and useful feedback to students is an important component of course implementation, we recommend that an alternative item be produced and tested in future iterations of the SCP process.

However, the items intended to group into `Learning Experience` (`Amount_Learned`, `Learning_Experience`), while they did group together, did not group together into a theoretical construct distinct from `Course_Delivery`. We therefore have termed the construct under which the `Course_Delivery` and `Learning_Experience` questions clustered `Implementation`, intending this as a word that suggests both delivery and experience.

However, both `Amount_Learned` and `Learning_Experience` are ''overall'' questions, as they were worded in the tested SCP. Since such overall questions are sometimes criticized in the literature as especially prone to bias, and these questions are not measuring learning experience separately from course delivery as intended, and since deleting the questions from the composite measure makes essentially no diference to the composite scores, CEPT(2) recommends removing those global items from the SCP. The resulting `Implemenation_Simple` composite score then looks very much like what CEPT1 intended to form the `Course_Delivery` composite when they designed the instrument. This suggests that one area different Faculties might consider for the second tier of questions is whether questions aimed at more specific aspects of the Learning Experience are priorities for teaching in their programs.

CEPT2 understands its remit from the Provost, acting on behalf of the University Senate, to be: Finetune and test a survey instrument, in support of a decision to implement a new SCP survey at Waterloo. As such, we regard the pilot test as a way to detect significant issues that would warn against replacing the existing tools with the new one. The opinion of the CEPT2 committee is that the pilot test results not only did not provide reasons against adoption of the new tool, but that results of the analysis also provide good reason for moving ahead with the new instrument, and valuable information that will help to improve the value and the fairness of teaching evaluation at Waterloo.

A Course-Level Qualitative Factor Analysis

In its final report endorsed by Senate, CEPT1 proposed a draft ''core" survey instrument for the cascaded model consisting of 13 questions. The questions were intended to gather student perceptions about how well the design and delivery of a course facilitated their learning. Based on an extensive review of the literature on effective teaching, CEPT1 organized these items into three conceptual categories: Course Design, Course Delivery and Learning Experience. The 13 items were amended by CEPT2 after conducting focus groups with students from each of the six Faculties, but the original conceptual framework (course design, course delivery, learning experience) did not change. One key research aim of the pilot test was to understand whether it would be legitimate to group the survey items into composite scores corresponding to these categories. A qualitative factor analysis was conducted to examine the extent to which the pilot test items were measuring the same or different underlying constructs.

For the first pass at a factor analysis, CEPT2 did not share information about how it was hoped that the questions would cluster with the Statistical Consulting and Research Unit, which carried out the analysis. Instead, the Unit was asked to carry out an overall factor analysis, and also one for each Faculty individually, in order to assess the extent to which questions might be interpreted differently by students in different Faculties.

The results of the first analysis were essentially similar to those described below, and left CEPT2 with the impression of the implications for composite measures as described in the main body of this report. A second analysis was carried out to check these impressions for accuracy. In the second version, we reordered Factor 2 and Factor 3 for Renison, achieving results more similar to those in the Faculties and at Conrad Grebel.

A.1 Correlation Matrix of Response Items

Prior to running the factor analysis, a correlation matrix of all nine core items on the pilot test was generated (see Figure 29). In the correlation matrix, a large and dark circle between items corresponds with a high correlation between the two items. The correlation matrix reveals that most of the items, with the exception of `Return_Grades`, share relatively high correlations (ranging from 0.55-0.83). This suggests the items are in fact measuring the same underlying construct. A closer look at the correlational values between specific items provides some evidence that there may be more than one construct underlying the data. For example, `Learning_Experience` correlates highly with `Amount_Learned` (r= 0.83) and `Stimulated_Interest` (r=0.8), but less with `Identified_LO` (r=0.59) and `LO_Assessed` (r=0.59). Further exploration was necessary to determine the underlying conceptual framework.

A.2 Factor Analysis Model

The course-level factor analysis model is mathematically stated as

$$\boldsymbol{\mu}_i = \boldsymbol{\Lambda} \boldsymbol{F}_i + \boldsymbol{\epsilon}_i, \tag{2}$$

where the variables in the model are:

- $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iq})$: The vector of average score in class i on each SCP item $j = 1, \ldots, q$. The average scores are assumed to be standardized, such that for each item we have , such that $\mathrm{E}(\mu_{ij}) = 0$ and $\mathrm{var}(\mu_{ij}) = 1$.

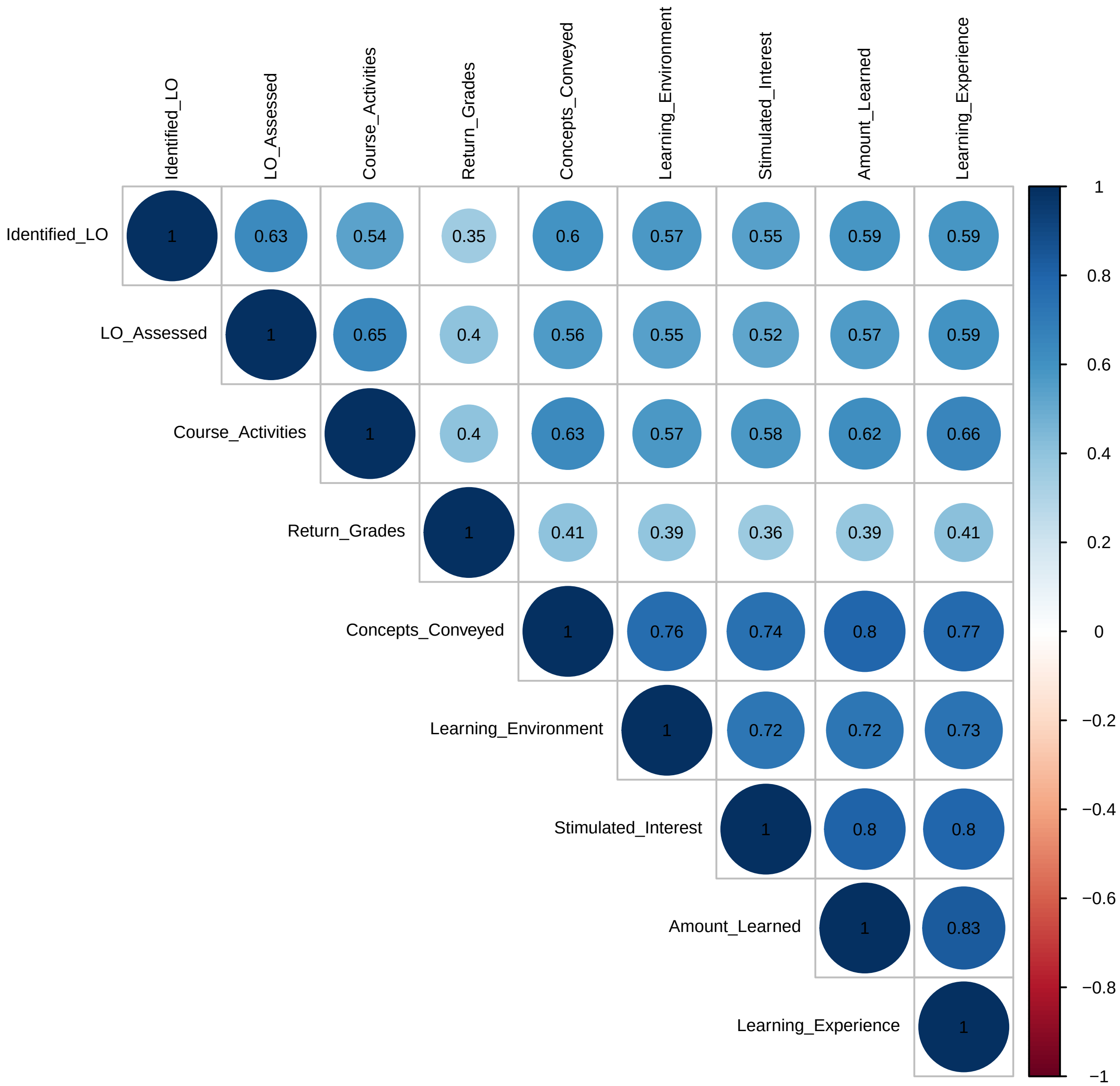


Figure 29: Visualization of the correlation among all response metrics.

- $\boldsymbol{F}_i = (F_{i1}, \dots, F_{iK})$: The vector of K latent factors determining the average score in course i. Each latent factor is given an independent standard normal distribution, such that $F_{ik} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$.
- $\boldsymbol{\Lambda}$: The $q \times K$ factor loading matrix.
- $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{iq})$: The vector of error terms in class i for each SCP item $j = 1, \dots q$. The error terms are assumed to be independent normals for each response item, such that $\epsilon_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma_j^2)$.

However, what is actually observed are the sample means $\bar{\boldsymbol{x}}_i = (\bar{x}_{i1}, \dots, x_{iq})$ per item per class, such that

$$\bar{x}_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(\mu_{ij}, \sigma_j^2/N_i), \tag{3}$$

where N_i is the size of class i. Taken together, equations \eqref{eq:fa_course} and \eqref{eq:fa_sample_mean} form the Factor Analysis model we have fit to the SCP survey data.

A.3 Number of Factors

To determine the number of factors, we looked at the average proportion of variance explained as a function of the number of latent factors. In writing this report we recognized most readers will not have a comprehensive background in statistics. As such, it was our intention to present the results in a manner that would best be understood by the majority of those interested in reading the report. With the advice of our stats consulting colleagues we thus elected to present results for the factor analysis in terms of the average proportion of variance explained, as opposed to the traditional plotting of eigenvalues. It was our view that this would be more intuitive to interpret for those with limited or no background in statistics.

In Figure 30 the y-axis displays for each Faculty,

$$\frac{\text{E}\left[\text{variance contributed by all the factors}\right]}{\text{total variance}} = \frac{\text{communality}}{\text{communality + uniqueness}}$$

This metric conveys similar information we would see with the eigenvalue scree plot that FA software typically produces. The graph reveals that the inclusion of three factors appears to be the most logical for our model. We can see that with three factors in our model there is an average of about 80% explained variance. You will also notice that the model shows close clustering of the Faculties with three factors included, and less clustering beyond three factors. Thus, adding more factors to the model does not appear to increase the average proportion of explained variance and we can see that it also results in less clustering across Faculties. In summary, Figure 30 shows us there is good evidence to move forward with three factors, since there is little explanatory power added to the model by adding additional factors (especially if one focuses on the Faculties, which have larger sample sizes, rather than the AFIW).

For this model the extraction technique we elected to use was oblique transformations (i.e. oblimin in R). This type of transformation was used to help guide clustering of the SCP items into distinct factors. We recognize that the trade-off of adopting an oblique extraction technique is that the factors are correlated; but we believe this is reasonable to assume given that all of the items included on the SCP are designed to measure aspects of effective teaching.

We can also check this type of trend within each response variable, as shown in Figure 31. The graphs displayed here offer further support for including three factors in our model; it is clear from these graphs that the model loses explanatory power beyond three Factors. We can see that across all the items (with the exception of Return_Grades) and Faculties the proportion of explained variance does not improve (increase) beyond three factors, this is evidenced by the levelling off we see once we move to four or five factors for each SCP item. Therefore, based on these analyses, we made the decision to proceed with a three factor model.

A.4 Factor Structure

We ran an explanatory factor analysis to investigate how the items on the SCP loaded onto the three different factors. We then fit the course-level factor model to each Faculty and investigated the factor

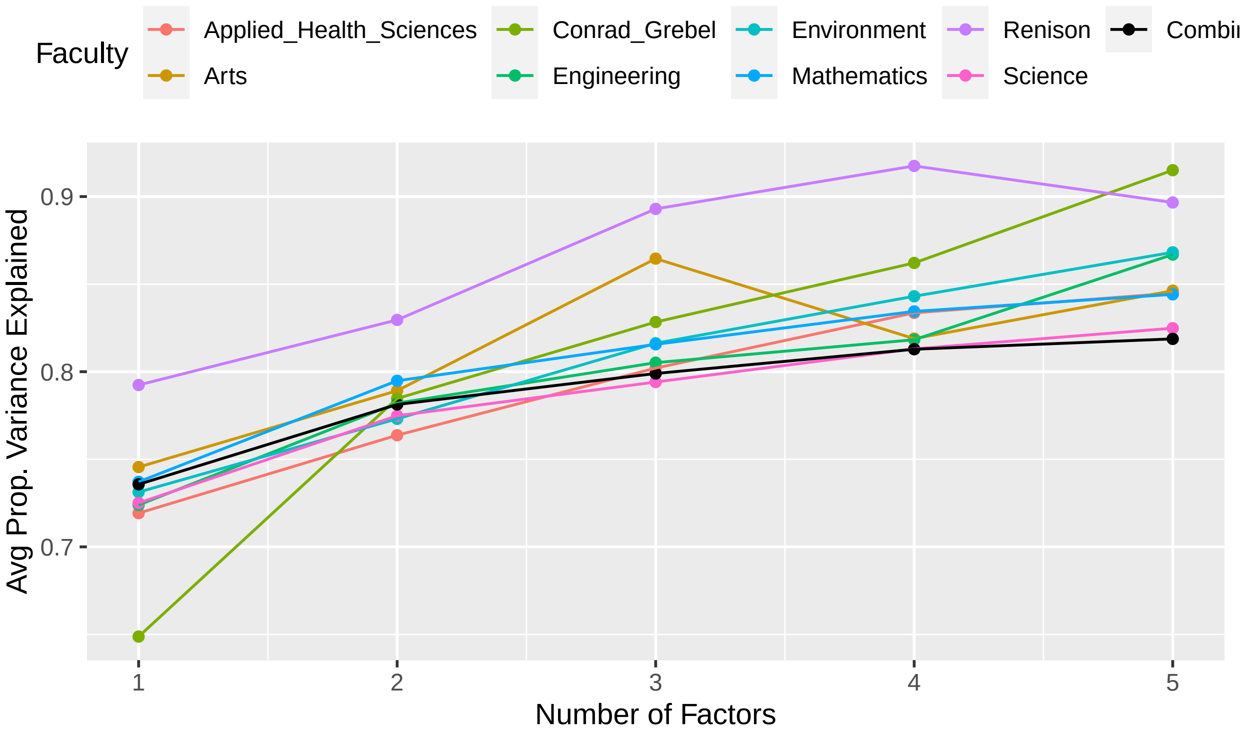


Figure 30: Average proportion of variance explained by latent factors, as a function of the number of factors in the (Course-level) model.

structure by extracting the rotated loading matrix.

Table 5: Factor Loadings. Faculty: Applied Health Sciences

Response_Item	Faculty	F1	F2	F3
Identified_LO	Applied_Health_Sciences	0.00	0.97	-0.06
LO_Assessed	Applied_Health_Sciences	0.09	0.69	0.19
Course_Activities	Applied_Health_Sciences	0.34	0.40	0.25
Return_Grades	Applied_Health_Sciences	0.00	0.00	0.82
Concepts_Conveyed	Applied_Health_Sciences	0.93	0.01	0.01
Learning_Environment	Applied_Health_Sciences	0.89	0.05	0.00
Stimulated_Interest	Applied_Health_Sciences	1.03	-0.09	-0.04
Amount_Learned	Applied_Health_Sciences	0.92	0.04	-0.02
Learning_Experience	Applied_Health_Sciences	0.84	0.07	0.10

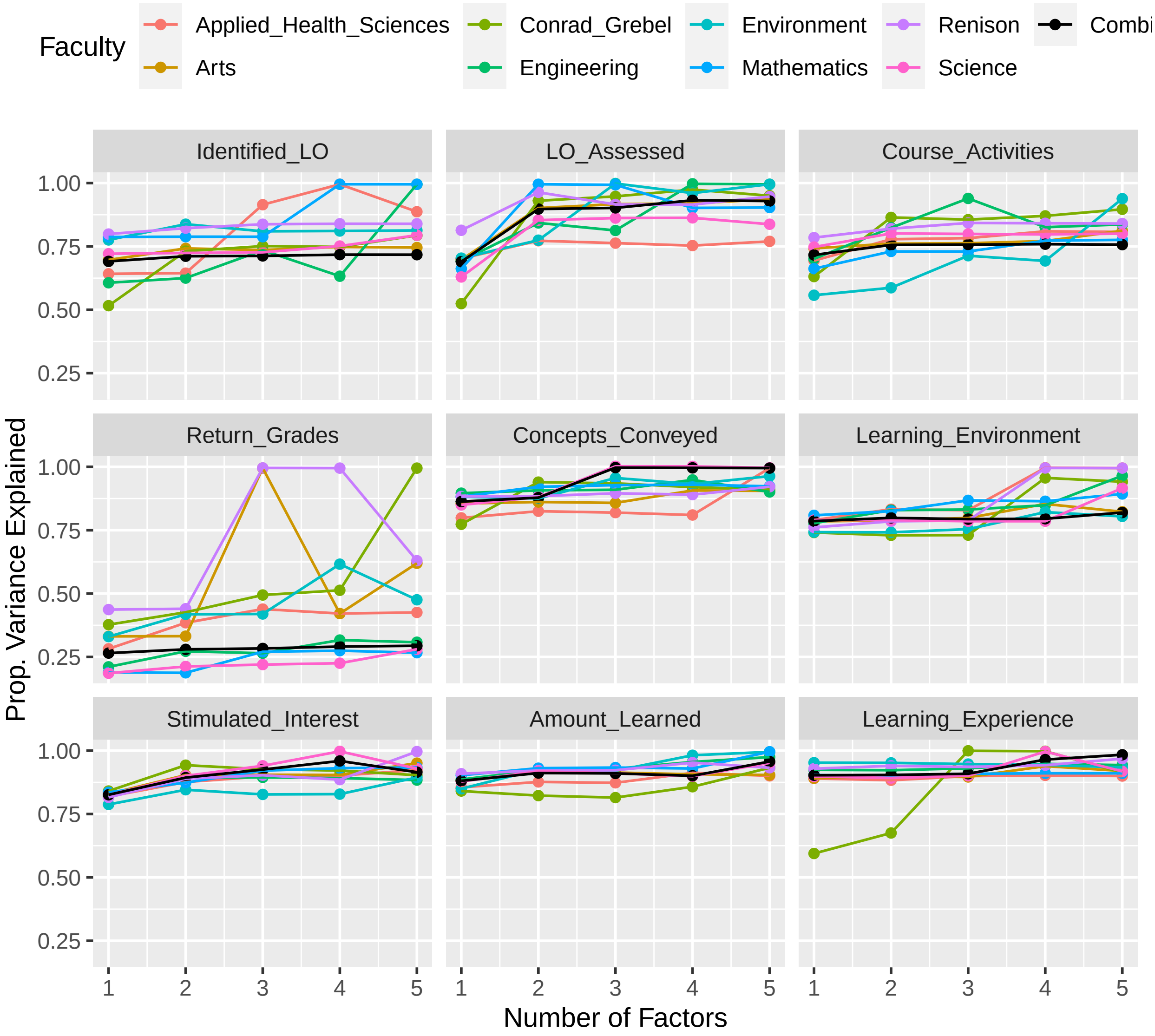


Figure 31: Proportion of variance of each response variable explained by latent factors, as a function of the number of factors in the (Course-level) model.

Table 6: Factor Loadings. Faculty: Arts

Response_Item	Faculty	F1	F2	F3
Identified_LO	Arts	0.02	0.90	-0.02
LO_Assessed	Arts	-0.05	0.94	0.04
Course_Activities	Arts	0.20	0.70	0.03
Return_Grades	Arts	0.00	-0.01	0.83
Concepts_Conveyed	Arts	0.91	0.01	0.03
Learning_Environment	Arts	0.86	0.08	0.00
Stimulated_Interest	Arts	1.04	-0.10	-0.02
Amount_Learned	Arts	0.88	0.07	0.02
Learning_Experience	Arts	0.85	0.08	0.03

Table 7: Factor Loadings. Faculty: Conrad Grebel

Response_Item	Faculty	F1	F2	F3
Identified_LO	Conrad_Grebel	-0.11	0.94	0.01
LO_Assessed	Conrad_Grebel	0.25	0.77	-0.14
Course_Activities	Conrad_Grebel	0.09	0.78	0.19
Return_Grades	Conrad_Grebel	0.10	0.03	0.80
Concepts_Conveyed	Conrad_Grebel	0.90	-0.02	0.10
Learning_Environment	Conrad_Grebel	0.79	0.12	0.07
Stimulated_Interest	Conrad_Grebel	0.95	-0.03	0.04
Amount_Learned	Conrad_Grebel	0.88	0.11	-0.01
Learning_Experience	Conrad_Grebel	1.01	-0.05	-0.05

Table 8: Factor Loadings. Faculty: Engineering

Response_Item	Faculty	F1	F2	F3
Identified_LO	Engineering	0.25	0.70	-0.04
LO_Assessed	Engineering	-0.05	0.97	0.02
Course_Activities	Engineering	0.04	0.87	0.05
Return_Grades	Engineering	0.01	0.00	0.89
Concepts_Conveyed	Engineering	0.87	0.10	-0.01
Learning_Environment	Engineering	1.00	-0.08	0.02
Stimulated_Interest	Engineering	1.01	-0.08	0.00
Amount_Learned	Engineering	0.84	0.13	0.00
Learning_Experience	Engineering	0.82	0.14	0.05

Table 9: Factor Loadings. Faculty: Environment

Response_Item	Faculty	F1	F2	F3
Identified_LO	Environment	0.40	0.36	0.25
LO_Assessed	Environment	-0.01	0.88	0.07
Course_Activities	Environment	0.03	0.95	-0.04
Return_Grades	Environment	0.03	0.01	0.72
Concepts_Conveyed	Environment	0.90	-0.01	0.09
Learning_Environment	Environment	0.92	0.04	-0.03
Stimulated_Interest	Environment	1.05	-0.07	-0.06
Amount_Learned	Environment	0.91	0.00	0.05
Learning_Experience	Environment	0.82	0.14	0.01

Table 10: Factor Loadings. Faculty: Mathematics

Response_Item	Faculty	F1	F2	F3
Identified_LO	Mathematics	0.61	0.35	0.00
LO_Assessed	Mathematics	0.01	0.87	0.04
Course_Activities	Mathematics	0.05	0.86	0.02
Return_Grades	Mathematics	0.00	0.02	0.87
Concepts_Conveyed	Mathematics	0.93	-0.01	0.05
Learning_Environment	Mathematics	0.99	-0.12	0.06
Stimulated_Interest	Mathematics	0.97	-0.02	-0.06
Amount_Learned	Mathematics	0.89	0.07	0.00
Learning_Experience	Mathematics	0.82	0.16	0.00

Table 11: Factor Loadings. Faculty: Renison

Response_Item	Faculty	F1	F2	F3
Identified_LO	Renison	0.92	0.22	-0.03
LO_Assessed	Renison	0.72	0.22	0.22
Course_Activities	Renison	0.80	0.23	0.10
Return_Grades	Renison	0.10	0.00	0.70
Concepts_Conveyed	Renison	0.98	-0.06	-0.06
Learning_Environment	Renison	0.67	-0.29	0.26
Stimulated_Interest	Renison	0.96	-0.16	-0.07
Amount_Learned	Renison	0.92	-0.05	0.03
Learning_Experience	Renison	0.82	-0.06	0.16

Table 12: Factor Loadings. Faculty: Science

Response_Item	Faculty	F1	F2	F3
Identified_LO	Science	0.24	0.73	-0.02
LO_Assessed	Science	-0.13	0.97	0.06
Course_Activities	Science	0.14	0.83	-0.02
Return_Grades	Science	0.03	0.00	0.84
Concepts_Conveyed	Science	0.87	0.08	0.03
Learning_Environment	Science	0.96	-0.04	0.01
Stimulated_Interest	Science	0.99	-0.03	-0.05
Amount_Learned	Science	0.90	0.03	0.04
Learning_Experience	Science	0.87	0.07	0.06

Table 13: Factor Loadings. Faculty: Combined

Response_Item	Faculty	F1	F2	F3
Identified_LO	Combined	0.26	0.67	-0.02
LO_Assessed	Combined	-0.08	0.97	0.03
Course_Activities	Combined	0.11	0.80	0.02
Return_Grades	Combined	0.01	0.00	0.86
Concepts_Conveyed	Combined	0.90	0.04	0.02
Learning_Environment	Combined	0.94	-0.03	0.02
Stimulated_Interest	Combined	1.00	-0.07	-0.02
Amount_Learned	Combined	0.88	0.07	0.01
Learning_Experience	Combined	0.83	0.12	0.03

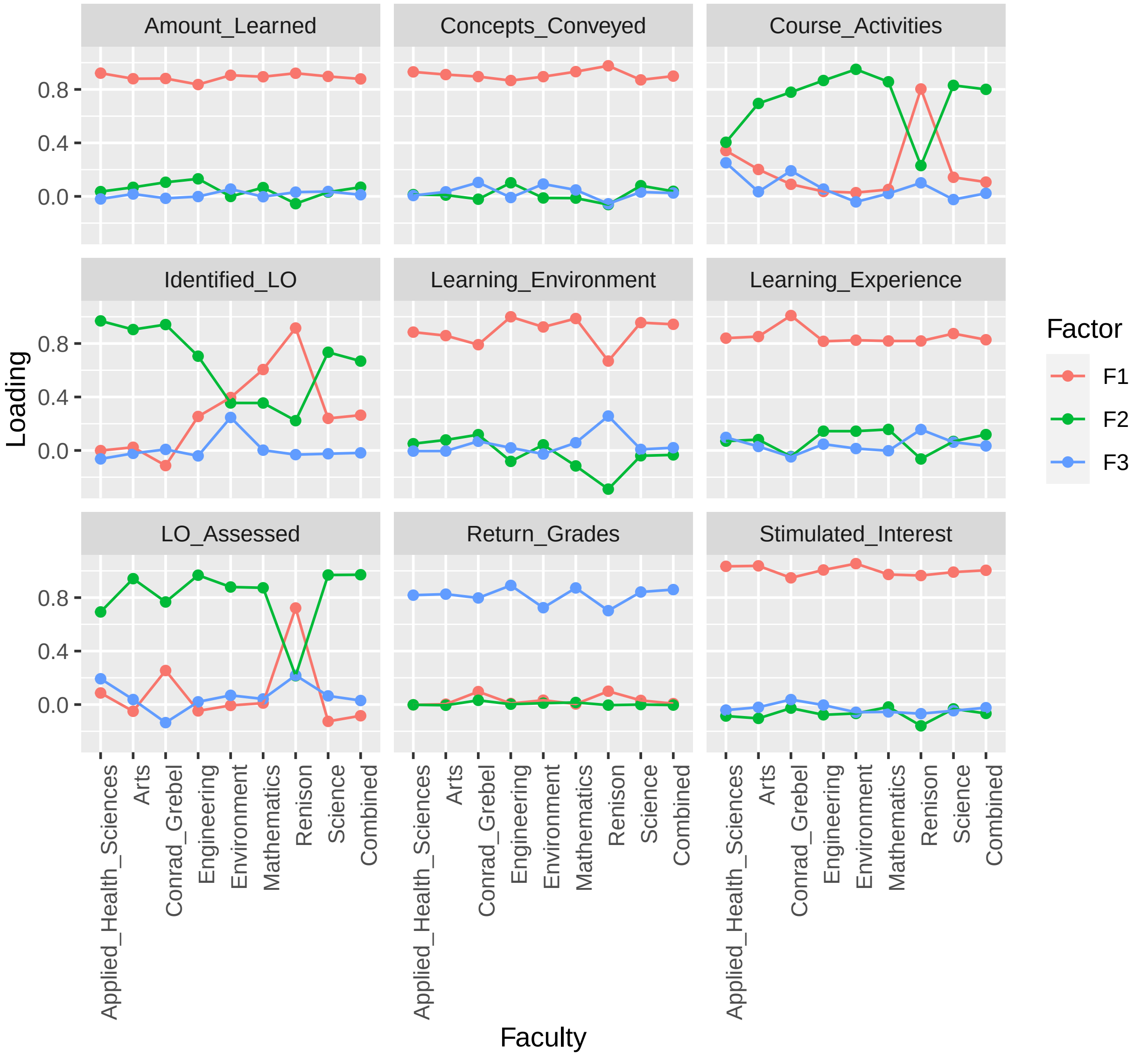


Figure 32: Factor loadings for each response variable. By choosing the order of factors, we make them as consistent as possible across all faculties.

A.5 Factor Analysis Results

Tables 5-12 show the factor loadings for each item by Faculty. A close look at these tables reveals the following key findings. For four of the six Faculties (AHS, Arts, Engineering, Science) and one of the affiliated colleges (Conrad Grebel) we see items loading on the same factors. Specifically, the following three items load heavily on factor two: `Identified_LO`, `Lo_Assessed` and `Course_Activities`. In those Faculties and at Grebel, five of the SCP items load heavily on factor one including: `Concepts_Conveyed`, `Learning_Environment`, `Stimulated_Interest`, `Amount_Learned`, `Learning_Experience`. The only item that loads heavily on factor three is `Return_Grades`. Table 13 is a summary table of the factor loadings for all Faculties and AFIW combined. This combined table displays the same combination of loadings on the three factors.

Environment, Math and Renison have slightly different factor loadings. The results for Environment and Math agreement with the larger group in having the same factors (`Concepts_Conveyed`, `Learning_Environment`, `Stimulated_Interest`, `Amount_Learned`, `Learning_Experience`) loading on factor one, with the difference that an additional item, `Identified_LO`, also loads on factor one. As with the earlier group, `LO_Assessed`, and `Course_Activities` load heavily on factor two, and again `Return_Grades` is the sole item loading on factor three.

Finally, the results for Renison given in Table 11 that it is only by force that one arrives at a three factor structure. All of the items aside from `Return_Grades` load heavily on Factor 1. The Renison results agree in treating `Return_Grades` as something of an outlier. If one forces three factors on the numbers, Factor 2 again appears to group `Identified_LO`, `LO_Assessed`, and `Course_Activities` on Factor 2, though `LO_Assessed` is equally correlated with `Return_Grades.` The most natural conclusion is that the Renison results suggest that students are not distinguishing Factors 1 and 2 in the way students are elsewhere at Waterloo, though the sample size is, of course, smaller for Renison than for other units.

The factor analysis findings are somewhat consistent with our theoretically predicted conceptual framework for organizing the SCP survey items.

In Section 4 we described the expected clustering of items into three dimensions of teaching effectiveness, both as informed by the work of CEPT1 and by our focus group research: course delivery, course design and learning experience. `Identified_LO`, `Course_Activities`, `LO_Assessed` and workload were designed to cluster together and to indicate perceptions relevant to course design. `Amount_Learned` and `Learning_Experience`, along with two open ended questions, were expected to measure perceptions of a learning experience dimension. The remaining dimensions were expected to cluster together and to indicate perceptions of the Course Delivery dimension. As a result of the focus group, the workload question was removed from our hoped for clustering for the course delivery dimension, and was instead treated as an explanatory variable.

The factor analysis reveals strong evidence for only two factors. The three items expected to cluster together to indicate Course Delivery did indeed cluster together as one of the factors. It was expected that the items `Amount_Learned`, `Concepts_Conveyed` and `Stimulated_Interest` would cluster together (as Course Delivery), and indeed they did. However, the items `Amount_Learned` and `Learning_Experience` were expected to cluster with each other (which they did), but to be distinguished from Course Delivery (i.e. to indicate perceptions of Learning Experience). Moreover, `Returned_Grades` was expected to form part of the Course Delivery factor. What the factor analysis suggests, instead, is that the `Return_Grades` item does not in fact cluster with any of the other questions. It also suggests that the questions do not in fact distinguish between learning experience and course delivery factors, but that the items (other than Returned_Grades) intended to measure those dimensions seem instead to be measuring a single underlying factor. We suggest the name ''Implementation'' for the factor that includes both course delivery and learning experience items (aside from Returned_Grades).

B Relationships Between Explanatory Variables and Composite Measure Scores

We carried out a course-level analysis of the relationships between the explanatory variables on the composite measures, parallel to the one in Section 8 for the individual items, in order to ensure that no unanticipated, problematic behaviour was observed. We saw nothing surprising, so do not include those results here in order to keep this already long report manageable. CEPT can provide the results to those who are interested upon request.

C Bivariate Regression Plots by Gender

Again in order to check that the results reported above were not hiding problematic results that could be detected with other analyses, we considered the relationship between the course-level average score and each individual explanatory factor by using simple linear regression separately for female and male instructors. In order to save space, we omit the detailed results here.

D Student-Level Regressions

For each response item, the detailed regression result is given in following tables.

Table 14: Student-level regression result of response item Identified LO.

term	estimate	std.error	statistic	p.value
(Intercept)	3.97	0.03	124.64	0.00
Instructor_GenderFemale	-0.02	0.01	-2.17	0.03
Student_GenderFemale	-0.04	0.01	-4.50	0.00
Student_GenderIdentified_Other	-0.09	0.04	-2.37	0.02
Student_GenderPrefer_No_Ans	-0.19	0.03	-6.78	0.00
WorkloadVery_Low	-0.42	0.04	-9.37	0.00
WorkloadLow	-0.12	0.02	-6.95	0.00
WorkloadHigh	0.02	0.01	1.45	0.15
WorkloadVery_High	-0.14	0.02	-8.32	0.00
Attendance_AmountAlmost_Never	-0.43	0.05	-8.77	0.00
Attendance_AmountLess_Half	-0.17	0.04	-4.11	0.00
Attendance_AmountMore_Half	0.12	0.03	4.44	0.00
Attendance_AmountAlmost_Always	0.22	0.02	9.16	0.00
Expected_Grade<60	-0.41	0.03	-13.11	0.00
Expected_Grade60-69	-0.16	0.02	-8.52	0.00
Expected_Grade80-89	0.17	0.01	14.80	0.00
Expected_Grade90+	0.24	0.01	16.47	0.00
Course_TypeElective	0.04	0.01	3.71	0.00
FacultyArts	0.03	0.02	1.35	0.18
FacultyConrad_Grebel	0.23	0.05	4.66	0.00
FacultyEngineering	-0.04	0.02	-2.05	0.04
FacultyEnvironment	0.07	0.02	2.75	0.01
FacultyMathematics	0.02	0.02	1.12	0.26
FacultyRenison	0.02	0.03	0.69	0.49
FacultyScience	-0.02	0.02	-0.83	0.41
Class_Size1-25	0.03	0.02	1.53	0.13
Class_Size26-50	-0.03	0.02	-1.98	0.05
Class_Size101-200	0.01	0.01	1.02	0.31
Class_Size200+	0.01	0.02	0.53	0.59
Attendance_TypeOnline	-0.02	0.03	-0.59	0.56

Table 15: Student-level regression result of response item LO Assessed.

term	estimate	std.error	statistic	p.value
(Intercept)	3.75	0.03	119.04	0.00
Instructor_GenderFemale	-0.04	0.01	-4.01	0.00
Student_GenderFemale	-0.01	0.01	-0.77	0.44
Student_GenderIdentified_Other	-0.06	0.04	-1.58	0.11
Student_GenderPrefer_No_Ans	-0.16	0.03	-5.64	0.00
WorkloadVery_Low	-0.48	0.04	-10.83	0.00
WorkloadLow	-0.15	0.02	-8.70	0.00
WorkloadHigh	-0.01	0.01	-1.24	0.21
WorkloadVery_High	-0.21	0.02	-12.94	0.00
Attendance_AmountAlmost_Never	-0.26	0.05	-5.42	0.00
Attendance_AmountLess_Half	-0.12	0.04	-2.86	0.00
Attendance_AmountMore_Half	0.10	0.03	3.87	0.00
Attendance_AmountAlmost_Always	0.19	0.02	8.08	0.00
Expected_Grade<60	-0.45	0.03	-14.43	0.00
Expected_Grade60-69	-0.23	0.02	-12.15	0.00
Expected_Grade80-89	0.22	0.01	19.93	0.00
Expected_Grade90+	0.34	0.01	23.35	0.00
Course_TypeElective	0.02	0.01	1.35	0.18
FacultyArts	0.12	0.02	6.52	0.00
FacultyConrad_Grebel	0.31	0.05	6.48	0.00
FacultyEngineering	0.07	0.02	3.62	0.00
FacultyEnvironment	0.08	0.02	3.37	0.00
FacultyMathematics	0.21	0.02	10.26	0.00
FacultyRenison	0.12	0.03	3.70	0.00
FacultyScience	0.10	0.02	5.17	0.00
Class_Size1-25	0.06	0.02	3.83	0.00
Class_Size26-50	0.00	0.02	-0.13	0.90
Class_Size101-200	0.02	0.01	1.66	0.10
Class_Size200+	0.01	0.02	0.44	0.66
Attendance_TypeOnline	-0.10	0.03	-3.91	0.00

Table 16: Student-level regression result of response item Course Activities.

term	estimate	std.error	statistic	p.value
(Intercept)	3.50	0.04	96.02	0.00
Instructor_GenderFemale	-0.05	0.01	-4.23	0.00
Student_GenderFemale	-0.04	0.01	-3.90	0.00
Student_GenderIdentified_Other	-0.02	0.04	-0.46	0.64
Student_GenderPrefer_No_Ans	-0.23	0.03	-7.15	0.00
WorkloadVery_Low	-0.46	0.05	-8.99	0.00
WorkloadLow	-0.19	0.02	-9.17	0.00
WorkloadHigh	-0.06	0.01	-5.01	0.00
WorkloadVery_High	-0.34	0.02	-17.98	0.00
Attendance_AmountAlmost_Never	-0.33	0.06	-5.81	0.00
Attendance_AmountLess_Half	-0.10	0.05	-2.12	0.03
Attendance_AmountMore_Half	0.15	0.03	4.96	0.00
Attendance_AmountAlmost_Always	0.25	0.03	9.17	0.00
Expected_Grade<60	-0.81	0.04	-22.46	0.00
Expected_Grade60-69	-0.35	0.02	-16.11	0.00
Expected_Grade80-89	0.33	0.01	25.46	0.00
Expected_Grade90+	0.51	0.02	30.14	0.00
Course_TypeElective	0.02	0.01	1.15	0.25
FacultyArts	0.18	0.02	8.01	0.00
FacultyConrad_Grebel	0.43	0.06	7.85	0.00
FacultyEngineering	0.14	0.02	6.20	0.00
FacultyEnvironment	0.09	0.03	3.12	0.00
FacultyMathematics	0.24	0.02	10.20	0.00
FacultyRenison	0.27	0.04	7.34	0.00
FacultyScience	0.16	0.02	6.80	0.00
Class_Size1-25	0.03	0.02	1.59	0.11
Class_Size26-50	0.00	0.02	-0.25	0.80
Class_Size101-200	0.03	0.01	1.92	0.05
Class_Size200+	-0.02	0.02	-0.93	0.35
Attendance_TypeOnline	-0.14	0.03	-4.78	0.00

Table 17: Student-level regression result of response item Return Grades.

term	estimate	std.error	statistic	p.value
(Intercept)	3.64	0.04	99.71	0.00
Instructor_GenderFemale	0.01	0.01	0.83	0.41
Student_GenderFemale	-0.01	0.01	-0.83	0.41
Student_GenderIdentified_Other	-0.10	0.04	-2.30	0.02
Student_GenderPrefer_No_Ans	-0.20	0.03	-6.15	0.00
WorkloadVery_Low	-0.34	0.05	-6.59	0.00
WorkloadLow	-0.08	0.02	-4.12	0.00
WorkloadHigh	-0.07	0.01	-5.44	0.00
WorkloadVery_High	-0.25	0.02	-13.06	0.00
Attendance_AmountAlmost_Never	-0.16	0.06	-2.80	0.01
Attendance_AmountLess_Half	-0.06	0.05	-1.34	0.18
Attendance_AmountMore_Half	0.16	0.03	5.27	0.00
Attendance_AmountAlmost_Always	0.20	0.03	7.21	0.00
Expected_Grade<60	-0.24	0.04	-6.65	0.00
Expected_Grade60-69	-0.08	0.02	-3.59	0.00
Expected_Grade80-89	0.15	0.01	11.35	0.00
Expected_Grade90+	0.22	0.02	12.93	0.00
Course_TypeElective	0.05	0.01	4.12	0.00
FacultyArts	0.18	0.02	8.32	0.00
FacultyConrad_Grebel	0.38	0.06	6.88	0.00
FacultyEngineering	0.08	0.02	3.55	0.00
FacultyEnvironment	-0.09	0.03	-3.13	0.00
FacultyMathematics	0.21	0.02	9.00	0.00
FacultyRenison	0.23	0.04	6.42	0.00
FacultyScience	0.31	0.02	13.25	0.00
Class_Size1-25	0.13	0.02	6.92	0.00
Class_Size26-50	0.09	0.02	4.56	0.00
Class_Size101-200	0.06	0.01	4.29	0.00
Class_Size200+	0.08	0.02	4.43	0.00
Attendance_TypeOnline	-0.17	0.03	-5.67	0.00

Table 18: Student-level regression result of response item Concepts Conveyed.

term	estimate	std.error	statistic	p.value
(Intercept)	3.58	0.03	103.75	0.00
Instructor_GenderFemale	-0.06	0.01	-5.35	0.00
Student_GenderFemale	-0.04	0.01	-3.44	0.00
Student_GenderIdentified_Other	-0.05	0.04	-1.29	0.20
Student_GenderPrefer_No_Ans	-0.21	0.03	-6.88	0.00
WorkloadVery_Low	-0.38	0.05	-7.78	0.00
WorkloadLow	-0.13	0.02	-6.66	0.00
WorkloadHigh	0.00	0.01	-0.07	0.94
WorkloadVery_High	-0.23	0.02	-13.11	0.00
Attendance_AmountAlmost_Never	-0.65	0.05	-12.10	0.00
Attendance_AmountLess_Half	-0.29	0.04	-6.43	0.00
Attendance_AmountMore_Half	0.20	0.03	6.67	0.00
Attendance_AmountAlmost_Always	0.36	0.03	13.93	0.00
Expected_Grade<60	-0.78	0.03	-22.97	0.00
Expected_Grade60-69	-0.30	0.02	-14.53	0.00
Expected_Grade80-89	0.25	0.01	20.43	0.00
Expected_Grade90+	0.37	0.02	23.21	0.00
Course_TypeElective	0.08	0.01	6.27	0.00
FacultyArts	0.17	0.02	7.89	0.00
FacultyConrad_Grebel	0.46	0.05	8.79	0.00
FacultyEngineering	0.10	0.02	4.86	0.00
FacultyEnvironment	0.14	0.03	5.29	0.00
FacultyMathematics	0.28	0.02	12.89	0.00
FacultyRenison	0.22	0.03	6.34	0.00
FacultyScience	0.14	0.02	6.18	0.00
Class_Size1-25	0.06	0.02	3.34	0.00
Class_Size26-50	0.02	0.02	0.91	0.36
Class_Size101-200	-0.02	0.01	-1.52	0.13
Class_Size200+	-0.06	0.02	-3.58	0.00
Attendance_TypeOnline	-0.27	0.03	-9.68	0.00

Table 19: Student-level regression result of response item Learning Environment.

term	estimate	std.error	statistic	p.value
(Intercept)	3.68	0.03	109.35	0.00
Instructor_GenderFemale	-0.04	0.01	-3.65	0.00
Student_GenderFemale	-0.01	0.01	-0.51	0.61
Student_GenderIdentified_Other	-0.12	0.04	-3.11	0.00
Student_GenderPrefer_No_Ans	-0.24	0.03	-7.87	0.00
WorkloadVery_Low	-0.36	0.05	-7.48	0.00
WorkloadLow	-0.08	0.02	-4.30	0.00
WorkloadHigh	-0.04	0.01	-3.22	0.00
WorkloadVery_High	-0.24	0.02	-14.06	0.00
Attendance_AmountAlmost_Never	-0.47	0.05	-8.94	0.00
Attendance_AmountLess_Half	-0.18	0.04	-4.17	0.00
Attendance_AmountMore_Half	0.25	0.03	8.63	0.00
Attendance_AmountAlmost_Always	0.36	0.03	14.25	0.00
Expected_Grade<60	-0.62	0.03	-18.76	0.00
Expected_Grade60-69	-0.27	0.02	-13.56	0.00
Expected_Grade80-89	0.21	0.01	17.64	0.00
Expected_Grade90+	0.30	0.02	19.50	0.00
Course_TypeElective	0.05	0.01	4.06	0.00
FacultyArts	0.13	0.02	6.46	0.00
FacultyConrad_Grebel	0.40	0.05	7.89	0.00
FacultyEngineering	0.05	0.02	2.33	0.02
FacultyEnvironment	0.08	0.03	3.22	0.00
FacultyMathematics	0.20	0.02	9.43	0.00
FacultyRenison	0.14	0.03	4.15	0.00
FacultyScience	0.08	0.02	3.55	0.00
Class_Size1-25	0.11	0.02	6.23	0.00
Class_Size26-50	0.05	0.02	2.63	0.01
Class_Size101-200	0.00	0.01	0.38	0.70
Class_Size200+	-0.09	0.02	-5.11	0.00
Attendance_TypeOnline	-0.27	0.03	-9.82	0.00

Table 20: Student-level regression result of response item Stimulated Interest.

term	estimate	std.error	statistic	p.value
(Intercept)	3.24	0.04	81.11	0.00
Instructor_GenderFemale	-0.12	0.01	-9.16	0.00
Student_GenderFemale	-0.06	0.01	-4.85	0.00
Student_GenderIdentified_Other	-0.04	0.05	-0.78	0.43
Student_GenderPrefer_No_Ans	-0.21	0.04	-5.88	0.00
WorkloadVery_Low	-0.54	0.06	-9.57	0.00
WorkloadLow	-0.21	0.02	-9.42	0.00
WorkloadHigh	0.03	0.01	2.32	0.02
WorkloadVery_High	-0.18	0.02	-8.51	0.00
Attendance_AmountAlmost_Never	-0.59	0.06	-9.50	0.00
Attendance_AmountLess_Half	-0.27	0.05	-5.20	0.00
Attendance_AmountMore_Half	0.29	0.03	8.45	0.00
Attendance_AmountAlmost_Always	0.50	0.03	16.62	0.00
Expected_Grade<60	-0.75	0.04	-19.06	0.00
Expected_Grade60-69	-0.35	0.02	-14.80	0.00
Expected_Grade80-89	0.28	0.01	20.02	0.00
Expected_Grade90+	0.42	0.02	23.09	0.00
Course_TypeElective	0.16	0.01	10.62	0.00
FacultyArts	0.15	0.02	6.21	0.00
FacultyConrad_Grebel	0.56	0.06	9.31	0.00
FacultyEngineering	0.07	0.03	2.85	0.00
FacultyEnvironment	0.14	0.03	4.52	0.00
FacultyMathematics	0.24	0.03	9.25	0.00
FacultyRenison	0.22	0.04	5.49	0.00
FacultyScience	0.17	0.03	6.79	0.00
Class_Size1-25	0.02	0.02	1.07	0.29
Class_Size26-50	0.02	0.02	0.90	0.37
Class_Size101-200	-0.01	0.02	-0.66	0.51
Class_Size200+	-0.09	0.02	-4.29	0.00
Attendance_TypeOnline	-0.23	0.03	-6.88	0.00

Table 21: Student-level regression result of response item Amount Learned.

term	estimate	std.error	statistic	p.value
(Intercept)	3.39	0.04	93.01	0.00
Instructor_GenderFemale	-0.09	0.01	-7.41	0.00
Student_GenderFemale	-0.03	0.01	-3.00	0.00
Student_GenderIdentified_Other	-0.08	0.04	-1.88	0.06
Student_GenderPrefer_No_Ans	-0.21	0.03	-6.51	0.00
WorkloadVery_Low	-0.70	0.05	-13.65	0.00
WorkloadLow	-0.27	0.02	-13.15	0.00
WorkloadHigh	0.07	0.01	5.16	0.00
WorkloadVery_High	-0.14	0.02	-7.44	0.00
Attendance_AmountAlmost_Never	-0.64	0.06	-11.35	0.00
Attendance_AmountLess_Half	-0.28	0.05	-6.00	0.00
Attendance_AmountMore_Half	0.31	0.03	10.02	0.00
Attendance_AmountAlmost_Always	0.51	0.03	18.89	0.00
Expected_Grade<60	-0.79	0.04	-22.17	0.00
Expected_Grade60-69	-0.31	0.02	-14.59	0.00
Expected_Grade80-89	0.22	0.01	17.62	0.00
Expected_Grade90+	0.33	0.02	19.80	0.00
Course_TypeElective	0.10	0.01	7.29	0.00
FacultyArts	0.17	0.02	7.70	0.00
FacultyConrad_Grebel	0.50	0.06	8.99	0.00
FacultyEngineering	0.09	0.02	4.16	0.00
FacultyEnvironment	0.09	0.03	3.13	0.00
FacultyMathematics	0.31	0.02	13.59	0.00
FacultyRenison	0.25	0.04	6.86	0.00
FacultyScience	0.15	0.02	6.60	0.00
Class_Size1-25	-0.01	0.02	-0.41	0.68
Class_Size26-50	-0.02	0.02	-0.80	0.42
Class_Size101-200	0.01	0.01	0.43	0.67
Class_Size200+	-0.04	0.02	-2.02	0.04
Attendance_TypeOnline	-0.28	0.03	-9.29	0.00

Table 22: Student-level regression result of response item Learning Experience.

term	estimate	std.error	statistic	p.value
(Intercept)	3.25	0.04	87.33	0.00
Instructor_GenderFemale	-0.09	0.01	-8.00	0.00
Student_GenderFemale	-0.05	0.01	-4.15	0.00
Student_GenderIdentified_Other	-0.06	0.04	-1.35	0.18
Student_GenderPrefer_No_Ans	-0.25	0.03	-7.39	0.00
WorkloadVery_Low	-0.58	0.05	-11.07	0.00
WorkloadLow	-0.20	0.02	-9.79	0.00
WorkloadHigh	-0.01	0.01	-0.63	0.53
WorkloadVery_High	-0.31	0.02	-16.28	0.00
Attendance_AmountAlmost_Never	-0.53	0.06	-9.12	0.00
Attendance_AmountLess_Half	-0.22	0.05	-4.56	0.00
Attendance_AmountMore_Half	0.27	0.03	8.47	0.00
Attendance_AmountAlmost_Always	0.45	0.03	16.13	0.00
Expected_Grade<60	-0.90	0.04	-24.58	0.00
Expected_Grade60-69	-0.37	0.02	-16.93	0.00
Expected_Grade80-89	0.29	0.01	22.40	0.00
Expected_Grade90+	0.46	0.02	26.97	0.00
Course_TypeElective	0.10	0.01	7.51	0.00
FacultyArts	0.19	0.02	8.17	0.00
FacultyConrad_Grebel	0.54	0.06	9.53	0.00
FacultyEngineering	0.15	0.02	6.46	0.00
FacultyEnvironment	0.15	0.03	5.05	0.00
FacultyMathematics	0.35	0.02	14.76	0.00
FacultyRenison	0.28	0.04	7.47	0.00
FacultyScience	0.19	0.02	7.99	0.00
Class_Size1-25	0.05	0.02	2.44	0.01
Class_Size26-50	0.01	0.02	0.52	0.61
Class_Size101-200	0.01	0.01	0.79	0.43
Class_Size200+	-0.01	0.02	-0.58	0.57
Attendance_TypeOnline	-0.21	0.03	-6.78	0.00

E Course-Level Regressions

For each response item, the detailed regression result on the course level is given respectively in following tables.

Table 23: Course-level regression result of response item Identified LO.

term	estimate	std.error	statistic	p.value
(Intercept)	4.16	0.06	70.89	0.00
Instructor_Gender	-0.02	0.02	-1.12	0.26
Student_Gender	-0.13	0.04	-2.87	0.00
Workload	-0.07	0.04	-1.96	0.05
Attendance_Amount	0.78	0.08	10.33	0.00
Expected_Grade	0.01	0.00	10.69	0.00
Course_Type	-0.06	0.03	-2.12	0.03
FacultyArts	0.06	0.04	1.71	0.09
FacultyConrad_Grebel	0.24	0.09	2.62	0.01
FacultyEngineering	0.04	0.04	0.96	0.34
FacultyEnvironment	0.19	0.05	4.11	0.00
FacultyMathematics	0.07	0.04	1.81	0.07
FacultyRenison	0.04	0.06	0.66	0.51
FacultyScience	0.04	0.04	0.94	0.35
Class_Size1-25	0.02	0.03	0.72	0.47
Class_Size26-50	-0.02	0.03	-0.81	0.42
Class_Size101-200	0.02	0.02	0.73	0.46
Class_Size200+	0.01	0.03	0.42	0.67
Attendance_Type	0.02	0.04	0.41	0.68

Table 24: Course-level regression result of response item LO Assessed.

term	estimate	std.error	statistic	p.value
(Intercept)	4.02	0.06	69.81	0.00
Instructor_Gender	-0.03	0.02	-1.57	0.12
Student_Gender	-0.15	0.04	-3.51	0.00
Workload	-0.12	0.04	-3.37	0.00
Attendance_Amount	0.69	0.07	9.35	0.00
Expected_Grade	0.02	0.00	13.30	0.00
Course_Type	0.01	0.03	0.20	0.84
FacultyArts	0.17	0.04	4.66	0.00
FacultyConrad_Grebel	0.33	0.09	3.77	0.00
FacultyEngineering	0.15	0.04	3.76	0.00
FacultyEnvironment	0.22	0.05	4.83	0.00
FacultyMathematics	0.27	0.04	6.88	0.00
FacultyRenison	0.12	0.06	1.96	0.05
FacultyScience	0.15	0.04	3.98	0.00
Class_Size1-25	0.07	0.03	2.16	0.03
Class_Size26-50	0.01	0.03	0.21	0.84
Class_Size101-200	0.02	0.02	0.84	0.40
Class_Size200+	0.01	0.03	0.37	0.71
Attendance_Type	-0.06	0.04	-1.54	0.12

Table 25: Course-level regression result of response item Course Activities.

term	estimate	std.error	statistic	p.value
(Intercept)	3.85	0.07	52.87	0.00
Instructor_Gender	-0.04	0.02	-1.93	0.05
Student_Gender	-0.16	0.05	-2.88	0.00
Workload	-0.23	0.04	-5.02	0.00
Attendance_Amount	0.95	0.09	10.20	0.00
Expected_Grade	0.03	0.00	15.66	0.00
Course_Type	0.00	0.04	-0.05	0.96
FacultyArts	0.22	0.05	4.94	0.00
FacultyConrad_Grebel	0.45	0.11	4.02	0.00
FacultyEngineering	0.23	0.05	4.45	0.00
FacultyEnvironment	0.26	0.06	4.41	0.00
FacultyMathematics	0.30	0.05	6.06	0.00
FacultyRenison	0.26	0.08	3.40	0.00
FacultyScience	0.21	0.05	4.52	0.00
Class_Size1-25	0.03	0.04	0.82	0.41
Class_Size26-50	-0.01	0.04	-0.19	0.85
Class_Size101-200	0.02	0.03	0.76	0.45
Class_Size200+	-0.03	0.03	-0.84	0.40
Attendance_Type	-0.04	0.05	-0.89	0.37

Table 26: Course-level regression result of response item Return Grades.

term	estimate	std.error	statistic	p.value
(Intercept)	3.92	0.09	45.20	0.00
Instructor_Gender	0.01	0.03	0.30	0.76
Student_Gender	-0.05	0.06	-0.83	0.41
Workload	-0.35	0.05	-6.51	0.00
Attendance_Amount	0.76	0.11	6.84	0.00
Expected_Grade	0.01	0.00	5.95	0.00
Course_Type	-0.03	0.04	-0.71	0.48
FacultyArts	0.22	0.05	4.07	0.00
FacultyConrad_Grebel	0.33	0.13	2.48	0.01
FacultyEngineering	0.14	0.06	2.34	0.02
FacultyEnvironment	0.05	0.07	0.76	0.45
FacultyMathematics	0.26	0.06	4.35	0.00
FacultyRenison	0.26	0.09	2.94	0.00
FacultyScience	0.27	0.06	4.87	0.00
Class_Size1-25	0.14	0.05	2.84	0.00
Class_Size26-50	0.03	0.04	0.70	0.49
Class_Size101-200	0.03	0.03	0.94	0.35
Class_Size200+	0.12	0.04	2.85	0.00
Attendance_Type	-0.12	0.06	-2.10	0.04

Table 27: Course-level regression result of response item Concepts Conveyed.

term	estimate	std.error	statistic	p.value
(Intercept)	4.21	0.08	56.09	0.00
Instructor_Gender	-0.07	0.02	-2.89	0.00
Student_Gender	-0.12	0.06	-2.16	0.03
Workload	-0.13	0.05	-2.90	0.00
Attendance_Amount	1.29	0.10	13.36	0.00
Expected_Grade	0.02	0.00	12.03	0.00
Course_Type	-0.12	0.04	-3.30	0.00
FacultyArts	0.18	0.05	3.82	0.00
FacultyConrad_Grebel	0.43	0.12	3.67	0.00
FacultyEngineering	0.18	0.05	3.32	0.00
FacultyEnvironment	0.30	0.06	4.99	0.00
FacultyMathematics	0.29	0.05	5.78	0.00
FacultyRenison	0.24	0.08	3.10	0.00
FacultyScience	0.18	0.05	3.73	0.00
Class_Size1-25	0.04	0.04	0.93	0.35
Class_Size26-50	-0.01	0.04	-0.22	0.83
Class_Size101-200	-0.02	0.03	-0.56	0.58
Class_Size200+	-0.08	0.04	-2.19	0.03
Attendance_Type	-0.19	0.05	-3.95	0.00

Table 28: Course-level regression result of response item Learning Environment.

term	estimate	std.error	statistic	p.value
(Intercept)	4.29	0.07	58.38	0.00
Instructor_Gender	-0.06	0.02	-2.61	0.01
Student_Gender	-0.11	0.05	-2.00	0.05
Workload	-0.21	0.05	-4.58	0.00
Attendance_Amount	1.14	0.09	12.11	0.00
Expected_Grade	0.02	0.00	11.76	0.00
Course_Type	-0.08	0.04	-2.12	0.03
FacultyArts	0.15	0.05	3.35	0.00
FacultyConrad_Grebel	0.37	0.11	3.30	0.00
FacultyEngineering	0.13	0.05	2.57	0.01
FacultyEnvironment	0.25	0.06	4.20	0.00
FacultyMathematics	0.23	0.05	4.56	0.00
FacultyRenison	0.16	0.08	2.07	0.04
FacultyScience	0.13	0.05	2.76	0.01
Class_Size1-25	0.10	0.04	2.32	0.02
Class_Size26-50	0.03	0.04	0.78	0.43
Class_Size101-200	0.00	0.03	0.13	0.90
Class_Size200+	-0.09	0.03	-2.71	0.01
Attendance_Type	-0.20	0.05	-4.17	0.00

Table 29: Course-level regression result of response item Stimulated Interest.

term	estimate	std.error	statistic	p.value
(Intercept)	4.10	0.09	46.73	0.00
Instructor_Gender	-0.12	0.03	-4.43	0.00
Student_Gender	-0.07	0.07	-1.02	0.31
Workload	0.01	0.05	0.22	0.83
Attendance_Amount	1.52	0.11	13.48	0.00
Expected_Grade	0.02	0.00	11.47	0.00
Course_Type	-0.26	0.04	-6.02	0.00
FacultyArts	0.13	0.05	2.32	0.02
FacultyConrad_Grebel	0.50	0.14	3.73	0.00
FacultyEngineering	0.12	0.06	2.01	0.04
FacultyEnvironment	0.30	0.07	4.25	0.00
FacultyMathematics	0.23	0.06	3.87	0.00
FacultyRenison	0.21	0.09	2.28	0.02
FacultyScience	0.19	0.06	3.44	0.00
Class_Size1-25	0.02	0.05	0.38	0.71
Class_Size26-50	0.01	0.04	0.27	0.79
Class_Size101-200	-0.01	0.03	-0.29	0.78
Class_Size200+	-0.11	0.04	-2.52	0.01
Attendance_Type	-0.17	0.06	-2.99	0.00

Table 30: Course-level regression result of response item Amount Learned.

term	estimate	std.error	statistic	p.value
(Intercept)	4.19	0.08	51.46	0.00
Instructor_Gender	-0.08	0.03	-3.23	0.00
Student_Gender	-0.12	0.06	-1.96	0.05
Workload	0.05	0.05	1.06	0.29
Attendance_Amount	1.46	0.10	13.99	0.00
Expected_Grade	0.02	0.00	11.30	0.00
Course_Type	-0.17	0.04	-4.25	0.00
FacultyArts	0.16	0.05	3.15	0.00
FacultyConrad_Grebel	0.45	0.13	3.55	0.00
FacultyEngineering	0.14	0.06	2.46	0.01
FacultyEnvironment	0.25	0.07	3.79	0.00
FacultyMathematics	0.31	0.06	5.52	0.00
FacultyRenison	0.24	0.08	2.87	0.00
FacultyScience	0.18	0.05	3.50	0.00
Class_Size1-25	-0.01	0.05	-0.26	0.80
Class_Size26-50	-0.02	0.04	-0.48	0.63
Class_Size101-200	0.02	0.03	0.63	0.53
Class_Size200+	-0.05	0.04	-1.18	0.24
Attendance_Type	-0.20	0.05	-3.71	0.00

Table 31: Course-level regression result of response item Learning Experience.

term	estimate	std.error	statistic	p.value
(Intercept)	3.94	0.08	47.09	0.00
Instructor_Gender	-0.10	0.03	-3.84	0.00
Student_Gender	-0.13	0.06	-2.14	0.03
Workload	-0.10	0.05	-1.91	0.06
Attendance_Amount	1.40	0.11	13.10	0.00
Expected_Grade	0.03	0.00	13.49	0.00
Course_Type	-0.16	0.04	-3.70	0.00
FacultyArts	0.20	0.05	3.92	0.00
FacultyConrad_Grebel	0.50	0.13	3.89	0.00
FacultyEngineering	0.23	0.06	3.95	0.00
FacultyEnvironment	0.32	0.07	4.82	0.00
FacultyMathematics	0.36	0.06	6.41	0.00
FacultyRenison	0.28	0.09	3.28	0.00
FacultyScience	0.23	0.05	4.30	0.00
Class_Size1-25	0.04	0.05	0.92	0.36
Class_Size26-50	0.00	0.04	0.02	0.98
Class_Size101-200	0.01	0.03	0.44	0.66
Class_Size200+	-0.02	0.04	-0.56	0.57
Attendance_Type	-0.12	0.05	-2.11	0.03

F SCP Pilot Test Survey Items Fall 2018

The first nine questions were measured on a 5-point Likert scale (ranging from strongly disagree to strongly agree; there will also be an additional response-category, labelled: “have no basis for rating”).

1. The instructor identified the intended learning outcomes for this course.
2. The intended learning outcomes were assessed through my graded work.
3. The course activities prepared me for the graded work.
4. Graded work was returned in a reasonable amount of time.
5. The instructor helped me to understand the course concepts.
6. The instructor created a supportive environment that helped me learn.
7. The instructor stimulated my interest in this course.
8. Overall, I learned a great deal from this instructor.
9. Overall, the quality of my learning experience in this course was excellent
10. The course workload demands were... (scale ranging from very low to very high)

Additional questions, for analysis purposes of the pilot test data, included the following:

11. What is your gender identity? (Note that this can also include gender expression as it relates to your gender identity). (Female, Male, Non-Binary, Agender, Genderqueer, Trans Male, Trans Female, Not Listed, Prefer Not to Answer)
12. On average, I attend class. . . (Almost Never/Less than half of the time/Half of the time/More than half of the time/Almost Always)
13. In terms of an expected grade in this course, I expect to get. . . (< 30, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84, 85-89, 90-94, 95+).
14. For me, this course is (required or elective).

For online courses, the question: ‘’On average, I attend class” was replaced with: ‘’On average, I engage in the prescribed weekly online work for this course”.

References

Abrami, P. C., d’Apollonia, S., and Rosenfield, S. (2007). The Dimensionality of Student Ratings of Instruction: What We Know and What We Do Not. In *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*, pages 385–456. Springer Netherlands, Dordrecht.

Adams, M. J. and Umbach, P. D. (2012). Nonresponse and Online Student Evaluations of Teaching: Understanding the Influence of Salience, Fatigue, and Academic Environments. *Research in Higher Education*, 53(5):576–591.

Algozzine, B., Beattie, J., Bray, M., Flowers, C., Gretes, J., Howley, L., Mohanty, G., and Spooner, F. (2004). Student evaluation of college teaching: A practice in search of principles. *College Teaching*, 52:134–141.

Bain, K. (2004). *What the Best College Teachers Do*. Harvard University Press.

Basow, S. (2000). Best and worst professors: Gender patterns in students' choices. *Sex Roles*, 45:405–417.

Beleche, T., Fairris, D., and Marks, M. (2012). Do Course Evaluations Truly Reflect Student Learning? Evidence from an Objectively Graded Post-Test. *Economics of Education Review*, 31(5):709–719.

Benton, S. L. and Cashin, W. E. (2012). Student Ratings of Teaching: A Summary of the Research. IDEA Paper. *IDEA Center*, 50:1–20.

Benton, S. L. and Cashin, W. E. (2014). Student Ratings of Instruction in College and University Courses. In *Higher education: Handbook of theory and research*, pages 279–326. Springer.

Boring, A., Ottoboni, K., and Stark, P. (2016). Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness. *ScienceOpen Research*.

Braskamp, L. A. and Ory, J. C. (1994). *Assessing Faculty Work: Enhancing Individual and Institutional Performance.* Jossey-Bass Inc., San Francisco, CA.

Carrell, S. and West, J. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy*, 118(3):409–432.

Centra, J. A. (2003). Will Teachers Receive Higher Student Evaluations by Giving Higher Grades and Less Course Work? *Research in Higher Education*, 44(5):495–518.

Centra, J. A. (2009). Differences in Responses to the Student Instructional Report: Is It Bias? *Princeton, NJ: Educational Testing Service*.

Chapman, D. D. and Joines, J. A. (2017). Strategies for Increasing Response Rates for Online End-of-Course Evaluations. *International Journal of Teaching and Learning in Higher Education*, 29(1):47–60.

Chickering, A. and Gamson, Z. (1987). The Seven Principles of Effective Undergraduate Education. *AAHE Bulletin*.

Christensen Hughes, J. and Mighty, J. (2010). *Taking Stock: Research on Teaching and Learning in Higher Education*. McGill-Queen's University Press, Montreal & Kingston: Queen's Policy Studies Series.

Clayson, D. E., Frost, T. F., and Sheffet, M. J. (2006). Grades and the Student Evaluation of Instruction: A Test of the Reciprocity Effect. *Academy of Management Learning & Education*, 5(1):52–65.

Cohen, P. A. (1980). Effectiveness of Student-Rating Feedback for Improving College Instruction: A Meta-Analysis of Findings. *Research in higher education*, 13(4):321–341.

Cook, C., Heath, F., and Thompson, R. L. (2000). A Meta-Analysis of Response Rates in Web-or Internet-Based Surveys. *Educational and psychological measurement*, 60(6):821–836.

Crews, T. B. and Curtis, D. F. (2011). Online Course Evaluations: Faculty Perspective and Strategies for Improved Response Rates. *Assessment & Evaluation in Higher Education*, 36(7):865–878.

Feldman, K. A. (1976). Grades and College Students' Evaluations of Their Courses and Teachers. *Research in Higher Education*, 4(1):69–111.

Feldman, K. A. (1993). College Students' Views of Male and Female College Teachers: Part II—Evidence from Students' Evaluations of Their Classroom Teachers. *Research in Higher Education*, 34(2):151–211.

Feldman, K. A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In P,

P. R. and C, S. J., editors, *The Scholarship of Teaching and Learning in Higher Education: An Evidenced-based Perspective*, pages 93–143. Springer.

Goos, M. and Salomons, A. (2017). Measuring Teaching Quality in Higher Education: Assessing Selection Bias in Course Evaluations. *Research in Higher Education*, 58(4):341–364.

Gravestock, P. and Gregor-Greenleaf, E. (2008). *Student Course Evaluations: Research, Models and Trends*. Toronto: Higher Education Quality Council of Ontario.

Hativa, N. (2014). *Student Ratings of Instruction: Recognizing Effective Teaching*. Oron Publications.

Isely, P. and Singh, H. (2005). Do Higher Grades Lead to Favorable Student Evaluations? *The Journal of Economic Education*, 36(1):29–42.

James, D. E., Schraw, G., and Kuch, F. (2015). Using the Sampling Margin of Error to Assess the Interpretative Validity of Student Evaluations of Teaching. *Assessment & Evaluation in Higher Education*, 40(8):1123–1141.

Lewis, K. G. (2001). Making Sense of Student Written Comments. *New Directions for Teaching and Learning*, 87:25–32.

MacNell, L., Driscoll, A., and Hunt, A. N. (2015). What's in a Name: Exposing Gender Bias in Student Ratings of Teaching. *Innovative Higher Education*, 40(4):291–303.

Marsh, H. and Dunkin, M. (1997). Students' Evaluations of University Teaching: A Multidimensional Perspective. In *Effective Teaching in Higher Education: Research and Practice*, pages 241–320. Agathon Press, Bronx, NY.

Marsh, H. W. (2007). Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases and Usefulness. In *The scholarship of teaching and learning in higher education: An evidence-based perspective*, pages 319–383. Springer.

Marsh, H. W. and Roche, L. A. (2000). Effects of Grading Leniency and Low Workload on Students' Evaluations of Teaching: Popular Myth, Bias, Validity, or Innocent Bystanders? *Journal of Educational Psychology*, 92(1):202.

McGowan, W. R. and Osguthorpe, R. T. (2011). Student and Faculty Perceptions of Effects of Midcourse Evaluation. *To improve the academy*, 29(1):160–172.

McGuire, S. Y. (2015). *Teach Students How to Learn: Strategies You Can Incorporate into Any Course to Improve Student Metacognition, Study Skills, and Motivation*. Stylus Publishing, LLC.

McPherson, M. A. and Jewell, R. T. (2007). Leveling the Playing Field: Should Student Evaluation Scores Be Adjusted? *Social Science Quarterly*, 88(3):868–881.

Mertler, C. A. and Reinhart, R. V. (2005). *Advanced and Multivariate Statistical Methods: Practical Application and Interpretation*. Pyrczak Publishing, Glendale, CA, 3 edition.

Ory, J. C. and Ryan, K. (2001). How do Student Ratings Measure up to a New Validity Framework? *New directions for institutional research*, 2001(109):27–44.

Osterlind, S. J. (2010). *Modern Measurement: Theory, Principles, and Applications of Mental Appraisal*. Upper Saddle River, NJ, 2 edition.

Ouellett, M. L. (2005). *Teaching Inclusively: Resources for Course, Department and Institutional Change in Higher Education*. New Forums Press.

Potvin, G., Hazari, Z., Tai, R. H., and Sadler, P. M. (2009). Unraveling Bias from Student Evaluations of Their High School Science Teachers. *Science Education*, 93(5):827–845.

Shih, T.-H. and Fan, X. (2008). Comparing Response Rates from Web and Mail Surveys: A Meta-Analysis. *Field methods*, 20(3):249–271.

Shih, T.-H. and Fan, X. (2009). Comparing Response Rates in E-Mail and Paper Surveys: A Meta-Analysis. *Educational research review*, 4(1):26–40.

Stark, P. and Freishtat, R. (2014). An Evaluation of Course Evaluations. *ScienceOpen Research*.

Superson, A. M. (2002). Sexism in the Classroom: The Role of Gender Stereotypes in the Evaluation of Female Faculty. *Theorizing Backlash: Philosophical Reflections on the Resistance to Feminism*, pages 201–213.

Svinicki, M. D. (2004). *Learning and Motivation in the Postsecondary Classroom.* Anker Publishing Company, San Francisco, CA.

Theall, M. and Franklin, J. (2001). Looking for Bias in All the Wrong Places: A Search for Truth or a Witch Hunt in Student Ratings of Instruction? *New directions for institutional research*, 2001(109):45–56.

Tucker, B., Jones, S., and Straker, L. (2008). Online Student Evaluation Improves Course Experience Questionnaire Results in a Physiotherapy Program. *Higher Education Research & Development*, 27(3):281–296.

Willits, F. and Brennan, M. (2017). Another Look at College Student's Ratings of Course Quality: Data from Penn State Student Surveys in Three Settings. *Assessment & Evaluation in Higher Education*, 42(3):443–462.

Winer, L., Di Genova, L., Vungoc, P.-A., and Talsma, S. (2012). Interpreting End-of-Course Evaluation Results. *Montreal: Teaching and Learning Services, McGill University*.

Wines, W. A. and Lau, T. J. (2006). Observations on the Folly of Using Student Evaluations of College Teaching for Faculty Evaluation, Pay, and Retention Decisions and Its Implications for Academic Freedom. *William and Mary Journal of Women and the Law*, 13:167.

Zabaleta, F. (2007). The Use and Misuse of Student Evaluations of Teaching. *Teaching in Higher Education*, 12(1):55–76.