An Analysis of Winter 2023
Student Course Perceptions Survey Responses

Prepared January 2024 by
Sonya Buffone, Director, Teaching Assessment Processes
Kathy Becker, Specialist, Teaching Assessment Processes
Glen McGee, Assistant Professor, Statistics and Actuarial Science
Trang Bui, Statistical Consultant, Statistical Consulting and Survey Research Unit

Contents

Introduction

As part of a broader operational and cultural shift to a holistic model of teaching assessment at the University of Waterloo, the new Student Course Perception (SCP) Survey was officially launched in Winter 2022. The SCP survey is the first step in implementing this holistic model, which is grounded in institutional teaching and learning priorities. As part of its commitment to the ongoing monitoring of the SCP survey, the Teaching Assessment Processes (TAP) office collaborated with the Statistical Consulting and Survey Research Unit (SCSRU) in the Faculty of Mathematics to analyze SCP data from Winter 2023. The present study is the second in the series, with the first being An Analysis of Winter 2022 SCP Survey Responses.

Sources of Data

Data for this analysis was drawn from three primary sources: Winter 2023 SCP survey data, responses to Waterloo's Equity Survey from Winter 2023 instructors included in the study, and Human Resources records on 'sex at birth' for Winter 2023 instructors whose students completed SCP surveys.[1] In total, 646 of the 1415 Winter 2023 instructors in this study did not complete the Equity Survey. This translates to 45.7% missing cases for the equity data. As such, any findings that are based on this incomplete dataset must be interpreted with a high degree of caution.

Research Aims

The analysis in this report was guided by the following research focus: to identify the extent to which key factors identified in the literature, and by our Waterloo community members, are correlated with SCP ratings assigned by student respondents. Specifically, this analysis explored the strength of associations between SCP ratings and:

1. Instructor-Level Variables: instructor Indigenous identity, instructor racial identity, instructor sex, instructor appointment, instructor time in Canada, and;
2. Course-Level Variables: class size, course type (i.e., online or in-person) and Faculty of course offering.

A detailed coding scheme is provided in Appendix A.

Primary Analyses

The report consists of three primary analyses:

1. A descriptive analysis of instructor demographics and course characteristics at Waterloo.

[1] Equity Survey Information

2. An analysis of the mean SCP ratings across different combinations of instructor-level and course-level variables.
3. An analysis of the difference in mean ratings between (i) white and racialized instructors and (ii) male and female instructors across different combinations of instructor-level and course-level variables.

Key Findings

Due to the above-noted nonresponse in Equity Survey data, we were unable to examine the impact of Indigenous identity on SCP ratings since data on instructor Indigenous identity was both incomplete and small. While also incomplete, data on instructor racial identity was large enough to allow some analysis.

Descriptive Analysis of Demographics and Course Characteristics

In total, 44,578 SCP surveys from 2459 section-instructor pairs were submitted by students from across the six Faculties and Renison. It is important to emphasize that, since one section of a course might have more than one instructor, and since one instructor might teach more than one section, the 2459 section-instructor pairs include the following:

- 1415 unique instructors
- 1538 unique courses
- 2316 distinct sections

Due to some issues the team faced when merging various data sources, some instructor-section pairs may be counted more than once in cases where they are cross-listed in different faculties and have different course-level info listed. Such cases were merged in the data set as a result.

Instructor-level Variable: Indigenous Identity

Eight Winter 2023 instructors identified as Indigenous in the Equity Survey (see Figure 1).

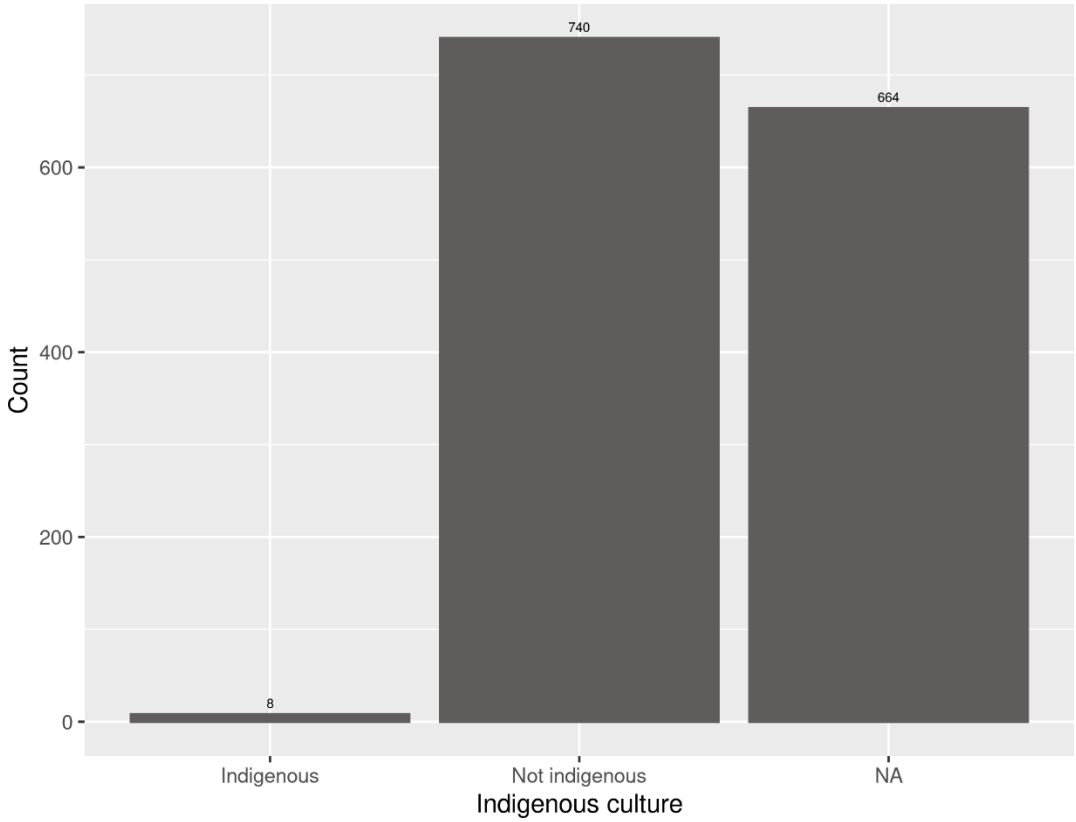

Figure 1: Number of Winter 2023 instructors by self-reported Indigenous identity – Note: Each Equity Survey respondent was asked to select all applicable racial identities. NA represents Winter 2023 instructors who did not complete the Equity Survey or preferred not to answer the Indigenous identity question.

Instructor-level Variables: Racial Identity and Time in Canada

Instructor Racial Identity

Turning to Figure 2, a further breakdown of the equity data shows the racial identities of Winter 2023 instructors. In cases where a respondent selected more than one racial identity, they were coded in each of the categories selected. For example, an instructor who selected both “East Asian” and “Another race category” would be included in those two racial identity categories.

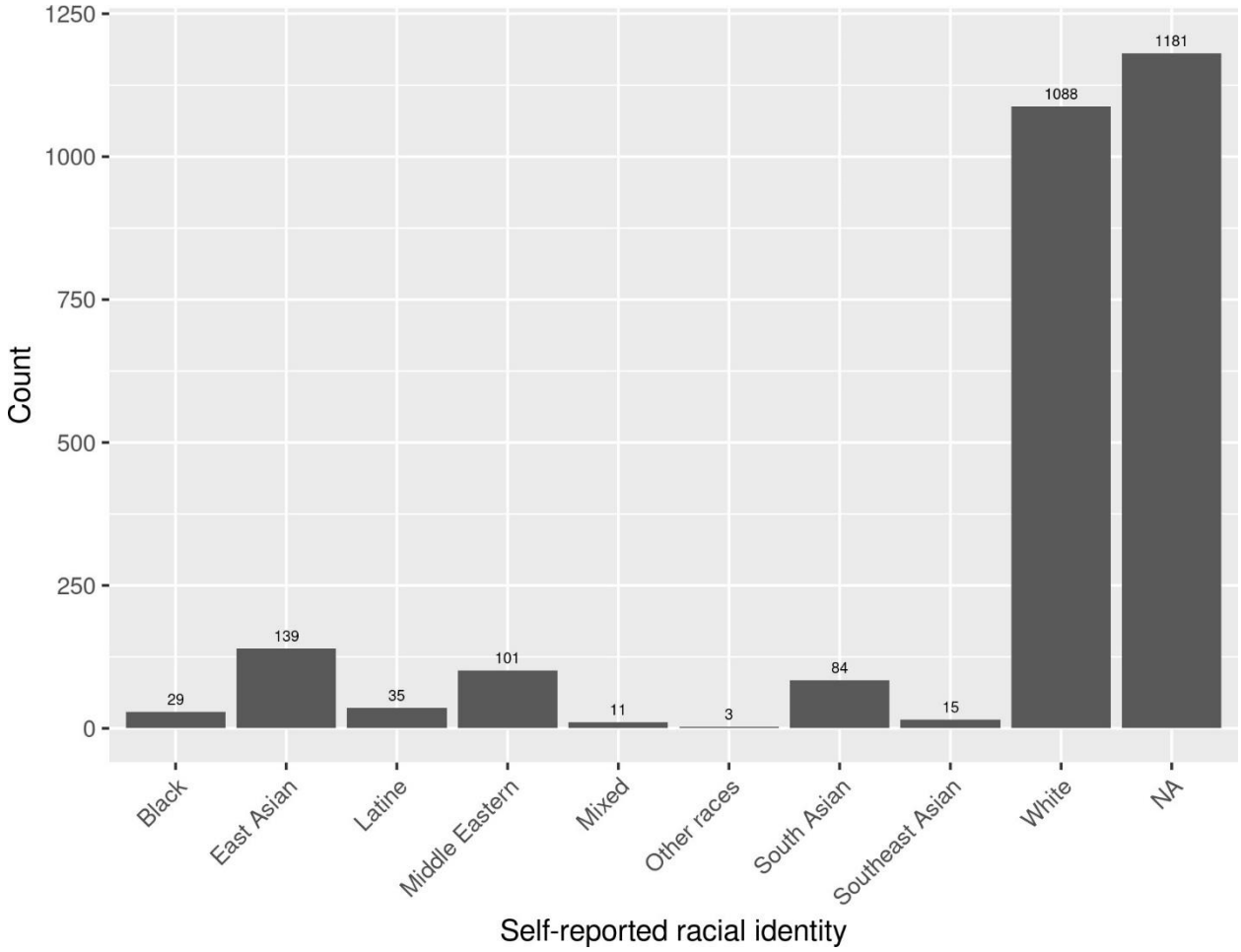

Figure 2: Number of Winter 2023 instructors by self-reported racial identity - Note: Each Equity Survey respondent was asked to select all applicable racial identities. NA represents Winter 2023 instructors who did not complete the Equity Survey or preferred not to answer the racial identity question.

Racialized Instructors by Faculty

We examined the number of racialized instructors across each Faculty who had SCP surveys in Winter 2023. As shown in Figure 3, setting missing cases aside, there were around four times more white than racialized instructors teaching in Winter 2023 across three of the six Faculties (Health, Arts, Environment). There were around two times more white than racialized instructors in one Faculty (Math), three times more white than racialized instructors in one Faculty (Science), and almost equal numbers of white and racialized instructors in one Faculty (Engineering).

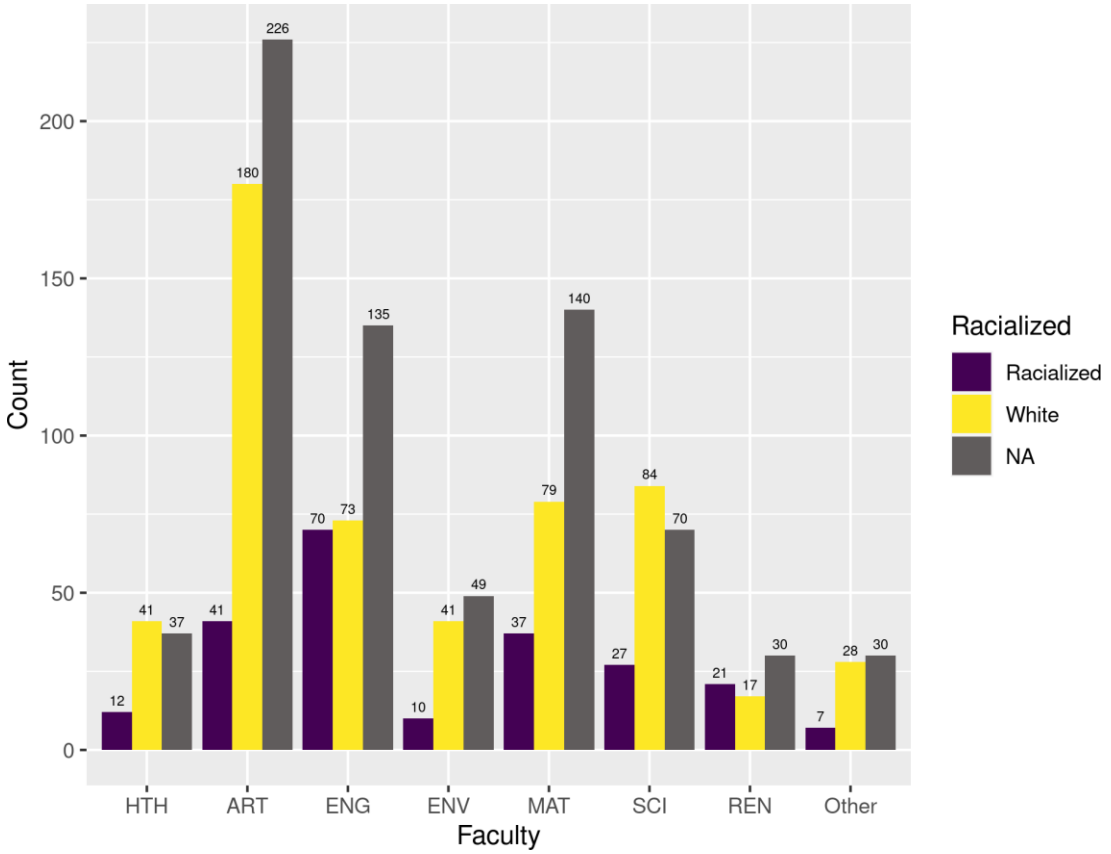

Figure 3: Number of Winter 2023 instructors by faculty and racial identity – Note: NA represents Winter 2023 instructors who did not complete the Equity Survey or preferred not to answer the racial identity question. “Other” includes University Colleges (except for Renison) and those that do not belong to any Faculty (e.g., GRAD).

Instructor Time in Canada

Figure 4 outlines the length of time in Canada for Winter 2023 instructors who responded to the Equity Survey. Overall, 405 instructors in the sample were born in Canada, 274 had been in Canada 5+ years and only 32 instructors reported being in Canada for <5 years. Again, it is important to be mindful of the proportion of missing cases (n= 704) for whom no information on time spent in Canada is available.

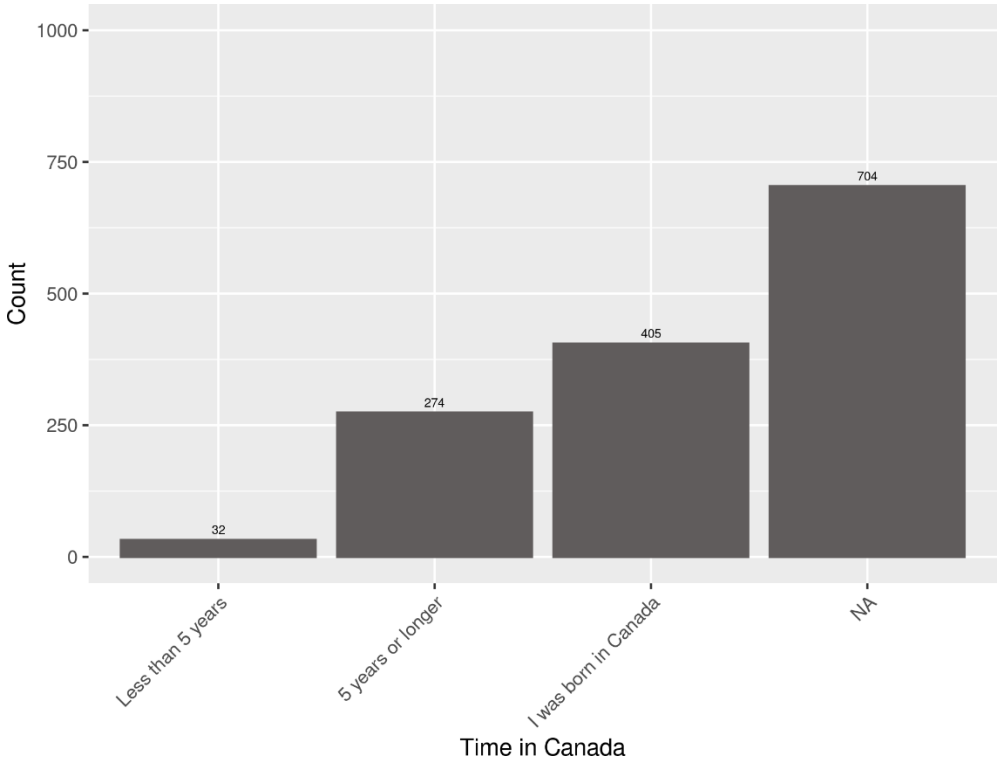

Figure 4: Number of Winter 2023 instructors by self-reported time in Canada. Note: NA represents Winter 2023 instructors who did not complete the Equity Survey or preferred not to answer the time in Canada question.

Instructor Racial Identity by Time in Canada

Figure 5 provides a further breakdown of the characteristics of instructors who responded to the Equity survey. As indicated by the yellow bars, we can see that of the respondents who are white, over two-thirds (72%) were born in Canada. This is in contrast with the much small number of racialized respondents (19%) who were born in Canada. For those respondents who were born outside of Canada, nearly three-quarters of those who are also racialized (72%) report being in Canada 5+ years compared to only one quarter of those who are white (25%) report being in Canada 5+ years.

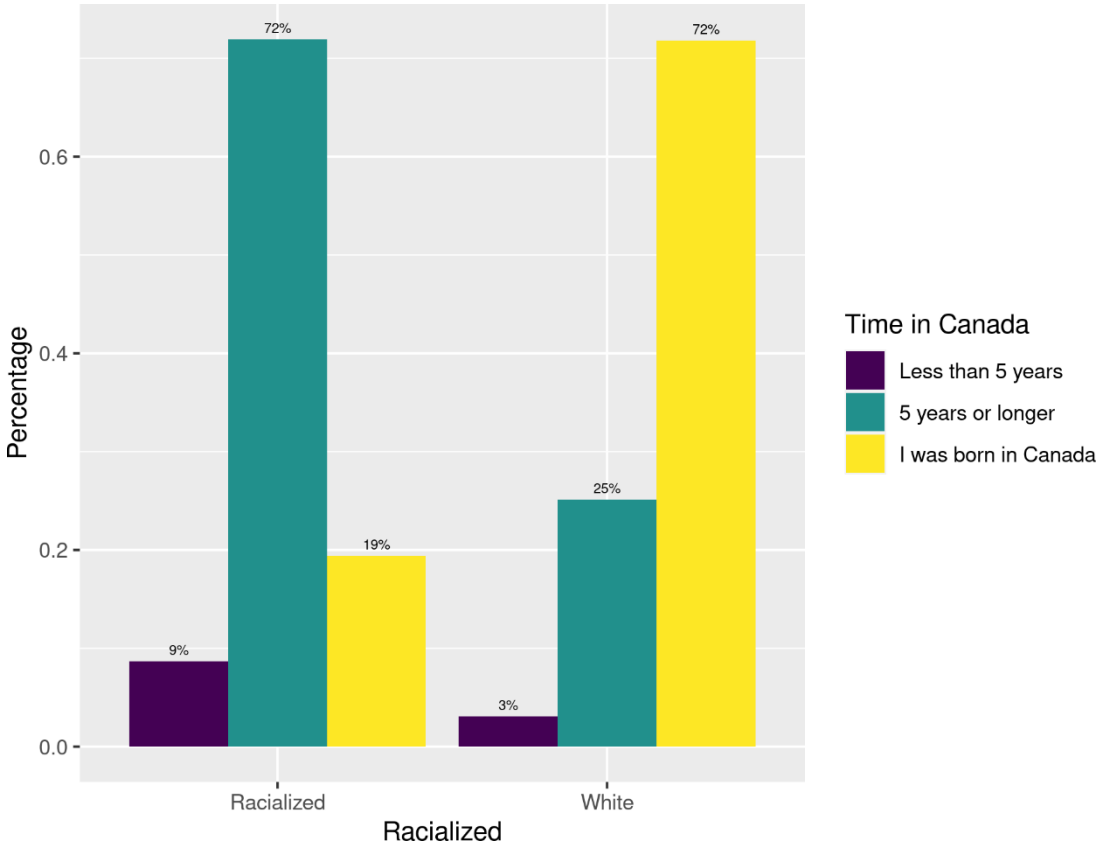

Figure 5: Number of white and racialized Winter 2023 instructors by time spent in Canada.

Instructor-level Variable: Sex

Instructor Sex

Figure 6 displays Human Resources data for 'sex at birth' for Winter 2023 instructors (542 female and 870 male). It's important to note that this data does not necessarily reflect instructor gender identity at the present time.

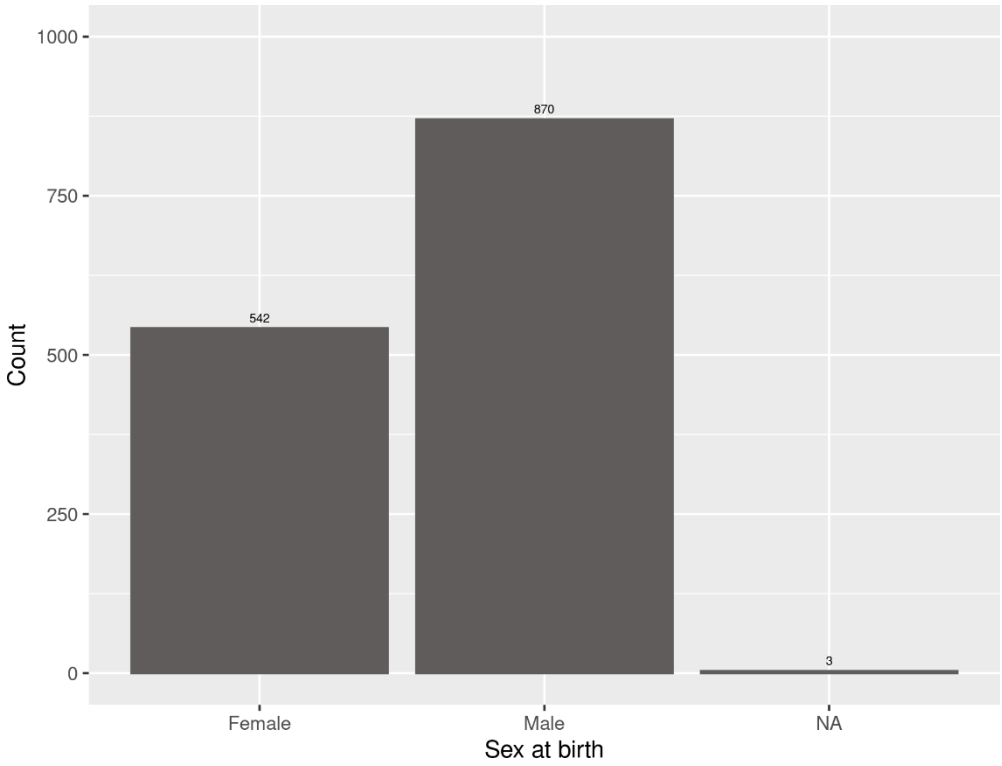

Figure 6: Number of Winter 2023 instructors by instructor sex at birth.

Instructor Sex by Faculty

Figure 7 shows the sex at birth for Winter 2023 instructors across each faculty. Male instructors outnumber female instructors across the STEM Faculties: Engineering, Math and Science. In two STEM Faculties, there are over three times more male than female instructors (Engineering: 212 males, 66 females, and Math: 193 males, 63 females). In one STEM Faculty, we see just under two times more male than female instructors (Science: 113 males and 67 females). In one Faculty, we see just under two times more male than female instructors (Environment: 63 males, 36 females). And in one Faculty we near-parity in the number of male and female instructors teaching in Winter 2023 (Arts: 234 males, 211 females).

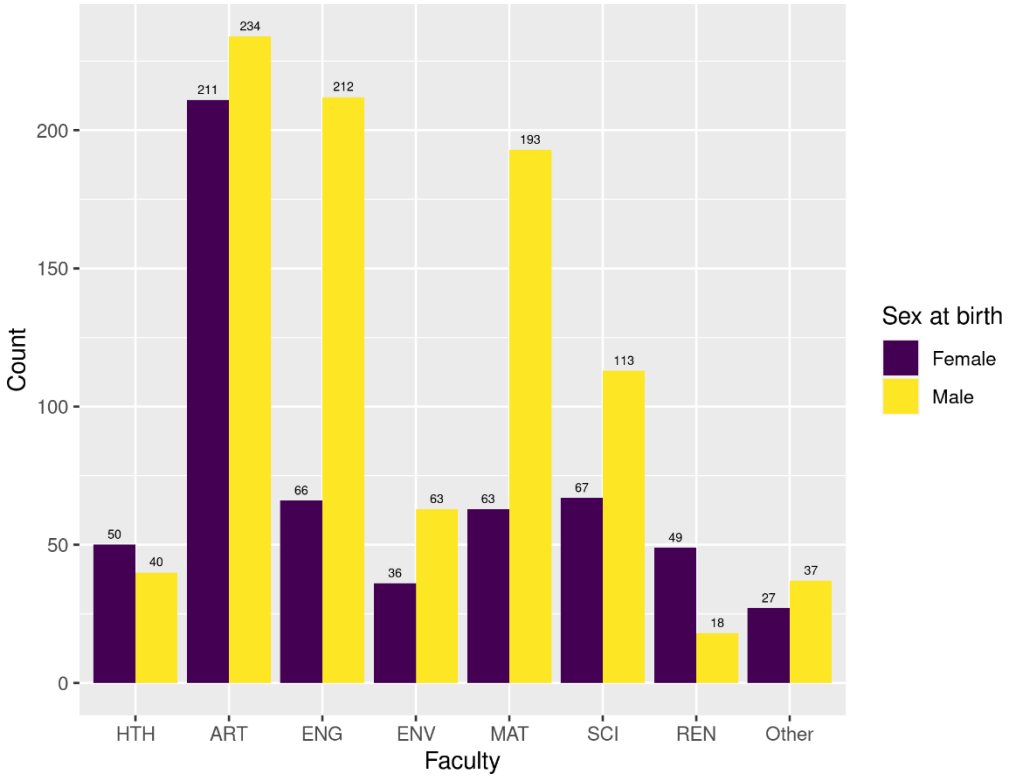

Figure 7: Number of Winter 2023 instructors by Faculty and Sex.

Instructor Sex by Appointment Type

Figure 8 further breaks down our sample characteristics by Winter 2023 instructor appointment type and sex at birth. In general, male instructors outnumber female instructors across all instructor types. Nearly four times more Winter 2023 Professors were male (198 males, 51 females) and two times more Winter 2023 Associate Professors were male (159 males, 79 females). The makeup of Winter 2023 instructors in remaining instructor types was less drastically imbalanced: Assistant Professors (69 males, 56 females), Lecturers (132 males, 87 females), and Sessionals (210 males, 150 females).

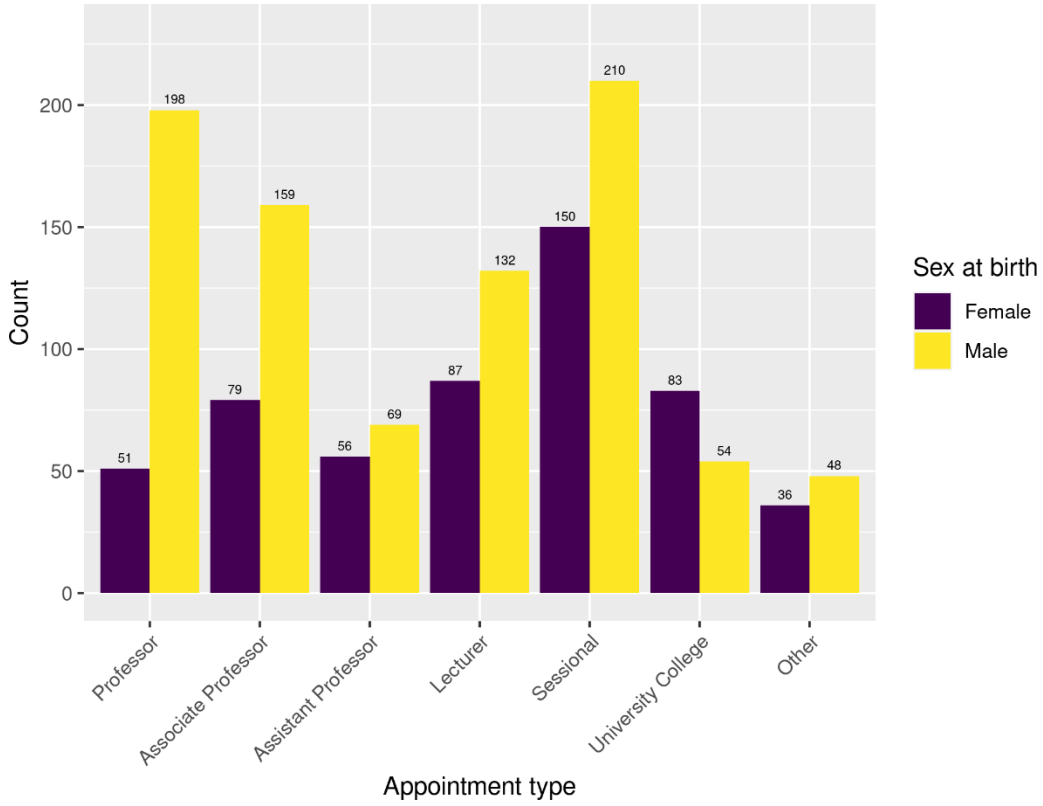

Figure 8: Number of Winter 2023 instructors by appointment type and sex at birth.

Course-level Variable: Class Size

Figure 9 outlines the number of Winter 2023 courses included in this study by class size. The bulk of courses (1,998) had 100 or fewer students; only a small number (57) had 201+ students enrolled.

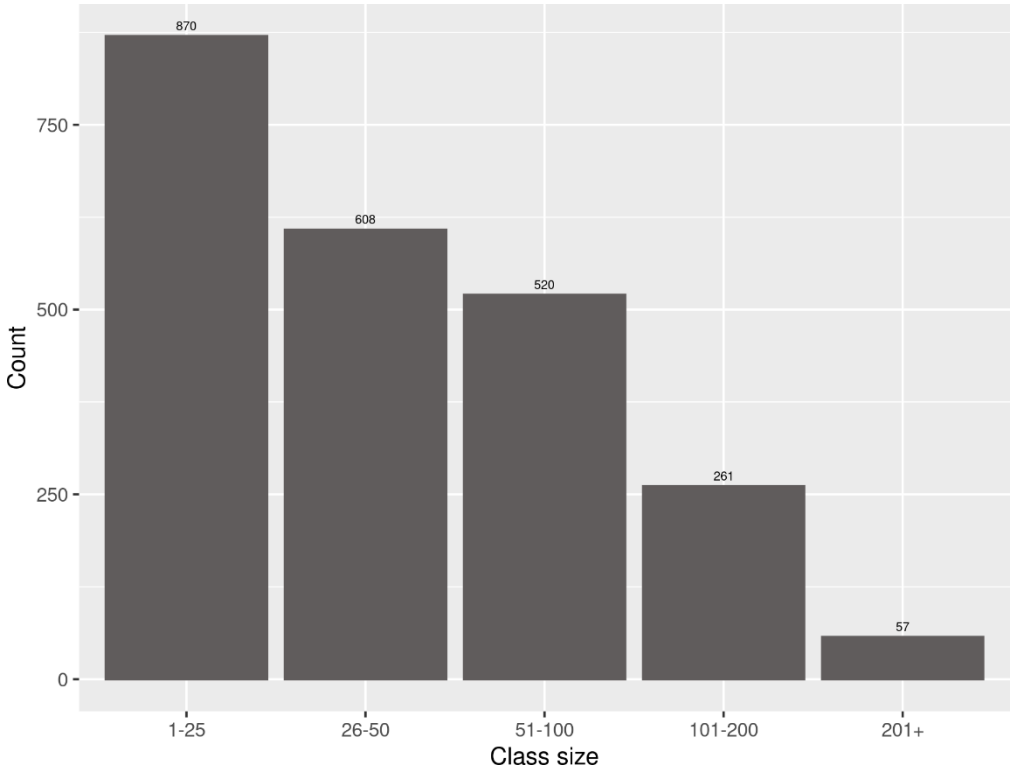

Figure 9: Number of Winter 2023 courses by class size.

Course-level Variable: Course Type by Faculty

The majority of Winter 2023 courses included in this study were held in person, as reflected in Figure 10. The Faculty of Arts had the highest number of online course offerings (124), but still held over five times that number of courses in person (667). Engineering held the second highest number of in-person courses (437), with only 15 courses offered online.

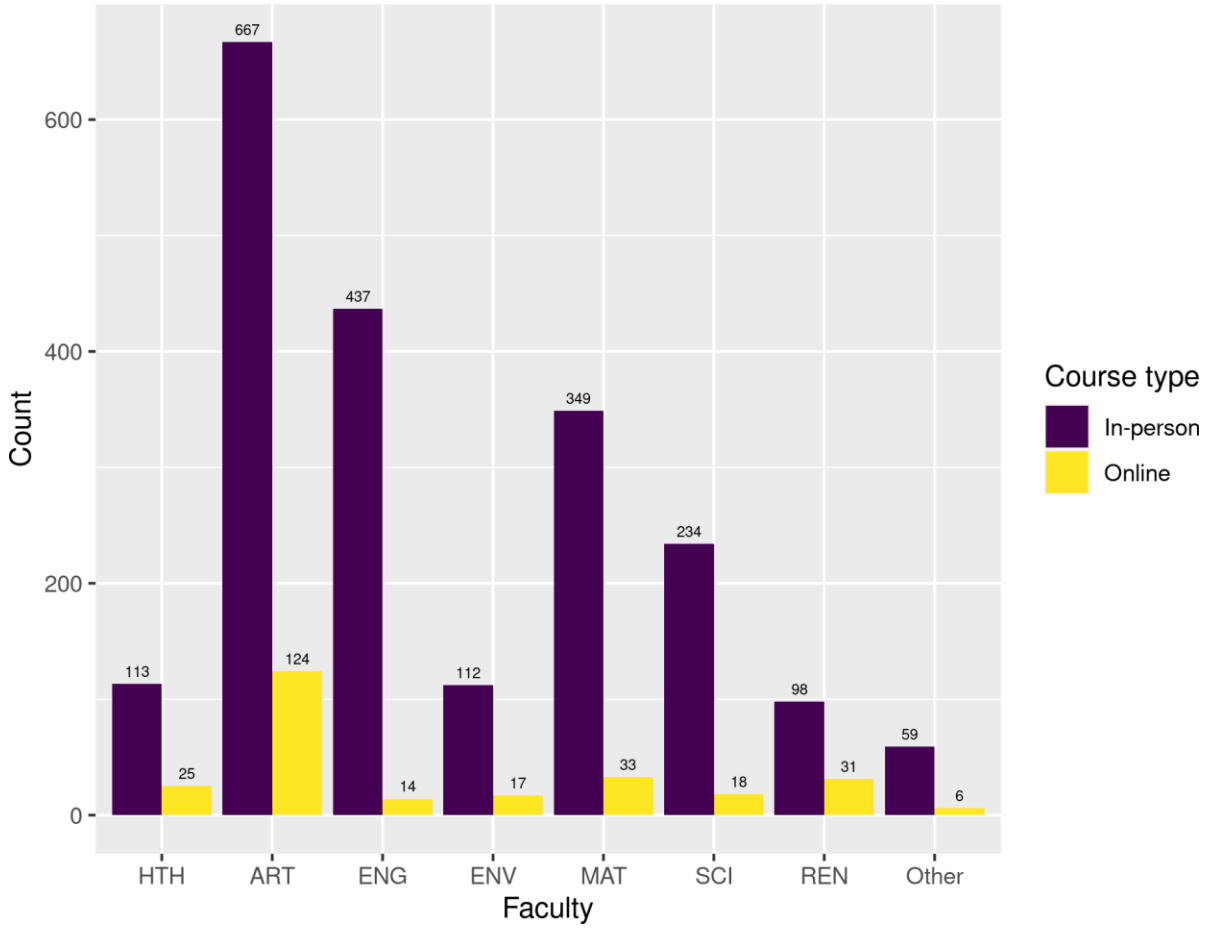

Figure 10: Number of Winter 2023 course sections by faculty and course type.

SCP Responses by Response Item

Figure 11 displays the frequency with which student respondents assigned each rating for the six core questions on the SCP. The 'No Basis for Rating' response was selected only 1-2% of the time. As is often the case with course evaluation surveys, the data is highly skewed, evidenced by the clustering of scores at the high-end of the scale. Specifically, respondents selected 'Agree' and 'Strongly Agree' between 71% and 83% of the time across all six items. Conversely, respondents selected 'Disagree' and 'Strongly Disagree' between 6% and 15% of the time across all six items. We can see that student respondents selected the mid-point ('Neutral') option on the five-point scale between 10% and 16% of the time across all six items.

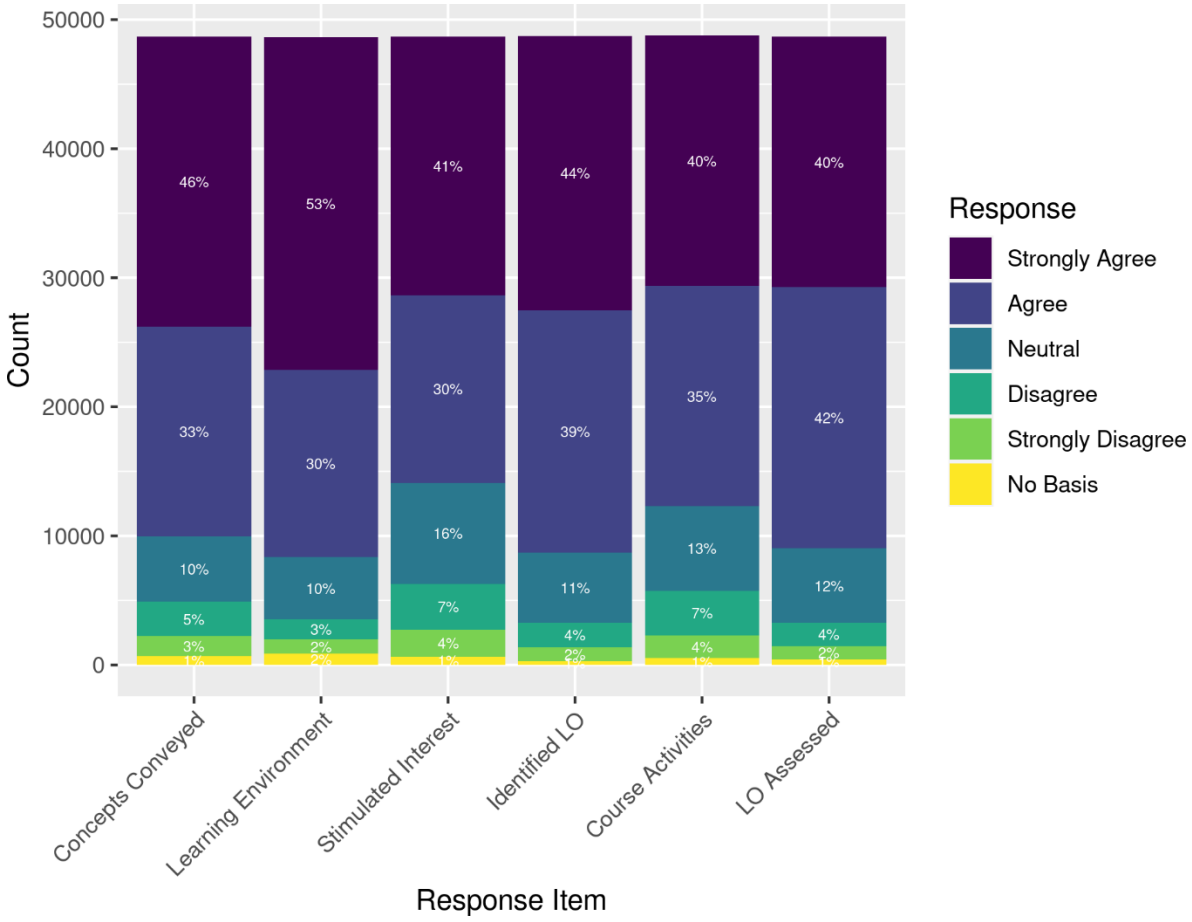

Figure 11: Proportion of Winter 2023 responses per response item.

SCP Mean Ratings

The following sections contain analyses of mean ratings (MR) of SCP scores for each response item and for various combinations of instructor-level and course-level variables. For a given combination of variables (e.g., assistant professors teaching online courses), the MR in that group is calculated using a mixed-effects model described in detail in Appendix B. Roughly speaking, the calculation is as follows:

- First, we assign a numerical value of 1-5 for each of the response categories from Strongly Disagree (score = 1) to Strongly Agree (score = 5).
- Next, the responses are averaged across students in each instructor/course-section pair. Thus, a multi-section course would have multiple pairs, as would a single section taught by multiple instructors. Students who leave a response item blank or respond with "No Basis for Rating" are not included in the average.
- Using a mixed-effects model approach a course-section-level random effect was identified. Course-section-level effects capture factors specific to the course section

offering, such as physical comfort of the classroom, student preference for the day or time of class offering, or other factors impacting student learning experience that are not related to teaching effectiveness. A score for each instructor is then obtained by assuming that the instructor/course-section average is the sum of the instructor score and the course-section-level effect and using the mixed-effects model to extract the former.

- Finally, the instructors' MR in the given group is the average of individual instructor scores across all instructors in the group.

The mixed-effects model also provides standard errors (SE) and 95% confidence intervals (CI) for the MR which account for the number of instructors, the number of course-sections taught by each instructor, and the number of student responses in each instructor/course-section pair.

Mean Ratings by Response Item and Class Size

Figure 12 displays the mean rating for each of the six response items accounting for class size. Reiterating what we observed in Figure 11, we can see that scores tend to cluster at the higher-end of the five-point scale, with most response items receiving scores between 3.9 and 4.5.

Students tend to rate their learning experiences more positively in smaller classes, as evidenced by higher ratings across all six survey items. This is not unexpected since smaller class sizes have more opportunity for individual student attention, and it aligns with the broader literature as well as past Waterloo SCP analyses. It also is interesting to observe very similar trends in the plotting of the data across all class sizes (the dotted line patterns). We see a peak for Learning_Environment across all class sizes (ranging from 4.0 in 200+ courses to close to 4.5 in courses with 25 or fewer students). Conversely, Stimulated_Interest (ranging from 3.75 for 200+ students to close to 4.3 for 1-25 students) and Course_Activities (ranging from 3.9 for 200+ students to 4.2 for 1-25 students) receive the lowest mean rating across all class sizes.

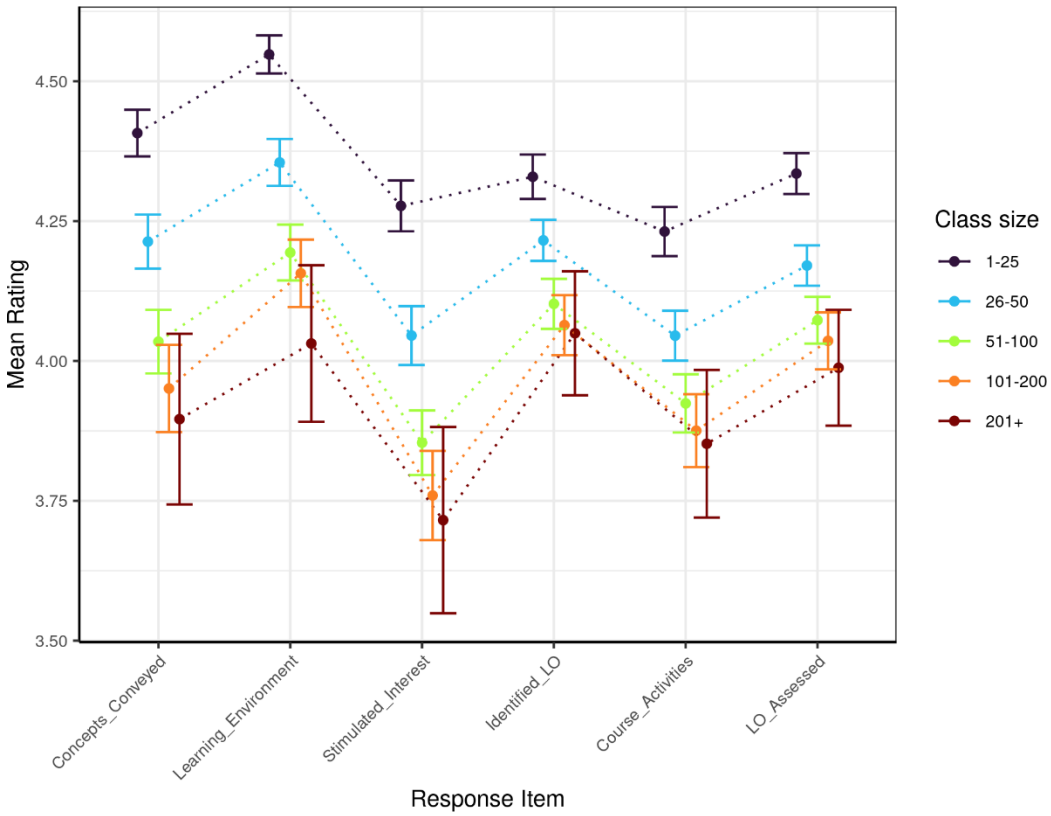

Figure 12: Mean Winter 2023 rating by response item and class size. Error bars correspond to 95% confidence intervals.

Mean Ratings by Response Item and Faculty of Course Offering

Figure 13 shows the mean ratings across the six core items by Faculty. Perhaps unsurprisingly, results here look similar to what we have observed throughout this report: in general, scores cluster at the higher end of the five-point scale accounting for Faculty (from 3.9 at the lowest end to about 4.5 at the higher end). We can see that across most Faculties Concepts_Conveyed, Learning_Environment, Identified_LOs and LOs_Assessed received higher scores while Stimulated_Interest and Course_Activities received slightly lower scores accounting for Faculty. It is important to emphasize that the differences in practical terms, on the 5-point Likert scale, are very small. For example, the mean rating for Course_Activities in Science was 3.9 and in Arts was 4.1, which translates to a negligible difference – just 0.2 on the 5-point scale.

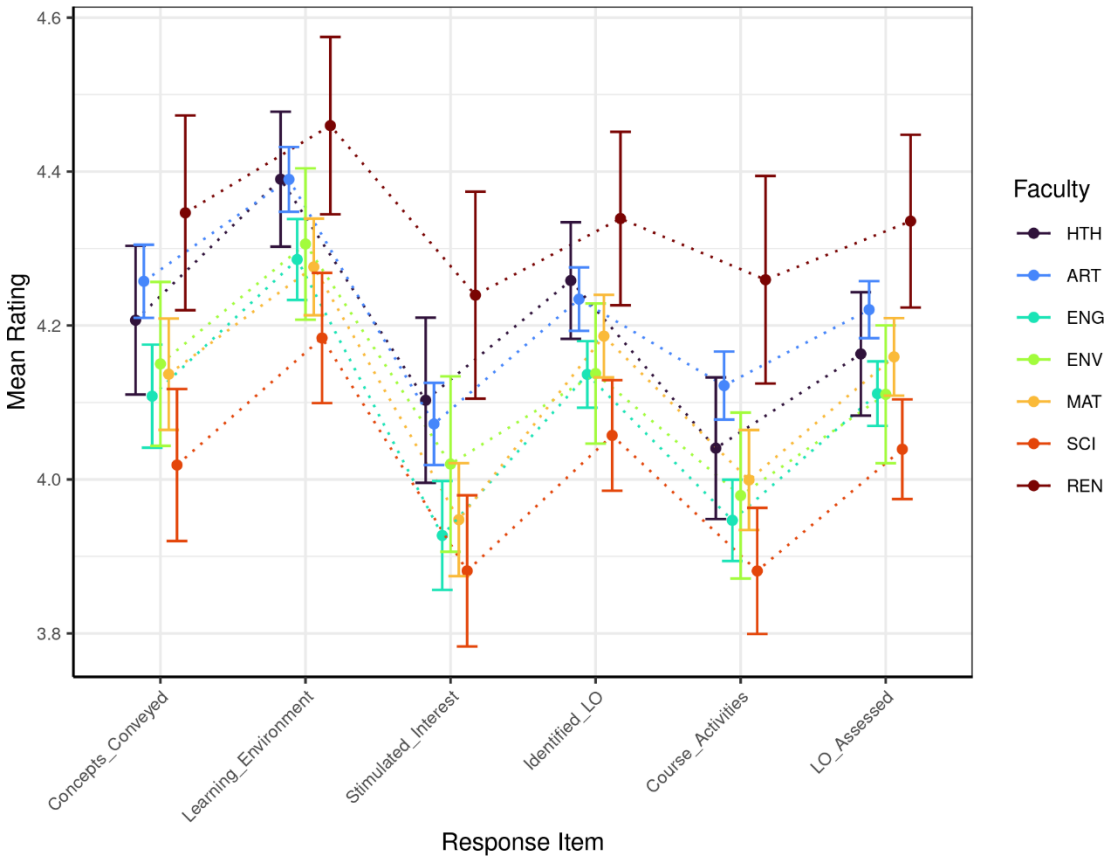

Figure 13: Mean Winter 2023 rating by response item and Faculty of course offering. Error bars correspond to 95% confidence intervals.

Mean Ratings by Response Item and Instructor Time in Canada

Figure 14 shows the mean ratings for response items taking the Instructor's time spent in Canada into account. The findings reveal that for instructors born in Canada, mean ratings are slightly higher (under 0.1 higher) for Stimulated_Interest. For instructors with less than 5 years in Canada, mean ratings are slightly lower (under 0.1 lower) for Course_Activities. Overall, the differences in scores across all three groups is relatively small ranging from about 0.1-0.2 points with extremely large confidence intervals for instructors with <5 years in Canada. The large confidence intervals reflect the small number of Equity Survey respondents who provided information about time spent in Canada (n=32). Interestingly, accounting for time spent in Canada, we see the same patterning in the data insofar as Learning_Environment received the highest scores across all three categories, while Stimulated_Interest and Course_Activities scored the lowest.

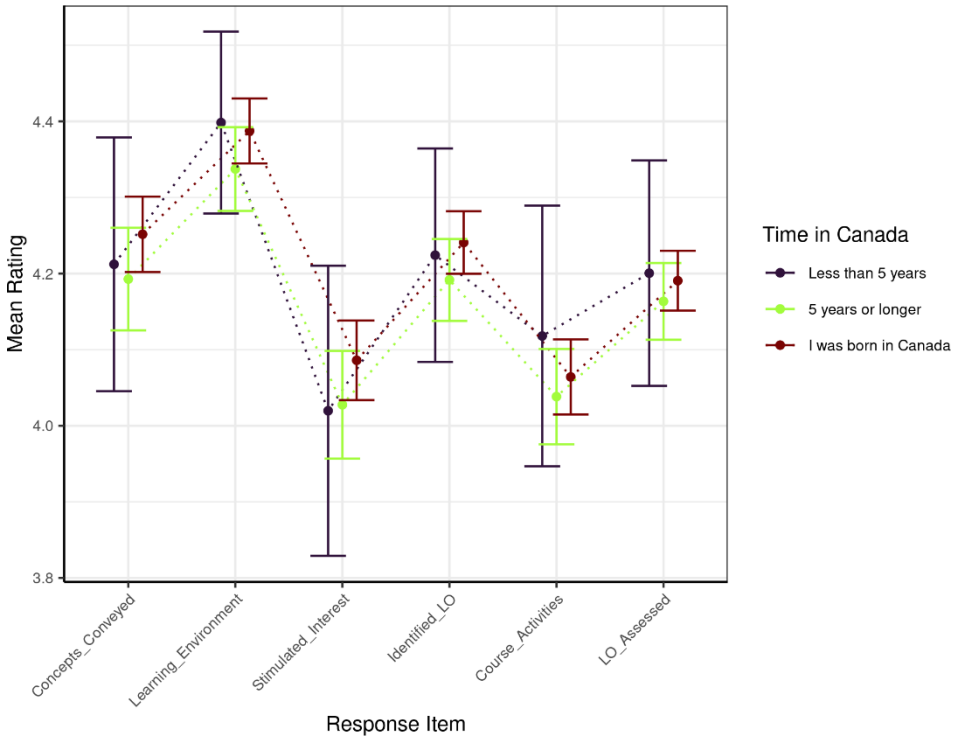

Figure 14: Mean Winter 2023 rating by response item and instructor's self-reported time in Canada. Error bars correspond to 95% confidence intervals.

Differences in Mean Ratings

This section examines differences in mean ratings (MR) assigned by students using different instructor- and course-level variables across the six core items on the SCP survey. This includes differences in ratings assigned by students for racialized instructors and white instructors, male instructors and female instructors, instructors of different instructor types, classes of different sizes, and different courses type (e.g., online or in-person). **It is important to note that the following analyses is exploratory and cannot be used to make causal claims about the effect of racial or sex on SCP ratings**. Nonetheless, findings provide a useful initial step in looking for systemic biases in course evaluations at the University of Waterloo.

Interpreting the Plots

- Compare plotted points to the zero line
    - The closer to the zero line, the smaller the difference in MR (a difference of 0 suggests no association).
- Compare error bars (which represent 95% confidence intervals) to the zero line:
    - When the confidence interval doesn't cross the zero line, it indicates a statistically significant difference in MR using a 95% confidence interval ($p \leq 0.05$)
    - When the confidence interval crosses the zero line, it indicates a difference in MR that is not statistically significant using a 95% confidence interval ($p > 0.05$)

- The confidence intervals show the difference in scores assigned by students to course surveys taught by those who experience systemic privilege (white/male) and those who experience systemic oppression (racialized/female). The position of the plot point relative to the zero line shows who experiences a benefit from the difference in scores:
    - When the plot point is above the zero line, the difference benefits those who experience systemic privilege.
    - When the plot point is below the zero line, the difference benefits those who experience systemic oppression.
- In cases where a certain combination of covariates does not exist (e.g., assistant professor + class size of 201 + Environment), no data will be displayed in the accompanying figure.
- To protect instructor anonymity, in cases where a certain combination of covariates exists but results in five or fewer instructor-section pairs (e.g., associate professor + class size of 1-25 + Environment), the box is displayed but will be blank.

Analysis of Differences: Instructor Racial Identity

Given the very small number of instructor course-pairs that emerged as we began to categorize this data set (i.e., when we apply more fine-grained classifications like racialized identity by class size, course type, appointment type etc.), the most meaningful approach to examine differences in scores for those who experience systemic oppression and others who experience systemic privilege was to compute a binary variable to represent racialized identity. This binary variable is described in Appendix A: Coding Scheme.

Although creating a binary variable for this important measure (racialized identity) is less than ideal, this was the most reasonable solution to use the data while still protecting instructor anonymity. Additionally, the small number of individuals within racialized groups precluded meaningful statistical tests at a deeper level.

This score was calculated as follows:

MR assigned to white instructors − MR assigned to racialized instructors

As highlighted previously in this report, results in this section should be interpreted with caution given the high number of Winter 2023 instructors who did not complete the Equity Survey (45.7%). As a result, it is difficult to attribute significance to any findings. Nonetheless, this analysis is an important first step to understanding how student responses to SCP surveys are impacted by racial biases and systemic privilege - something that we have only begun to explore at Waterloo. An additional note of caution is worth repeating: **the analyses conducted are exploratory only and do not, under any circumstances, claim to make causal or definitive inferences.**

Difference in MR for White and Racialized Instructors

Figure 16 displays the difference in mean ratings for white and racialized instructors across all response items. Error bars cross the zero-line for all response items, signifying that any differences are not statistically significant at the 95% confidence interval. Table 1 contextualizes these scores, displaying the differences in mean ratings on the five-point scale for racialized and white instructors. Overall, the difference in ratings ranges from 0.01-0.04 points on the five-point Likert Scale. **These small differences, and the perceived tendency of some performance assessors to over-interpret small differences in SCP scores, warns of the dangers of such practices and speaks strongly in favor of not reporting scores beyond one-decimal point.**

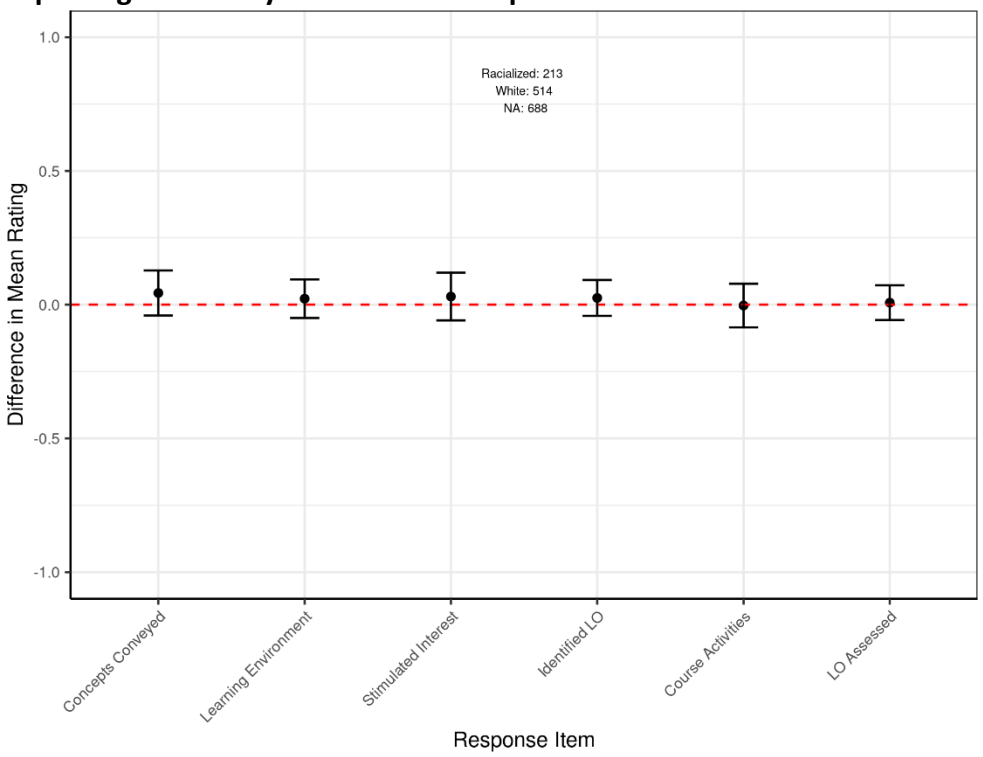

Figure *15*: Difference in mean Winter 2023 rating for white and racialized instructors by response item.

Table 1: Difference in mean Winter 2023 ratings for white and racialized instructors by response item, with 95% confidence interval (C.I.)

	White Avg.	Racialized Avg.	Difference	White C.I.	Racialized C.I.	Difference C.I.
Concepts Conveyed	4.24	4.19	0.04	(4.19, 4.28)	(4.11, 4.26)	(-0.04, 0.13)
Learning Environment	4.37	4.35	0.02	(4.33, 4.41)	(4.29, 4.41)	(-0.05, 0.09)
Stimulated Interest	4.07	4.03	0.03	(4.02, 4.12)	(3.95, 4.11)	(-0.06, 0.12)
Identified LO	4.22	4.20	0.03	(4.19, 4.26)	(4.14, 4.26)	(-0.04, 0.09)
Course Activities	4.05	4.05	0.00	(4.01, 4.1)	(3.98, 4.12)	(-0.08, 0.08)
LO Assessed	4.18	4.17	0.01	(4.14, 4.21)	(4.11, 4.23)	(-0.06, 0.07)

Difference in MR for White and Racialized Instructors by Faculty of Course Offering

In Figure 18 we explore the difference in mean ratings for white and racialized instructors in each Faculty. Overall, the difference in mean ratings for white and racialized instructors is not statistically significant accounting for Faculty. The average difference in ratings across all six items is generally less than 0.1 across all Faculties, with the exception of ENV. In Environment we see a larger difference (about 0.25) across all response items and very large error bars, which is reflective of the small sample (41 white instructors and 10 racialized instructors).

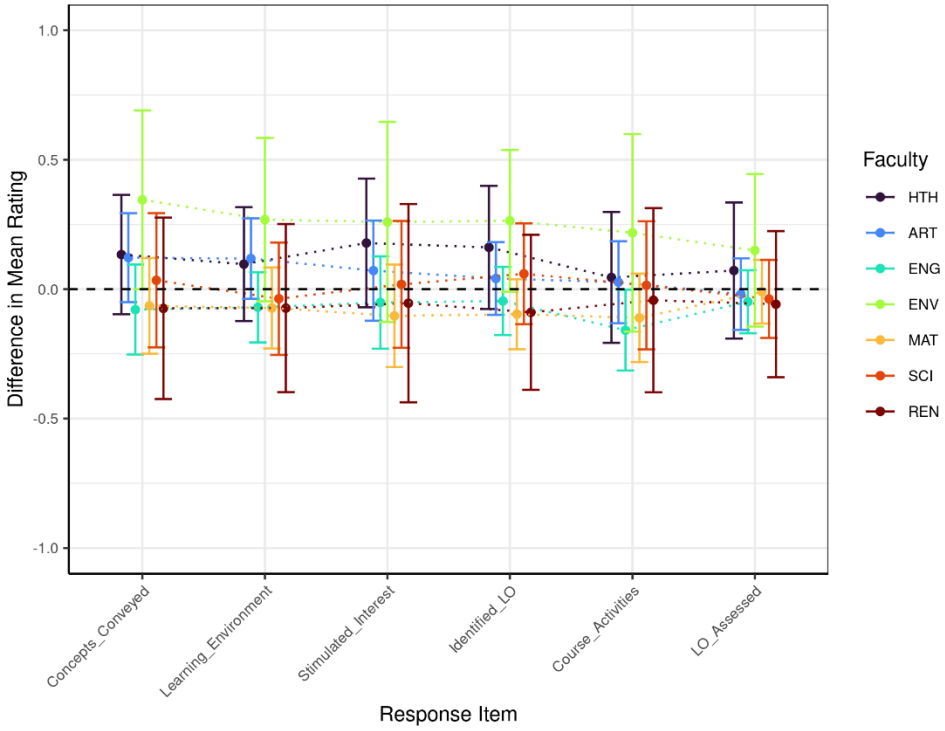

Figure 16: Difference in mean Winter 2023 ratings for white and racialized instructors by response item and Faculty of course offering. Error bars correspond to 95% confidence intervals.

Difference in MR for White and Racialized Instructors by Class Size and Instructor Appointment Type

Figure 20 displays the difference in mean ratings given by student respondents for white and racialized instructors while also taking instructor appointment and class size into consideration. In all instances, differences observed are not statistically significant (as evidenced by the error bar crossing or touching the zero line). And in most cases, the non-significant difference is very small (as evidenced by the data points falling close to or on the zero line).

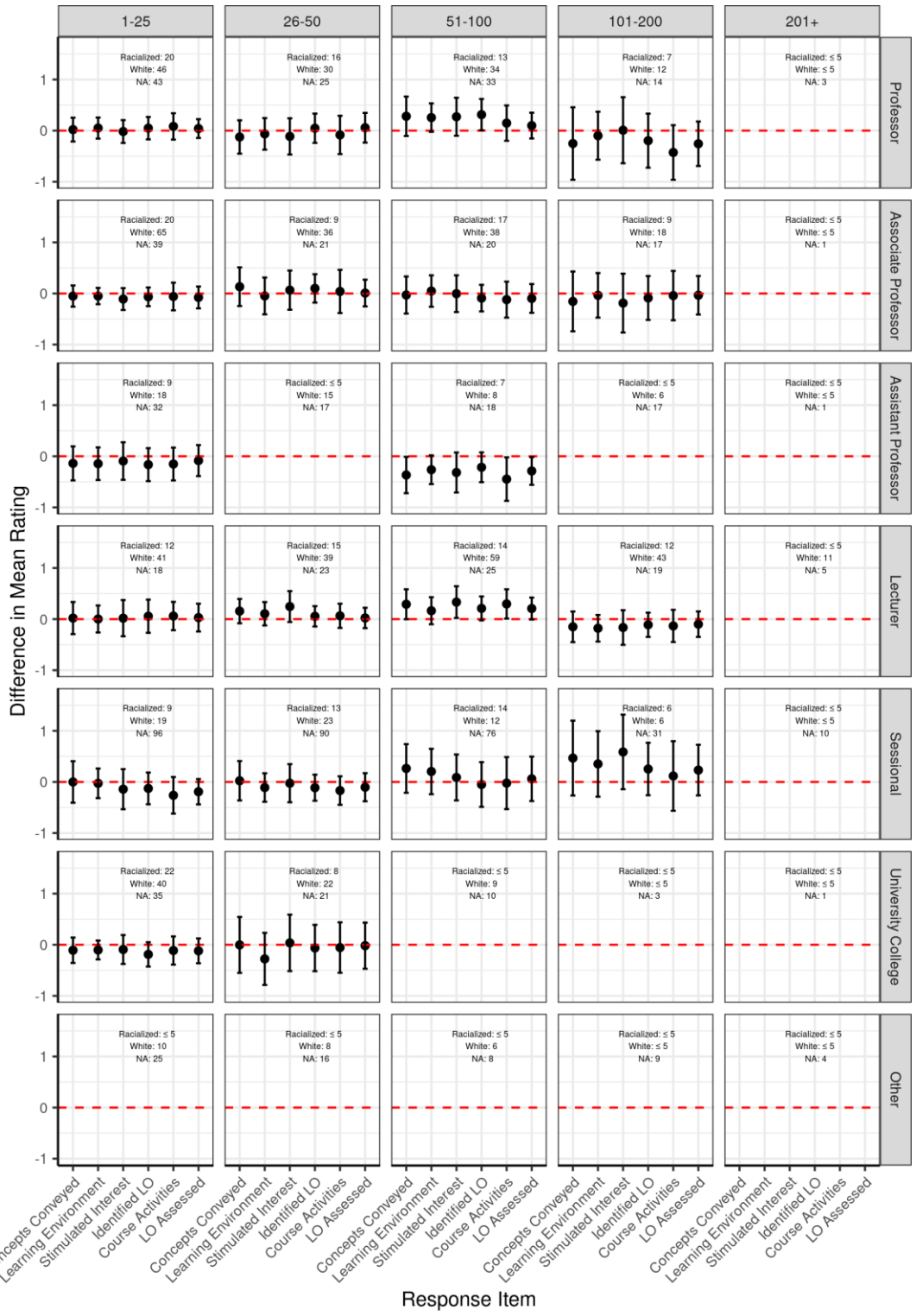

Figure 17: Difference in mean Winter 2023 ratings for white and racialized instructors by response item, instructor appointment type, and class size. Error bars correspond to 95% confidence intervals.

Difference in MR for White and Racialized Instructors by Instructor Appointment Type and Course Type

Figure 22 shows the difference in mean ratings given by students for racialized and white instructors accounting for course type (online or in-person) and instructor appointment type. One finding that stands out is that confidence intervals are larger for online classes as compared to those for in-person classes. This is indicative of the small number of online classes (n=268) included in the study. In all instances except two, differences observed are not statistically significant (as evidenced by the error bar crossing or touching the zero line). And in most cases, the non-significant difference is very small (as evidenced by the data points falling close to or on the zero line). The two instances that suggest statistical significance (where the error bar does not cross the zero line) are within the plot for Sessionals teaching Online for the Learning Environment and Identified LO items. The data points for these items indicate that student ratings for this question when they have a white instructor are higher (about 0.5) than when they have a racialized instructor. However, given the small sample size (n= 7 racialized Sessionals and n= 11 white Sessionals), attributing too much weight to this finding would not be prudent. The TAP office will continue to monitor these findings.

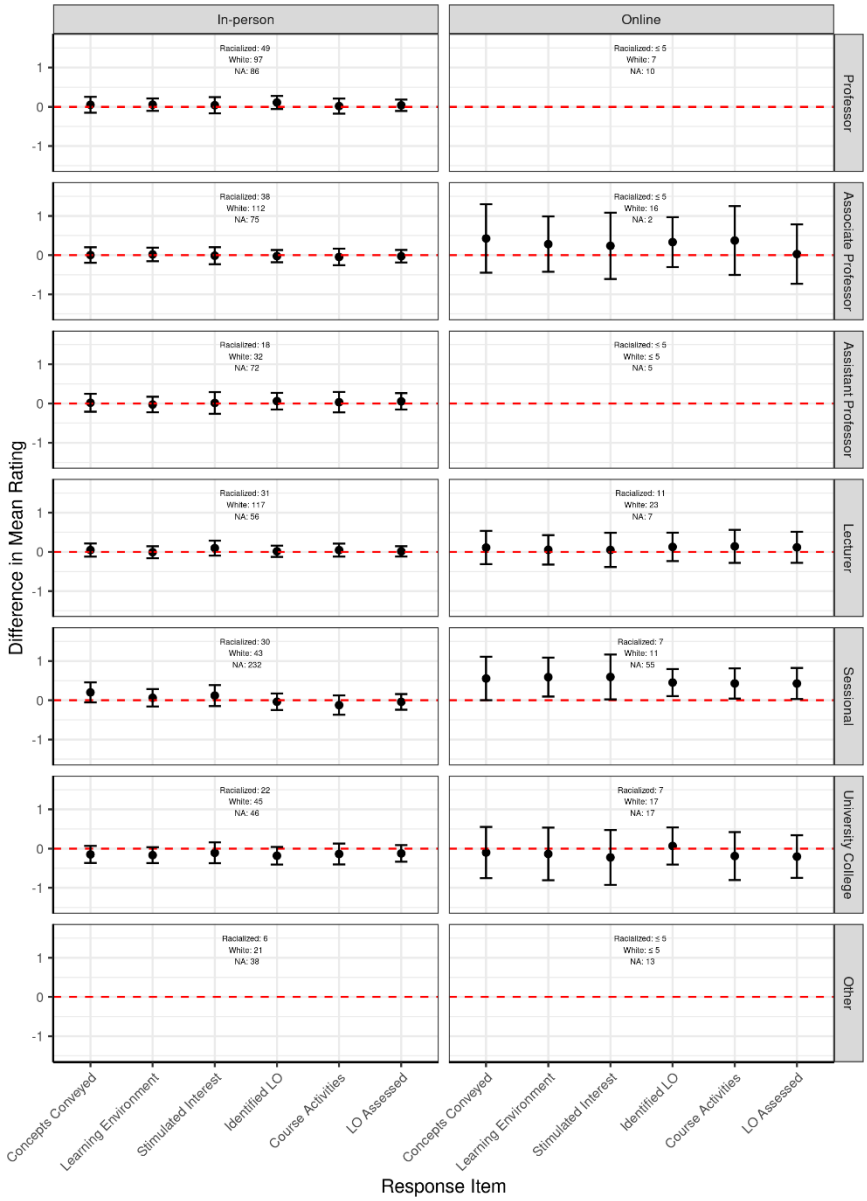

Figure 18: Difference in mean Winter 2023 ratings for white and racialized instructors by response item, instructor appointment type, and course type. Error bars correspond to 95% confidence intervals.

Difference in MR for White and Racialized Instructors by Class Size and Course Type

Figure 24 shows the difference in mean ratings given by students to racialized and white instructors accounting for course type (online or in-person) and class size. In all instances, differences observed are not statistically significant (as evidenced by the error bar crossing or touching the zero line). And in most cases, the non-significant difference is very small (as evidenced by the data points falling close to or on the zero line).

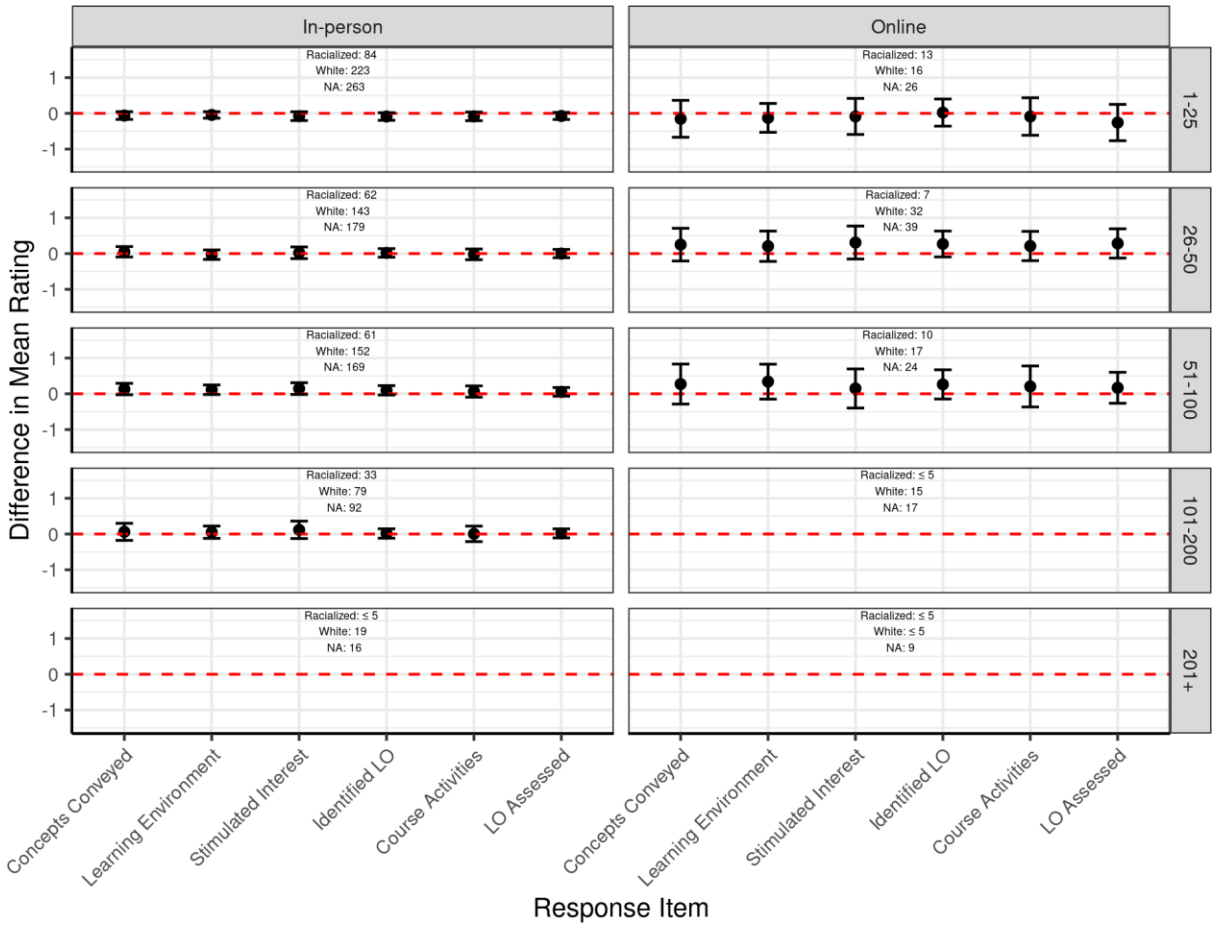

Figure 19: Difference in mean Winter 2023 ratings for white and racialized instructors by response item, class size, and course type. Error bars correspond to 95% confidence intervals.

Difference in MR for White and Racialized Instructors by Class Size

Figure 26 displays the difference in mean ratings given by students to white and racialized instructors by class size. In four of five class size categories (1-25, 26-50, 101-200, and 201+), differences observed are not statistically significant (as evidenced by the error bar crossing or touching the zero line). And in most cases, the non-significant difference is very small (as evidenced by the data points falling close to or on the zero line). However, in the fifth class size category (51-100), differences observed are statistically significant (as evidenced by the error bar not crossing or touching the zero line) for two items: Concepts_Conveyed, and Learning_Environment. For these two items, students in classes of 51-100 students rated white instructors somewhat higher (under 0.2) than racialized instructors. However, given the small number of courses with 51-100 students (n=520 or 22% of all included courses), attributing too much weight to this finding would not be prudent. The TAP office will continue to monitor these findings.

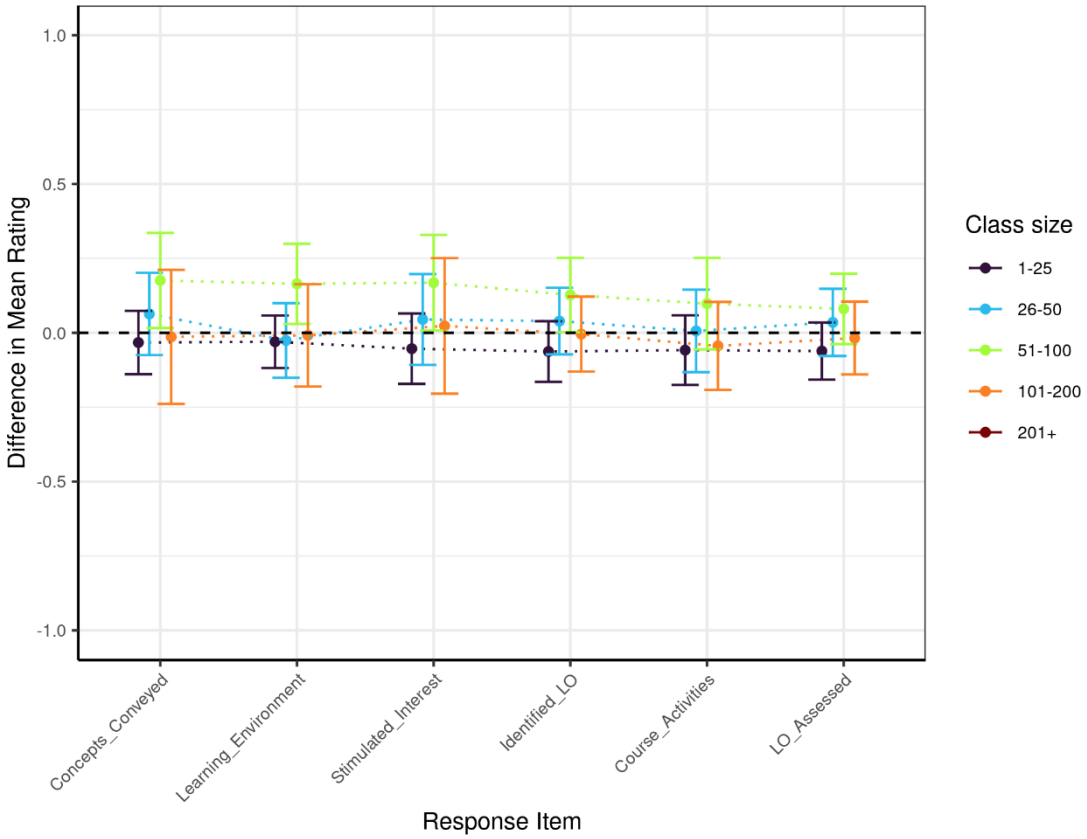

Figure 20: Difference in mean Winter 2023 ratings for white and racialized instructors by response item and class size. Error bars correspond to 95% confidence intervals. Note that there were few classes with 201+ students (n=57) included in this study, and therefore results are not plotted.

Difference in MR for White and Racialized Instructors by Instructor Time in Canada

Figure 28 provides a visual of the difference in mean ratings for all six items accounting for instructor racial identity and time in Canada. In all instances, differences observed are not statistically significant (as evidenced by the error bar crossing or touching the zero line). And in most cases, the non-significant difference is very small (as evidenced by the data points falling close to the zero line).

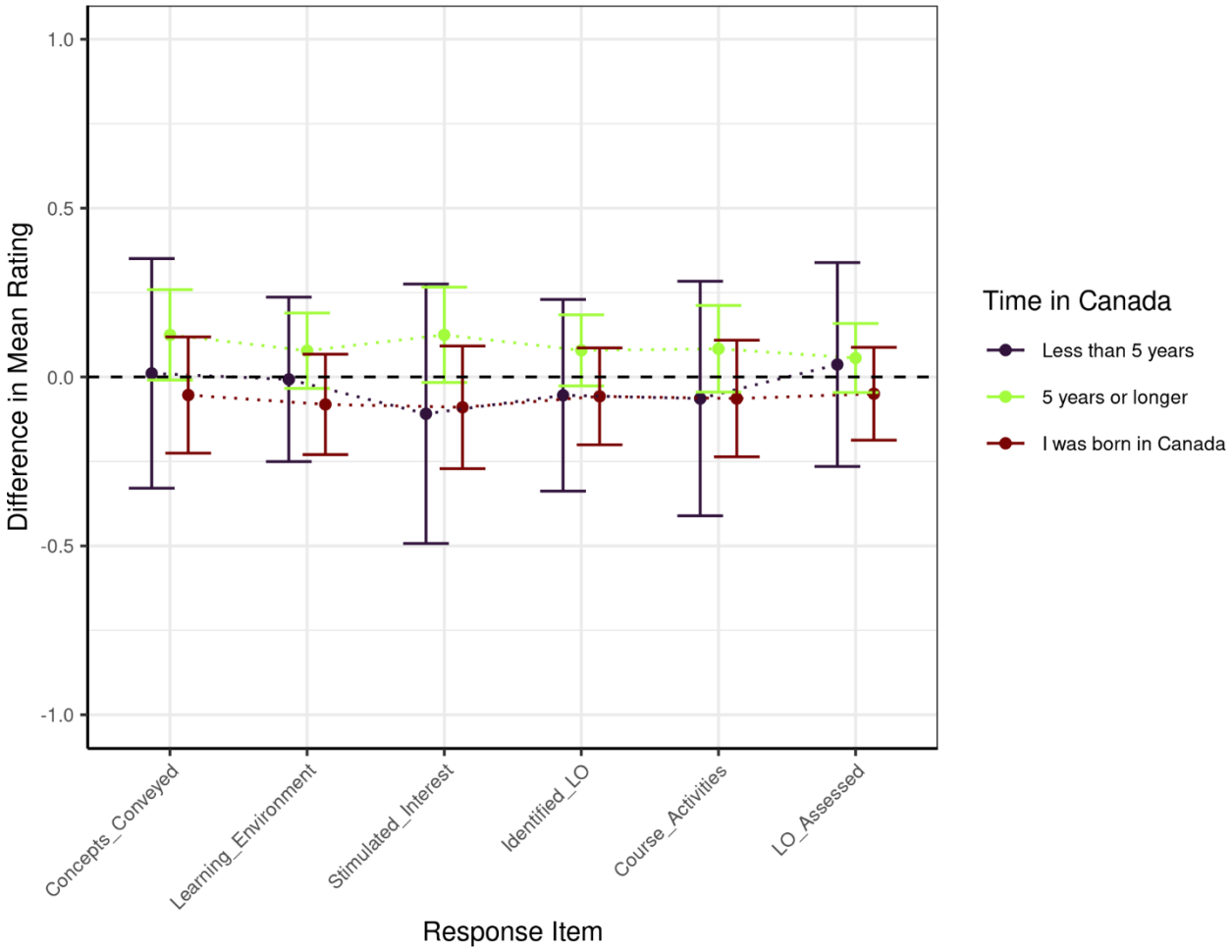

Figure 21: Difference in mean Winter 2023 ratings for white and racialized instructors by response item and instructor's time spent in Canada. Error bars correspond to 95% confidence intervals.

In the literature on student course evaluation surveys, there has been some evidence that instructor accent, among other things like cultural familiarity etc. may impact student perceptions of their course experience (Hamermesh & Parker, 2005; Ogier, 2005; Paquette-Smith et al, 2023). The consensus on whether this is a bias in and of itself, or if it is indeed an impediment to student learning, is contentious. Using the same logic we applied in the Winter 2022 analysis, we explored the possibility that respondent bias against instructors born outside of Canada may result in lower mean ratings. We still do not have data on instructor's first language, or a method to measure 'cultural familiarity' in the equity survey, so again we relied on "Time spent in Canada" (<5 years, 5+years, or born in Canada), an imprecise proxy measure. We ultimately collapsed this into a dichotomous variable by combining two categories (<5 years and 5+ years) into one, leaving us with Instructors Born in Canada and Instructors Born Outside Canada.

Difference in MR for Instructors Born in Canada and Born Outside Canada by Instructor Racial Identity

MR assigned to instructors born in Canada − MR assigned to instructors born outside Canada

Figure 30 shows the difference in mean ratings given by students to instructors born in or outside of Canada by instructor racial identity. In all instances, differences observed are not statistically significant (as evidenced by the error bar crossing or touching the zero line). And in most cases, the non-significant difference is very small (as evidenced by the data points falling close to the zero line). The large confidence intervals reflect the small sample size generated from this parsing of the data.

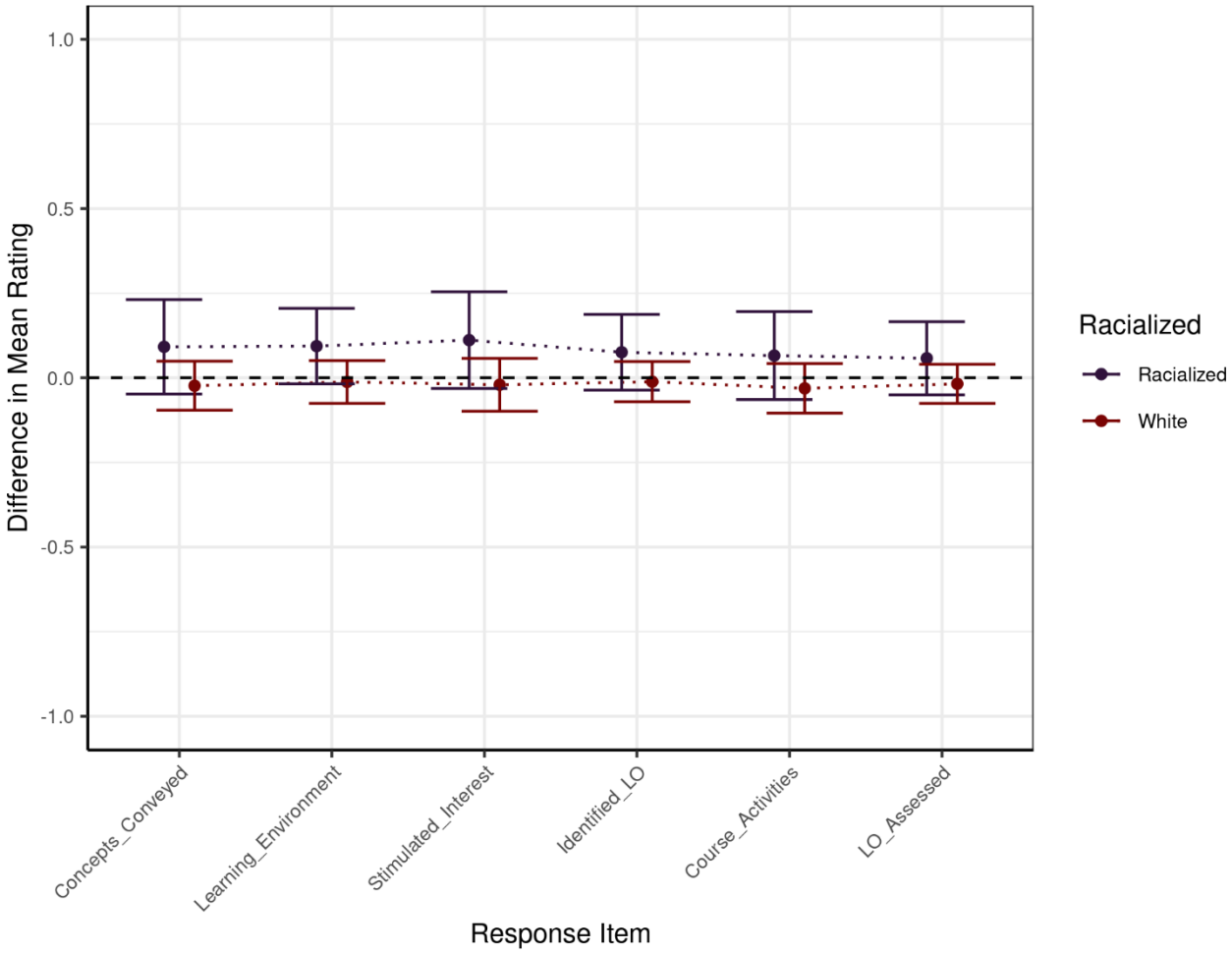

Figure 22: Difference in mean Winter 2023 ratings for instructors born in Canada and instructors born outside Canada by response item and instructor racial identity. Error bars correspond to 95% confidence intervals.

Difference in MR for Instructors Born in Canada and Instructors Born Outside Canada by Response Item

To further test the dichotomized 'Born in Canada' measure outlined above, we explored the difference in mean ratings assigned by students to instructors born inside and outside of Canada (see Figure 32). In all instances, differences observed are not statistically significant (as evidenced by the error bar crossing or touching the zero line). And in most cases, the non-significant difference is very small (as evidenced by the data points falling close to the zero line). This contrasts with our Winter 2022 analysis, which suggested differences in mean ratings that were significant for instructors born in Canada and instructors born outside Canada for two survey items: Concepts_conveyed and Simulated interest. However, it is still far too early to draw any conclusions from the data, and we will continue to investigate this relationship.

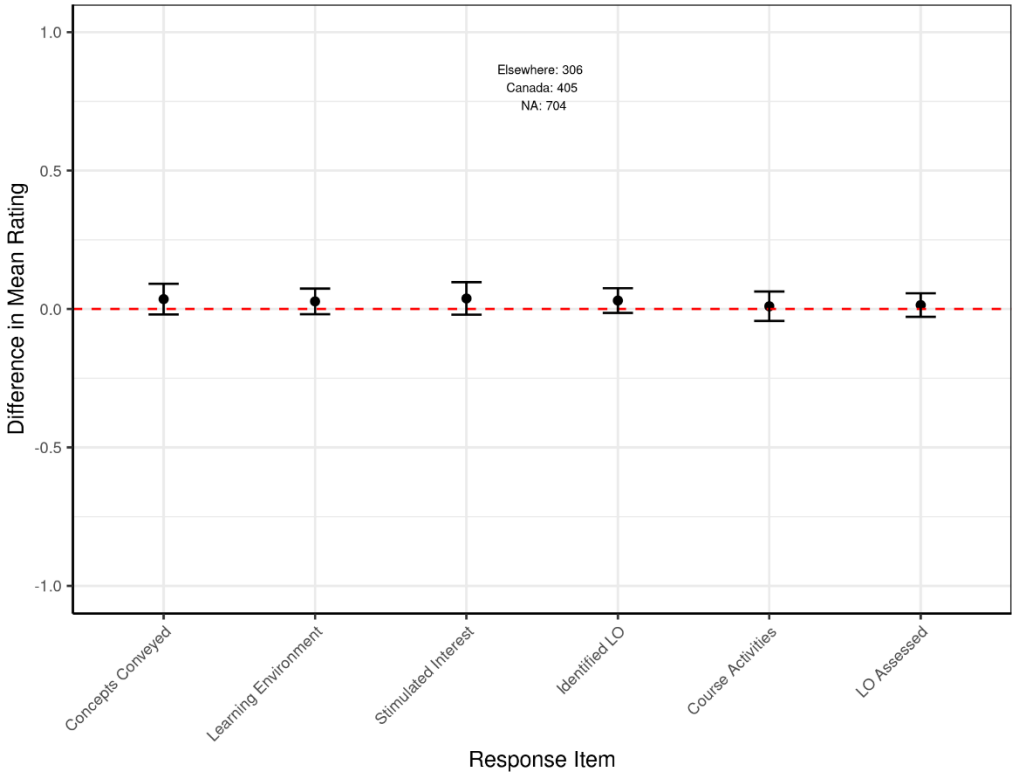

Figure 23: Difference in mean Winter 2023 ratings for instructors born in Canada and instructors born outside Canada by response item. Error bars correspond to 95% confidence intervals.

Analysis of Differences: Instructor Sex

The sex-based difference in ratings is calculated as follows:

MR assigned to male instructors – MR assigned to female instructors

Difference in MR for Male and Female Instructors

Figure 35 shows the difference in mean ratings assigned by student respondents to male and female instructors across all response items. In all instances, differences observed are not statistically significant (as evidenced by the error bar crossing or touching the zero line). And in most cases, the non-significant difference is very small (as evidenced by the data points falling close to the zero line).

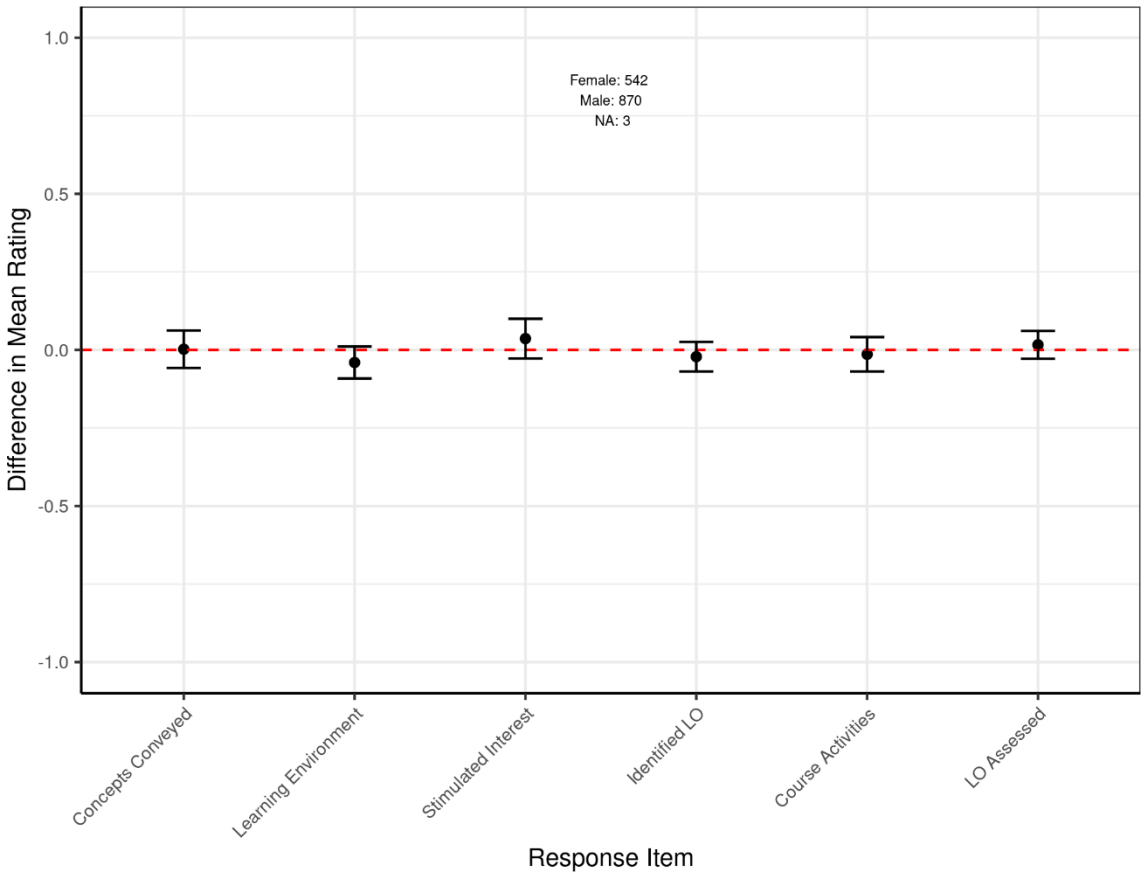

Figure 34: Difference in mean Winter 2023 ratings for male and female instructors per response item. Error bars correspond to 95% confidence intervals.

A clearer picture of the non-significant differences in mean ratings for male and female instructors is evident in Table 3, which also displays the confidence intervals for each response item. Again, across the six core items we can see how the difference in mean ratings on the five-point scale is very small (ranging from -0.04 to 0.2).

Table 2: Difference in mean Winter 2023 ratings for male and female instructors per response item, with 95% confidence interval (C.I.).

	Male Avg.	Female Avg.	Difference	Male C.I.	Female C.I.	Difference C.I.
Concepts Conveyed	4.17	4.17	0.00	(4.13, 4.21)	(4.13, 4.22)	(-0.06, 0.06)
Learning Environment	4.31	4.35	-0.04	(4.28, 4.34)	(4.31, 4.39)	(-0.09, 0.01)
Stimulated Interest	4.02	3.99	0.04	(3.98, 4.06)	(3.93, 4.04)	(-0.03, 0.1)
Identified LO	4.18	4.20	-0.02	(4.15, 4.21)	(4.16, 4.24)	(-0.07, 0.03)
Course Activities	4.02	4.04	-0.01	(3.99, 4.06)	(3.99, 4.08)	(-0.07, 0.04)
LO Assessed	4.17	4.15	0.02	(4.14, 4.19)	(4.11, 4.18)	(-0.03, 0.06)

Difference in MR Assigned to Male and Female Instructors by Faculty of Course Offering

We also explored mean ratings assigned by students to male and female instructors accounting for the Faculty of course offering (Figure 38). In all instances except one, differences observed are not statistically significant (as evidenced by the error bar crossing or touching the zero line). And in most cases, the non-significant difference is very small (as evidenced by the data points falling close to the zero line). The only finding that displays statistical significance is for Stimulated_Interest in Arts, where the error bar does not cross the zero line. The data point for this item indicates that student ratings for this question when they have a male instructor is higher (about 0.1) than when they have a female instructor. Other than this single response item in the Arts faculty, we see a close clustering of scores on the 0 line across remaining Faculties and all response items.

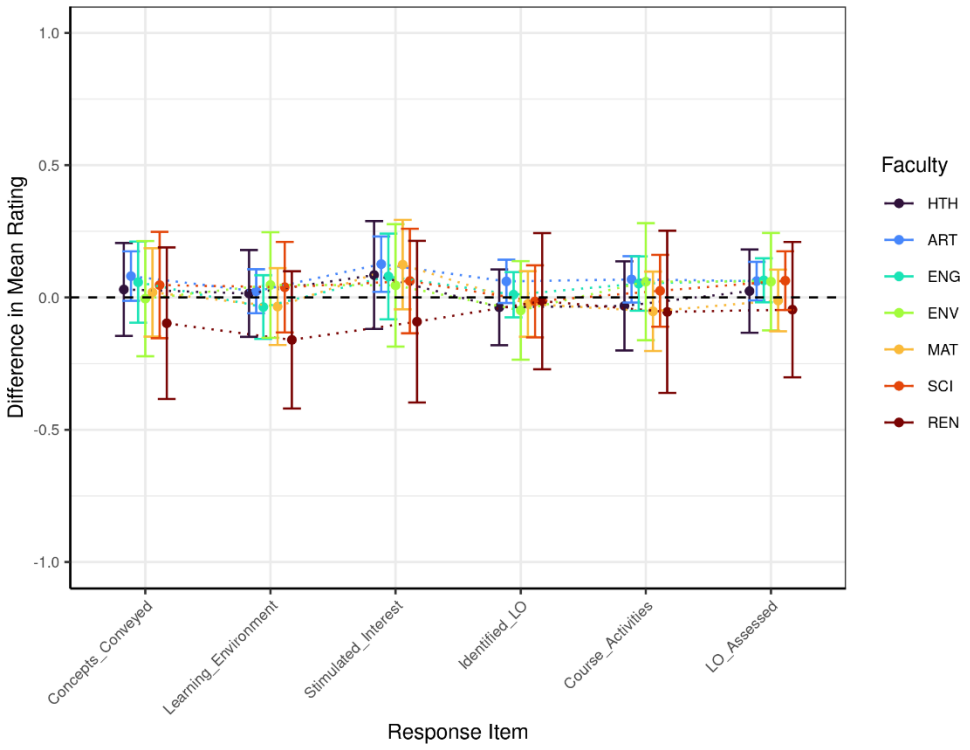

Figure 36: Difference in mean Winter 2023 ratings for male and female instructors by response item and Faculty of course offering. Error bars correspond to 95% confidence intervals.

Difference in MR for Male and Female Instructors by Faculty of Course Offering and Class Size

We also examined mean ratings at the Faculty level for male and female instructors accounting for class size (Figure 39). In all instances except one, differences observed are not statistically significant (as evidenced by the error bar crossing or touching the zero line). And in most cases, the non-significant difference is very small (as evidenced by the data points falling close to the zero line). Not surprisingly given the findings noted in the previous section, the one finding that displays statistical significance is for Stimulated_Interest in Arts in classes of 51-100, where the error bar does not cross the zero line. The data point for this item indicates that student ratings for this question are higher (about 0.2) for a class of 51-100 students in Arts when they have a male instructor as opposed to when they have a female instructor. Other than this single response item in the Arts faculty for classes of this specific size, any differences in mean ratings for male and female instructors are not statistically significant when we account for class size and Faculty.

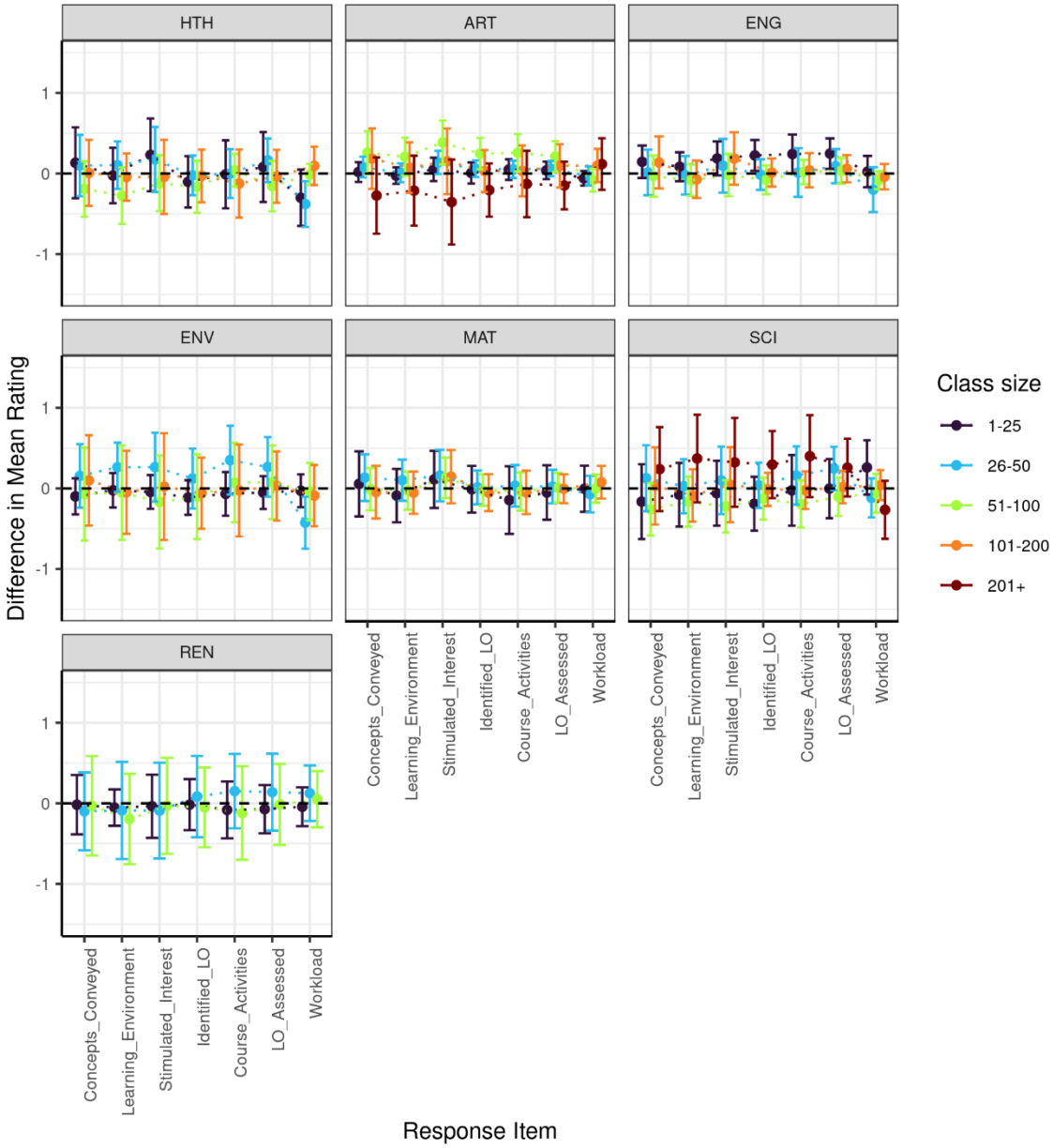

Figure 38: Difference in mean Winter 2023 ratings for male and female instructors by response item, class size, and Faculty of course offering. Error bars correspond to 95% confidence intervals.

Difference in MR for Male and Female Instructors by Faculty of Course Offering and Course Type

Figure 41 depicts the difference in mean ratings for male and female instructors at the Faculty level accounting for course type (Online or In-person). These results should be considered within the context of the sample sizes. If we refer back to Figure 10, we can see that the majority of courses offered in Winter 2023 were in-person, across all Faculties. In all instances, differences observed are not statistically significant (as evidenced by the error bar crossing or touching the zero line). And in most cases, the non-significant difference is very small (as evidenced by the data points falling close to the zero line). Among these non-significant

differences, findings in two Faculties (Engineering and Science) suggest that these students tend to give higher ratings to female online instructors than male online instructors. However, it's important to consider the small number of online courses in these Faculties (Engineering = 14 and Science = 18).

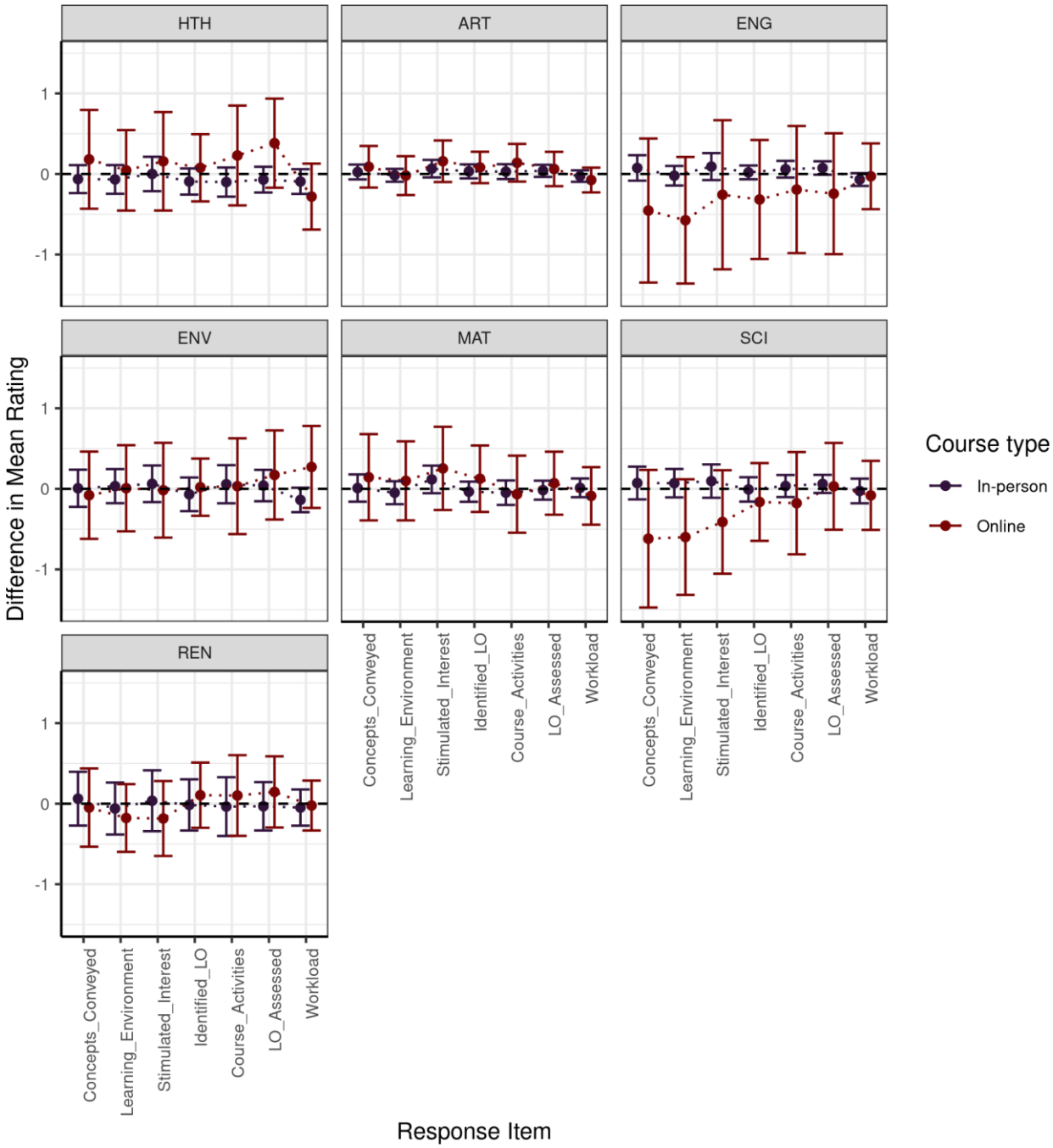

Figure 40: Difference in mean Winter 2023 ratings for male and female instructors by response item, course type, and Faculty of course offering. Error bars correspond to 95% confidence intervals.

Difference in MR for Male and Female Instructors by Faculty of Course Offering and Instructor Appointment

At the Faculty level we also examined the difference in mean ratings for male and female instructors accounting for instructor appointment type (Figure 43). In all instances except one, differences observed are not statistically significant (as evidenced by the error bar crossing or touching the zero line). And in most cases, the non-significant difference is very small (as evidenced by the data points falling close to the zero line). The one finding that displays statistical significance is for Stimulated_Interest in Math classes taught by Lecturers, where the error bar does not cross the zero line. The data point for this item indicates that student ratings for this question are slightly higher (under 0.2) when they have a male instructor as opposed to when they have a female instructor. Other than this single response item in the Math faculty for Lecturers, differences in mean ratings for male and female instructors are not statistically significant when we account for class size and Faculty.

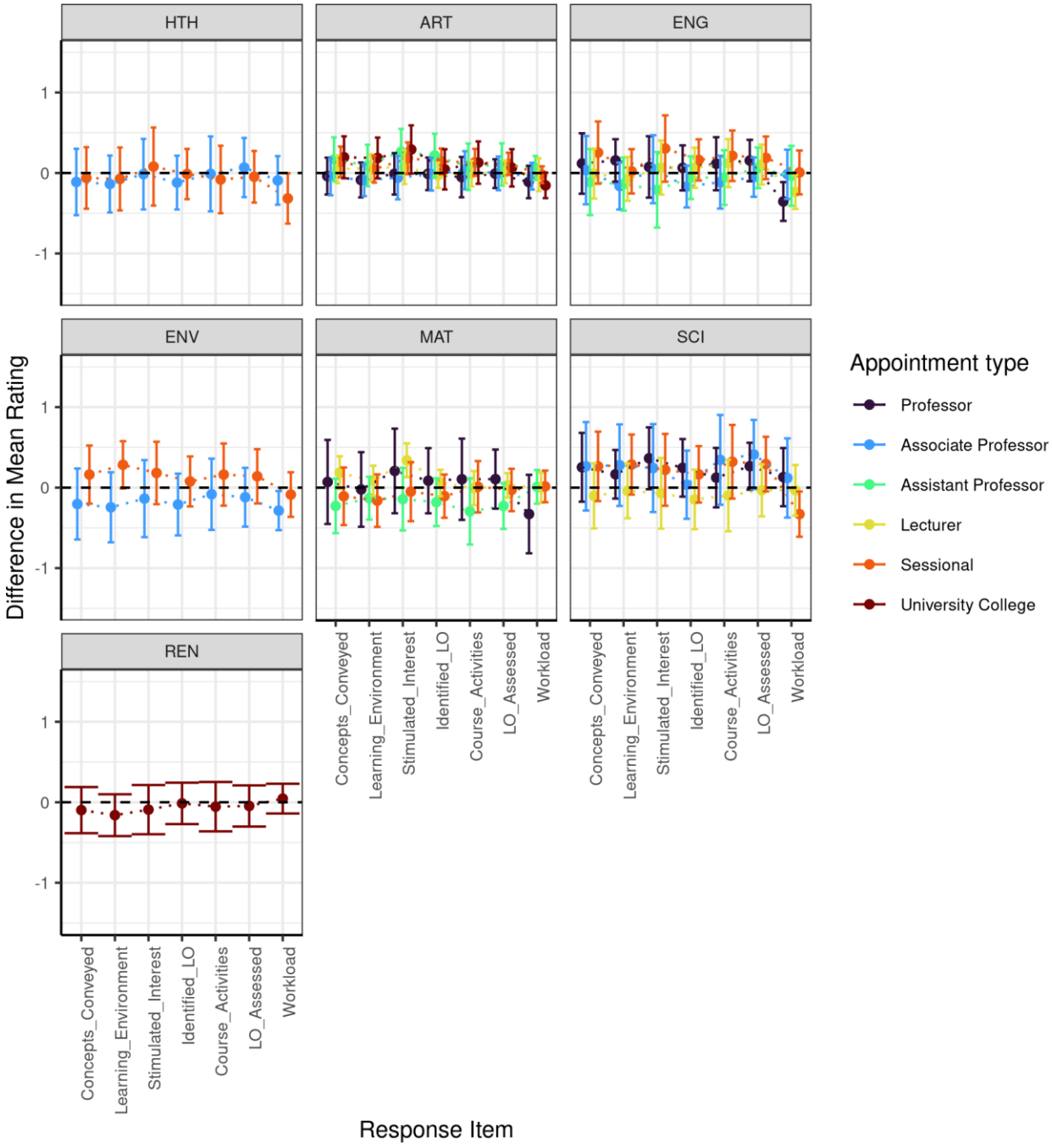

Figure 42: Difference in mean Winter 2023 ratings assigned for male and female instructors by response item, instructor appointment type, and Faculty of course offering. Error bars correspond to 95% confidence intervals.

Difference in MR for Male and Female Instructors by Class Size

In Figure 45 the difference in mean ratings for male and female instructors across all six survey items is plotted, accounting for class size. It's noteworthy that for larger classes (101-200 and 200+), the sample size is small, which is reflected by the larger confidence intervals; these should be interpreted with extra caution.

In all instances except one, differences observed are not statistically significant (as evidenced by the error bar crossing or touching the zero line). And in most cases, the non-significant difference is very small (as evidenced by the data points falling close to the zero line). The one finding that displays statistical significance is for LO_Assessed in classes of 26-50 students, where the error bar does not cross the zero line. The data point for this item indicates that student ratings for this question are slightly higher (under 0.1) for a class of 26-50 students when they have a male instructor as opposed to when they have a female instructor.

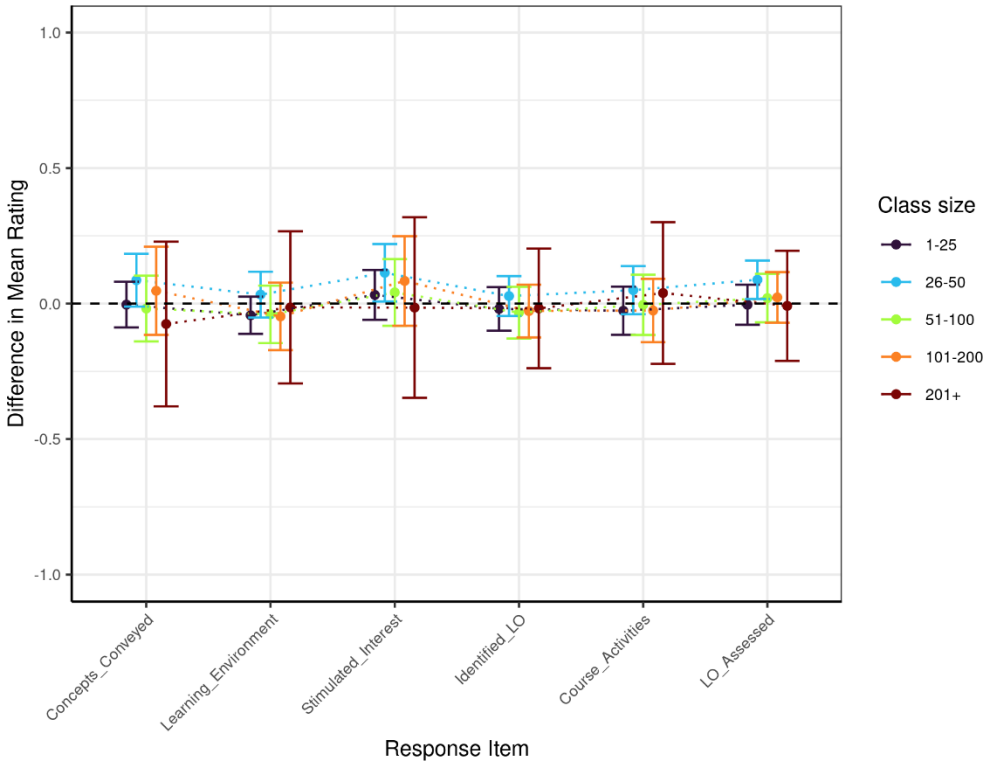

Figure 44: Difference in mean Winter 2023 ratings for male and female instructors by response item and class size. Error bars correspond to 95% confidence intervals.

Difference in MR for Male and Female Instructors by Instructor Time in Canada

Figure 47 displays the difference in mean ratings for male and female instructors by instructor time in Canada. For both groups of instructors who report not being born in Canada (Less than 5 years and 5 years or longer), we observe extremely large error bars reflective of the small number of Winter 2023 instructors included in this study who reported not being born in Canada (n=306).

Within the group of instructors who report being in Canada less than 5 years, we observe that students rated female instructors slightly higher (0.1-0.2) across most response items than male instructors; however, this difference is not statistically significant (as evidenced by the error bars crossing the zero line).

For instructors who were not born in Canada, but who have been in Canada 5 years or longer, we observe statistically significant findings (as evidenced by the error bars not crossing the zero line) for the following five response items: Concepts_Conveyed, Learning_Environment, Stimulated_Interest, Identified_LO, and Course_Activities. Across these items, students assigned female instructors who have been in Canada 5 years or longer higher ratings (less than 0.25 higher) than male instructors who have been in Canada 5 years or longer.

For instructors who were born in Canada, we observe that the difference in ratings assigned by students for male and female instructors is closer to 0 than for the other groups.

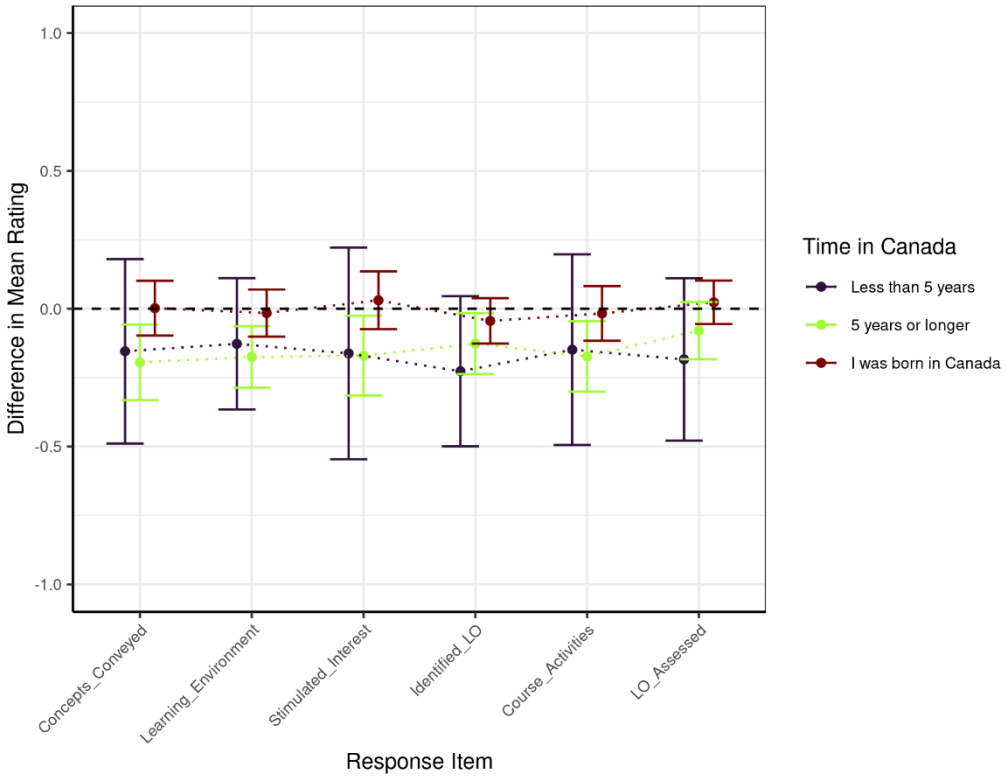

Figure 24: Difference in mean Winter 2023 ratings for male and female instructors by response item and instructor's time spent in Canada. Error bars correspond to 95% confidence intervals.

Difference in MR for Male and Female Instructors by Class Size and Instructor Appointment Type

Figure 49 displays the difference in mean ratings assigned by student respondents to female and male instructors accounting for instructor appointment type and class size. In general, these differences are very close to 0, and across nearly all cases, they are <0.5 points on the 5-point Likert scale. Also noteworthy is the fact that these differences are not statistically significant. This is evidenced by the confidence intervals crossing the 0 line for all plots, with the exception of one item in one specific circumstance (discussed below).

Student respondents in classes with 101-200 students rated male Professors about 0.5 points higher than female Professors for LO_Assessed, and this difference is statistically significant.

However, given the small sample size (n= 6 female Professors and n= 27 male Professors), attributing too much weight to this finding would not be prudent. The TAP office will continue to monitor these findings.

Also worth highlighting is that although previous analyses (Winter 2022 & Pilot 2018) showed that student respondents in classes with 101-200 students rated male Lecturers about 0.5 points higher than female Lecturers, we did not observe this trend in the Winter 2023 analysis.

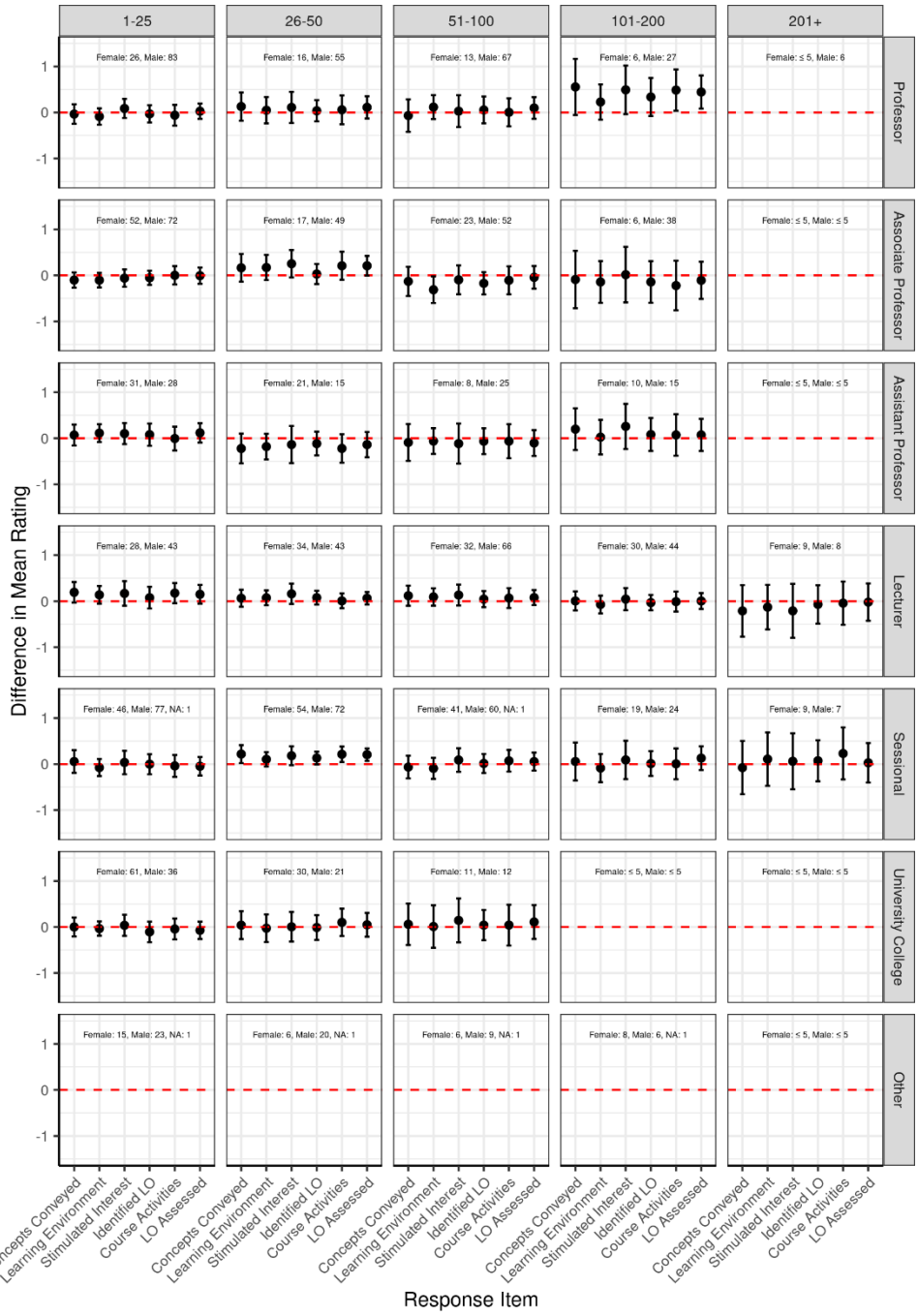

Figure 49: Difference in mean Winter 2023 ratings for male and female instructors by response item, instructor appointment type, and class size. Error bars correspond to 95% confidence intervals.

Difference in MR for Male and Female Instructors by Course Type, and Instructor Appointment Type

Figure 51 shows the difference in mean ratings for female and male instructors, accounting for course type (Online or In-person) and instructor appointment type. In all instances, differences observed are not statistically significant (as evidenced by the error bar crossing or touching the

zero line). And in most cases, the non-significant difference is very small (as evidenced by the data points falling close to the zero line). The larger error bars in the plots for Online courses taught by either Professors or Associate Professors are reflective of the small number of these types of cases.

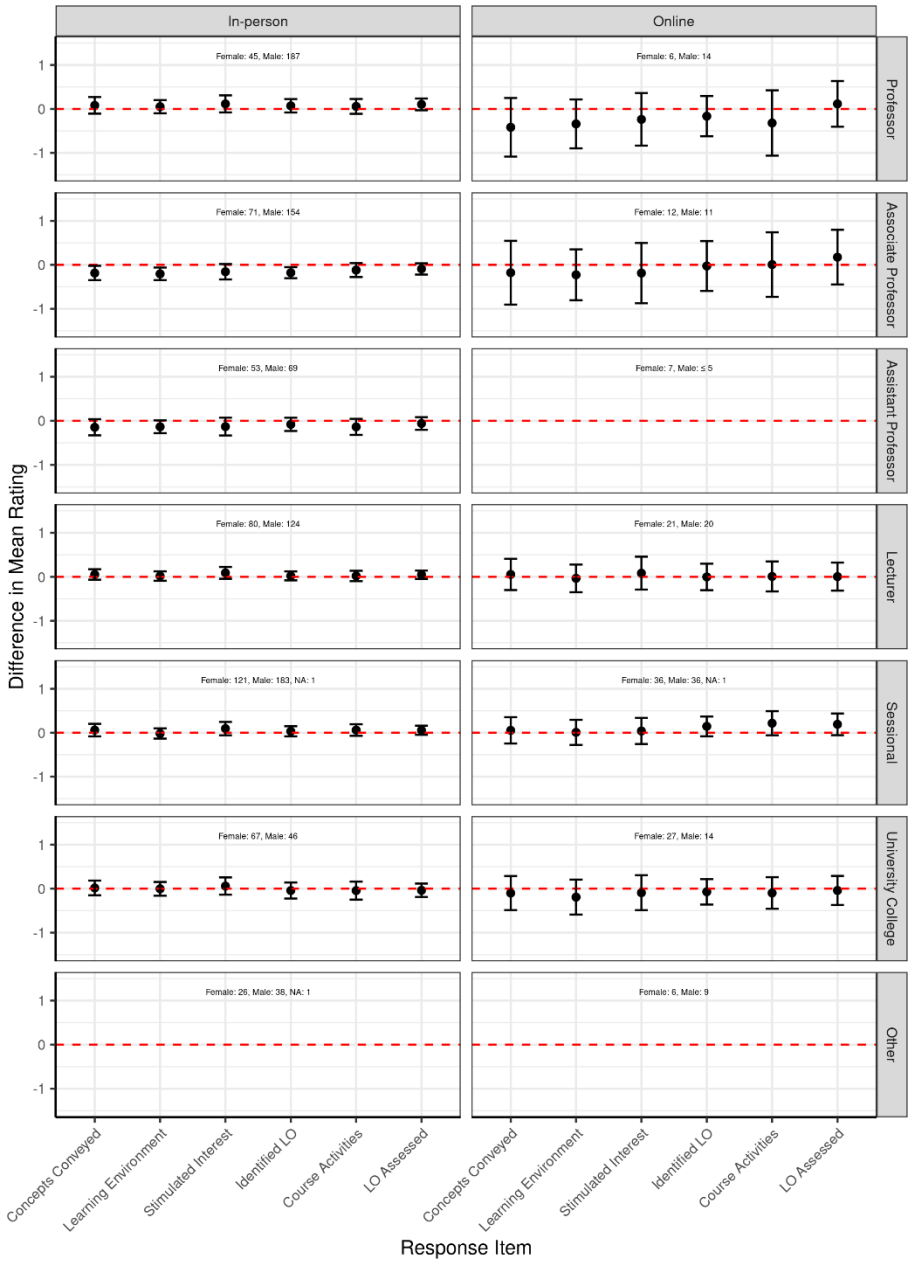

Figure 51: Difference in mean Winter 2023 ratings for male and female instructors by response item, instructor appointment type and course type. Error bars correspond to 95% confidence intervals.

Difference in MR for Male and Female Instructors by Course Type and Class Size

Figure 53 shows the differences in mean ratings for female and male instructors accounting for course type (Online or In-person) and class size. In all instances, differences observed are not statistically significant (as evidenced by the error bar crossing or touching the zero line). And in most cases, the non-significant difference is very small (as evidenced by the data points falling close to the zero line).

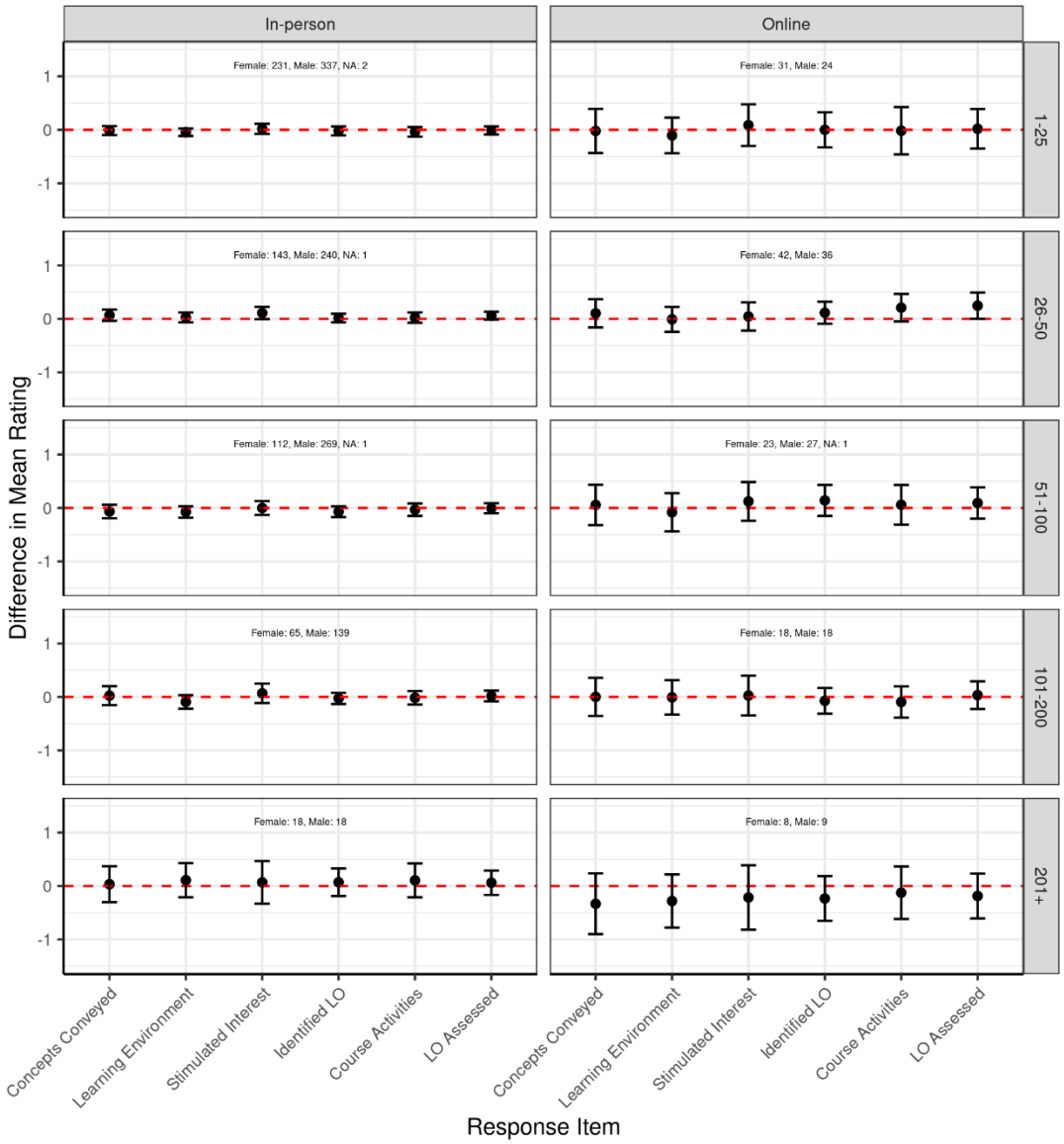

Figure 53: Difference in mean Winter 2023 ratings for male and female instructors by response item, class size, and course type. Error bars correspond to 95% confidence intervals.

Recap of Main Findings

This analysis involved numerous statistical tests to examine how various instructor characteristics (racial identity, time in Canada, and appointment type) and course attributes (class size, Faculty of course offering, course type) impacted Winter 2023 student responses to SCP surveys. The following key findings emerged from this analysis:

- Nearly half (45.7%) of Winter 2023 instructors did not complete the Equity Survey. As a result, any analyses based on racial or Indigenous identity are based on incomplete data, and therefore must be interpreted with great caution.
- We were unable to meaningfully examine the impact of instructor Indigenous identity on mean ratings due to the small number of Winter 2023 instructors who indicated Indigenous identity in the Equity Survey (n=8).
- The dataset included in this analysis contained substantially more white instructors than racialized instructors across all Faculties, and only 32 of the instructors who responded to the Equity Survey reported being in Canada <5 years.
- There were more male instructors than female instructors across all appointment categories.
- Most of the Winter 2023 courses included in this analysis (89%) were delivered in person, as opposed to online.
- The bulk of Winter 2023 courses included in this analysis (86%) had 100 or fewer students.
- The smaller Winter 2023 courses included in this analysis tended to receive higher mean ratings than the larger courses.
- Consistent with research on this topic in general, mean ratings across all six response items included in our dataset clustered at the higher-end of the five-point scale. In other words, Winter 2023 students tended to assign scores between 4-5 on the five-point scale across all SCP survey items. This finding was true across each Faculty and Renison.
- Overall, we found two statistically significant differences in mean ratings between racialized and white instructors. However, given the small sample sizes in each, attributing too much weight to these findings would not be prudent. The TAP office will continue to monitor these findings.
- The Winter 2023 analysis did not provide evidence to suggest time spent in Canada, a proxy variable used to examine possible language and/or cultural biases, is associated with lower scores assigned by students.
- The small differences observed in Winter 2023 mean ratings for male and female instructors was found to be statistically non-significant accounting for class size, course type, or instructor appointment.
- The overall difference in mean ratings assigned by students to male and female instructors was not statistically significant accounting for course type, class size, or instructor appointment type. Analysis did find that in classes of 101-200 students,

female professors received lower scores than male professors teaching classes of the same size (difference of 0.05 or less) across all six survey items, with only one of those differences being statistically significant (LO_Assessed). It is important to note the sample size for this group was very small (female = 6, male = 27), and therefore caution is necessary when drawing conclusions for this finding, but we will continue to investigate this finding in future analyses.

Concluding Remarks

The Winter 2023 analysis of the SCP survey data is the second to explore the impact of systemic bias on SCP ratings in the Waterloo context. The TAP office remains committed to exploring these issues in-depth so we can understand how SCP ratings are impacted by bias; a third study is planned for Winter 2024 data. Given the cross-sectional nature of this dataset, and especially in light of the unreliability of this dataset due to the large number of Winter 2023 instructors who did not complete the Equity Survey, it is ill advised to draw firm conclusions from this study or take any sort of action at this time. This study does, however, suggest important lines of inquiry the TAP office will pursue in the future that may lead to action.

- Overall, the findings continue to emphasize that comparison of decimal-point differences in ratings is a problematic practice that should be avoided. A stark reminder that decisions made on the basis of decimal-point differences in ratings are likely to bolster (rather than combat) systemic bias.
- Racial disparity in Faculty teaching appointments is not unique to Waterloo. According to Connor (2022):

> In recent years, many colleges and universities have begun to increase efforts to improve diversity, equity, and inclusion (DEI) […] Although many institutions, […] have improved their efforts, retaining Black faculty continues to be a significant challenge. National data show glaring disparities in this trend. According to the National Center for Education Statistics, Black faculty comprise roughly 6 percent of all faculty in colleges and universities, despite representing 13 percent of the national population. Eight percent of junior faculty are Black, with 5 percent of assistant professors being Black women and 3 percent being Black men. These numbers drop as professorial ranks increase. Only 4 percent of full professors are Black, evenly split at 2 percent each for Black men and women. This drop-off is especially significant for Black women, who comprise 5 percent of assistant professors compared to 3 percent for Black men (American Sociological Association, 2022).[2]

- Mirroring what we observed in the 2018 pilot test, and the Winter 2022 analysis, males included in this study continue to outnumber females quite substantially across all

[2] Source: https://www.asanet.org/footnotes-article/black-faculty-and-radical-retention/#:~:text=According%20to%20the%20National%20Center,percent%20of%20the%20national%20population.

instructor types. The disparity we observe in Faculty teaching appointments is also not unique to Waterloo, as research has long-documented, particularly in higher education, that women are overrepresented in the lowest ranks and often among the most unstable faculty positions, including non-tenure track roles and adjunct positions (AAUW, 2020; August & Waltman, 2004; Equal Rights Advocates, 2003).

- Our work continues to be impacted by the large number of Winter 2023 instructors (45.7%) who did not complete the Equity Survey. It is our hope that future such analyses will benefit from higher Equity Survey response rates.

Limitations

As outlined above, 646 of the 1415 Winter 2023 instructors in this study did not complete the Equity Survey. This translates to 45.7% missing cases for the equity data. As such, any findings based on this incomplete dataset must be interpreted with a high degree of caution.

Though some of the deeper analyses of specific instructor groups and mean ratings did not reveal high instances of statistically significant differences in scores for racialized and white instructors this should not be interpreted to suggest bias in scores does not exist. The lack of significant findings may result from incomplete Equity Survey data, or it may suggest an alternative explanation for the differences in mean ratings. Or it may simply be that the differences are small, and the smaller sample sizes of the deeper analyses did not have the power to deem the differences statistically significant. Either way, further analysis is necessary and remains a priority for the TAP office.

A further note on the limitations to this analysis also relates to the high number of missing cases for the equity data, which prohibits us from exploring deeper intersectional analyses for key variables like race and gender (e.g., what differences in scores are observed between racialized women versus white men). The TAP office hopes to explore intersectional analyses in the future.

References

August, L., & Waltman, J. (2004). Culture, climate, and contribution: Career satisfaction among female faculty. Research in Higher Education, 45(2), 177–192. https://doi.org/10.1023/b:rihe.0000015694.14358.ed

Buser, W., Batz-Barbarich, C.L. & Kearns, J.K (2023). Evaluation of Women in Economics: Evidence of Gender Bias Following Behavioral Role Violations. *Sex Roles,* 86: 695-710 *https://doi.org/10.1007/s11199-022-01299-w*

Chevalier, J. (2020, December 16). The 2020 report of the committee on the status of women in the economics profession. American Economic Association. AAUW

Connor, T. (2022). Black Faculty and Radical Retention. *Footnotes: A Magazine of the American Sociological Association*, 50(2) https://www.asanet.org/footnotes-article/black-faculty-and-radical-retention/#:~:text=According%20to%20the%20National%20Center,percent%20of%20the%20national%20population.

Equal Rights Advocates. (2003). Creating gender equity in academia: Equal rights advocates' higher education legal advocacy project roundtable report. University Women http://universitywomen.law.stanford.edu/reports/Creating_gender_equity_academia.pdf

Hamermesh, D.S. and Parker, A. (2005), "Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity", *Economics of Education Review*, 24(4): 369-376.

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.

Ogier, J. (2005) Evaluating the effect of a lecturer's language background on a student rating of teaching form, *Assessment & Evaluation in Higher Education*, 30:5, 477-488, DOI: 10.1080/02602930500186941

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

Paquette-Smith, M., Buckler, H. & Johnson, E.K. (2023). How Sociologistic Factors Shape Children's Subjective Impressions of Teacher Quality. *Quarterly Journal of Experimental Psychology,* 76(3): https://doi.org/10.1177/17470218221094312

Wang, L., & Gonzalez, J.A. (2020). Racial/ethnic and national origin bias in SET. International Journal of Organizational Analysis, 28 (4): 843-855. DOI 10.1108/IJOA-06-2019-1793

Appendix A: Coding Scheme

Variable	Variable label	Values included in the corresponding label	Additional note
Indigenous identity	Indigenous	Winter 2023 instructors who selected "Yes" for the Equity Survey question, "Do you identify as an Indigenous person?"	
	Not indigenous	Winter 2023 instructors who selected "No" for the Equity Survey question, "Do you identify as an Indigenous person?"	
	NA	Winter 2023 instructors who either left blank or selected "I prefer not to answer" for the Equity Survey question, "Do you identify as an Indigenous person?"	
Self-reported racial identity	Racialized	Winter 2023 instructors who selected "Yes" for the Equity Survey question, "Do you identify as an Indigenous person?" and/or selected at least one of the following for the Equity Survey question, "Please select the racial category or categories with which you primarily identify." Black e.g., African, Caribbean, Black Canadian, Afro-Latine, African American, or other African descent East Asian e.g., Chinese, Korean, Japanese, or other East Asian descent Latine e.g., Latin American,	

		Hispanic descent Middle Eastern e.g., Afghan, Egyptian, Iranian, Lebanese, Turkish, Kurdish, or other Arab or Persian descent South Asian e.g., East Indian, Pakistani, Bangladeshi, Sri Lankan, Indo-Caribbean, or other South Asian descent Southeast Asian e.g., Filipino, Vietnamese, Cambodian, Thai, Malaysian, Indonesian, or other Southeast Asian descent Another race category (please specify)	
	White	Winter 2023 instructors who did not select "Yes" for the Equity Survey question, "Do you identify as an Indigenous person?" and/or selected only the following for the Equity Survey question, "Please select the racial category or categories with which you primarily identify." White e.g., British, German, Ukranian, or other European descent	
	NA	Winter 2023 instructors who either left blank or selected "I prefer not to answer" for the Equity Survey question, "Please select the racial category or categories with which you primarily identify."	
Time in Canada	Less than 5 years	Winter 2023 instructors who selected one of the following for the Equity Survey question, "How	

		long have you lived in Canada?" Less than 1 year 1 year to less than 3 years 3 years to less than 5 years	
	Born in Canada	Winter 2023 instructors who selected the following for the Equity Survey question, "How long have you lived in Canada?" I was born in Canada	
	Born Elsewhere	Winter 2023 instructors who selected one of the following for the Equity Survey question, "How long have you lived in Canada?" Less than 1 year 1 year to less than 3 years 3 years to less than 5 years 5 years or longer	
	NA	Winter 2023 instructors who either left blank or selected "I prefer not to answer" for the Equity Survey question, "How long have you lived in Canada?"	
Sex	Male	Winter 2023 instructors for whom Human Resources data on sex indicated male	This data is best described as "sex at birth" and does not necessarily reflect instructor gender identity at the present time.
	Female	Winter 2023 instructors for whom Human Resources data on sex indicated female	
Instructor group	Professor	Winter 2023 instructors for whom	In the case

		Human Resources data on primary job profile or additional job profile indicated “Professor”	where a Winter 2023 instructor belonged to more than one instructor group, the order of overriding is “Professor, Associate Professor, Assistant Professor, Lecturer, Sessional, University College”. That is, if an instructor has “Assistant Professor” in their primary job profile and “Conrad Grebel Staff 9120”, they are assigned to the Assistant Professor group.
	Associate Professor	Winter 2023 instructors for whom Human Resources data on primary job profile or additional job profile indicated “Associate Professor”	
	Assistant Professor	Winter 2023 instructors for whom Human Resources data on primary job profile or additional job profile indicated “Assistant Professor”	
	Lecturer	Winter 2023 instructors for whom Human Resources data on primary job profile or additional job profile indicated “Lecturer” or “Clinical Lecturer”	
	Sessional	Winter 2023 instructors for whom Human Resources data on primary job profile or additional job profile indicated "Special (Sessional) Faculty", "Adjunct Lecturer", "Instructor III 5380", "Instructor IV 5385", "Instructor V 5386", "Instructor I 5360", "Adjunct Professor", "Adjunct Associate Professor", "Adjunct Assistant Professor", "Research Assistant Professor", "Research Professor", "Research Associate 1560", "Post-Doctoral Fellow (International) 1566", "Post-Doctoral Fellow (Domestic) 1565", "Graduate Teaching Assistant 1671", "Graduate Research Assistant 1672", "Visiting Lecturer", or "Visiting Scholar"	

	University College	Winter 2023 instructors for whom Human Resources data on primary job profile or additional job profile indicated "Renison Coll Faculty - Monthly 9150", "Renison Coll Staff - Monthly 9170", "Renison Coll Casual 9180", "St Jeromes Faculty 9190", "St Jeromes Staff 9200", "Conrad Grebel Faculty 9110", "Conrad Grebel Casual 9130", "Conrad Grebel Staff 9120", "St Pauls Coll Faculty 9210", "St Pauls Coll Faculty 9210", "St Pauls Coll Staff - Monthly 9230"	
Response item	Concepts_Conveyed	Student responses to the SCP Survey item, “The instructor(s) helped me to understand the course concepts.”	All response items are rated by students on a 5-point scale as follows: Strongly Disagree Disagree Neutral Agree Strongly Agree Strongly Disagree Each also includes a non-response item, No Basis for Rating
	Learning_Environment	Student responses to the SCP Survey item, “The instructor(s) created a supportive environment that helped me learn (Supportive environments enable students to feel included and valued regardless of any aspect of their identity).”	
	Stimulated_Interest	Student responses to the SCP Survey item, “The instructor(s) stimulated my interest in this course.”	
	Identified_LO	Student responses to the SCP Survey item, “The intended learning outcomes were identified (Learning outcomes/objectives articulate what students should be able to know, do, and/or value by the end of a course).”	

	Course_Activities	Student responses to the SCP Survey item, “The course activities prepared me for the graded work.”	
	LO_Assessed	Student responses to the SCP Survey item, “The intended learning outcomes were assessed through my graded work.”	
Faculty of course offering	HTH	Winter 2023 courses offered by the Faculty of Health	
	ART	Winter 2023 courses offered by the Faculty of Arts	
	ENG	Winter 2023 courses offered by the Faculty of Engineering	
	MAT	Winter 2023 courses offered by the Faculty of Mathematics	
	SCI	Winter 2023 courses offered by the Faculty of Science	
	REN	Winter 2023 courses offered by Renison University College	
	Other	Winter 2023 courses offered by All other University Colleges other than Renison	
Course type	In-person	Courses whose campus code belongs to one of the following: "UW", "REN", "STJ", "STP", "CGC", "OFF", "BLND", "BLNDJ", "BLNDR"	
	Online	Courses whose campus code belongs to one of the following: "ONLN", "ONLNG", "ONLNJ”, “ONLNP", "ONLNR"	

Appendix B: Model Description

Decisions around modeling were grounded in both statistical principles as well as practical and logistical considerations that made sense for the data and research questions posed.

The following model was used to conduct the MR analysis and not the descriptive analysis of demographics and course characteristics. Let y_{ijk} denote the score given by student k on instructor i teaching course-section j in a subgroup of interest, e.g., for all instructors in the Math faculty, or all lecturers teaching a class of more than 200 students. We adopt a mixed-effects model given by

$$y_{ijk} = \beta_0 + \beta_1 X_i + \mu_i + \alpha_j + \epsilon_{ijk},$$

Where:

- β_0 is the intercept or the baseline for a (sub)population
  - X_i is an instructor-level covariate of interest like sex or race,
  - β_1 is the difference in mean scores between the two levels of the covariate of interest, e.g., mean difference in score comparing male instructors to female instructors
    - μ_i is the instructor-level random effect, which is modelled as

$$\mu \overset{iid}{\sim} Normal(0, \sigma_\mu^2)$$

- α_j is the course-section-level random effect, which is modelled as

$$\alpha_j \overset{iid}{\sim} Normal(0, \sigma_\alpha^2)$$

- ϵ_{ijk} is the student-level error term, which is modelled as

$$\epsilon_{ijk} \overset{iid}{\sim} Normal(0, \sigma_\epsilon^2)$$

The random-effects structure accounts for within-instructor correlation as well as within-course-section correlation. In particular, we allow for between-course-section variability, since different sections — held at different times of day and in different classrooms — may lead to differences in scores.

Other Models Considered (but not used)

Before arriving at the final model described above, we considered a number of additional models:

Model 1:

$$y_{rk} = \mu_r + \epsilon_{rk}$$

Where $\mu_r \overset{iid}{\sim} N(\mu, \sigma_\mu^2)$ is the random effect of an instructor-section pair, μ is the true subgroup mean score, and ϵ_{rk} is the student-level random effect. This model was not used because it underestimates the variability introduced by the assignment of instructor and course-section.

Model 2:

$$y_{rk} = \beta_0 + \beta_1 X_i + \mu_i + \omega_c + \epsilon_{ick}$$

Where ωc is the course-level random effect, instead of the course-section-level random effect. We chose the proposed model instead of model 2 because this model assumes that students in the same course taught by the same instructor would give scores in the same distribution. However, we have observed that students in different sections of the same course can rate the instructor differently, potentially because of the location and time of the section.

Model 3:

$$yicjk = \beta 0 + \beta 1Xi + \mu i + \omega c + \alpha cj + \epsilon icjk$$

Where αcj is the section-level random effect nested in course c This model is not chosen because there are not enough data of courses with multiple sections.

Model 4:

$$yijk = \beta 0 + \beta 1Xi1 + \beta 2Xi2 + \beta 3X1X2 + \mu i + \alpha j + \epsilon ijk$$

Where $X1$ and $X2$ are the two covariates of interest: sex and race. This model is more complicated than the proposed model as it includes the interaction effect of two covariates. However, this model was not used since model fitting encounters non-convergence issue, possibly caused by lack of data to estimate the interaction effect.

Implementation

The model is fitted using the lmer() function from the lme4 R package. Part of the codes used for model fitting are given below.

- The mean estimates for each subgroup of interest (e.g., all instructors in the Math faculty) is obtained using the formula
$$Score(1 \vee instructor_{id}) + (1 \vee section_{id})$$

- The sex and racial difference for each subgroup is obtained using the formula Score ~ Sex + (1|instructor_id) + (1|section_id) and Score ~ Race + (1|instructor_id) + (1|section_id)

Appendix C: List of Figures