An Analysis of Winter 2022
Student Course Perceptions Survey Responses

Prepared February 2023 by
Sonya Buffone, Director, Teaching Assessment Processes
Kathy Becker, Specialist, Teaching Assessment Processes
Martin Lysy, Director, Statistical Consulting and Collaborative Research Unit
Glen McGee, Assistant Professor, Statistics and Actuarial Science
Meixi Chen, Statistical Consultant, Statistical Consulting and Collaborative 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 campus-wide in Winter 2022. The SCP survey is the first step in implementing this 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 Collaborative Research Unit (SCCRU) in the Faculty of Mathematics to analyze SCP data from Winter 2022.

Sources of Data

Data for this analysis was drawn from three primary sources: Winter 2022 SCP survey data, responses to Waterloo's Equity Survey from Winter 2022 instructors included in the study, and Human Resources records on 'sex at birth' for Winter 2022 instructors whose students completed SCP surveys.[1] In total, 1056 of the 2481 Winter 2022 instructors in this study did not complete the Equity Survey. This translates to 42% 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 among campus stakeholders, are correlated with SCP ratings assigned by Waterloo students. 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.

Of particular interest is whether there are differences in the SCP ratings by instructors' sex or racial identity sex after controlling for other instructor- or course-level variables.

Primary Analyses

The report consists of three primary analyses:

1. A descriptive analysis of the demographics and course characteristics at UW.

---

[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. Preliminary findings indicate a small but statistically significant difference in the ratings for white instructors and racialized instructors for two of six response items: on the five-point scale, white instructors received an average of 0.12 (4.24,4.13) points higher for Concepts_Conveyed and 0.12 (4.08, 3.96) points higher for Stimulated_Interest. Instructor sex data was complete, and analysis did not find statistically significant evidence of difference in ratings for male and female instructors accounting for class type, course size and instructor appointment type.

Descriptive Analysis of Demographics and Course Characteristics

In total, 40,316 SCP surveys from 2462 section-instructor pairs were submitted by students from across the six Faculties and Renison. It is important to emphasize, given that one section of a particular course might have more than one instructor, the 2462 section-instructor pairs include 1375 unique instructors, 1548 unique courses and 2244 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 2022 instructors identified as Indigenous in the Equity Survey (see Figure 1).

Figure 1: Number of Winter 2022 instructors by self-reported Indigenous identity – Note: Each Equity Survey respondent was asked to select all applicable racial identities. NA represents Winter 2022 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 indicated by Winter 2022 instructors in the Equity Survey. 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.
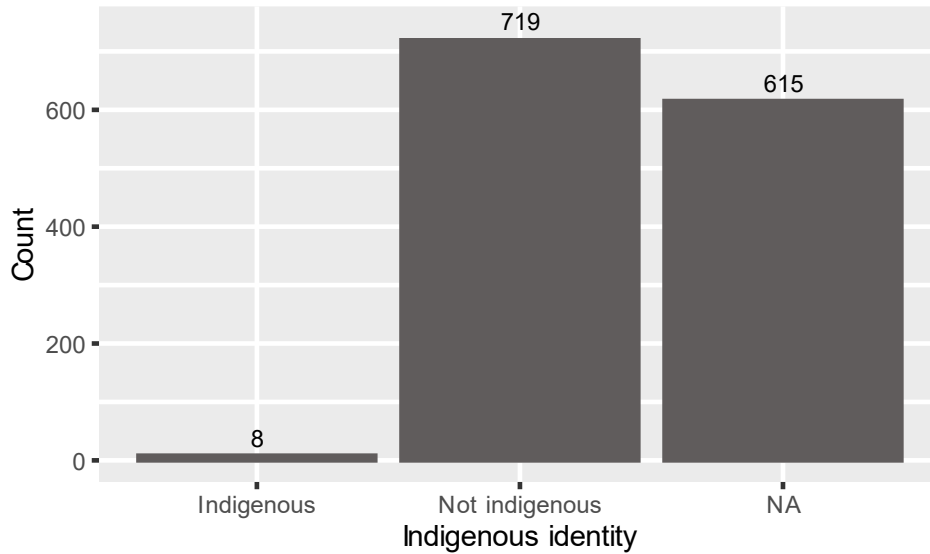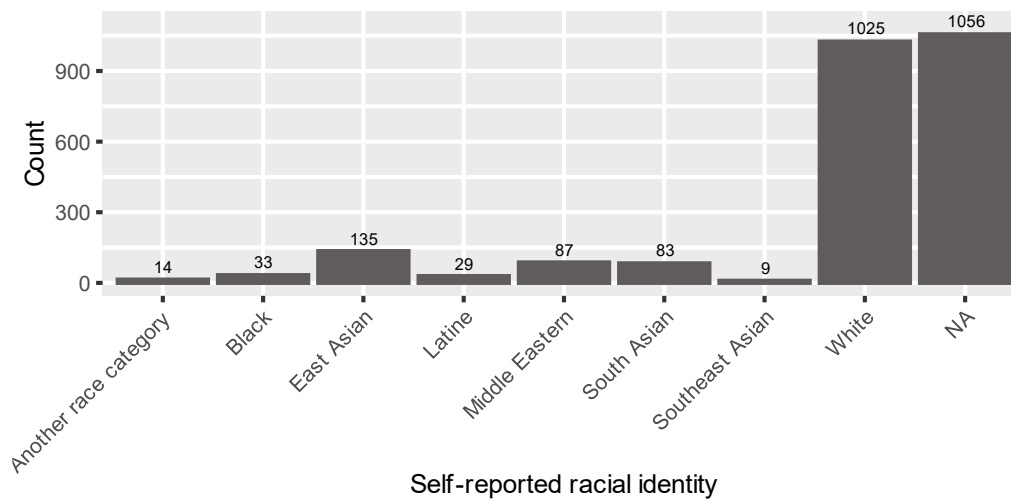

Figure 2: Number of Winter 2022 instructors by self-reported racial identity - Note: Each Equity Survey respondent was asked to select all applicable racial identities. NA represents Winter 2022 instructors who did not complete the Equity Survey or preferred not to answer the racial identity question.

https://uwaterloo.ca/teaching-assessment-processes

Racialized Instructors by Faculty
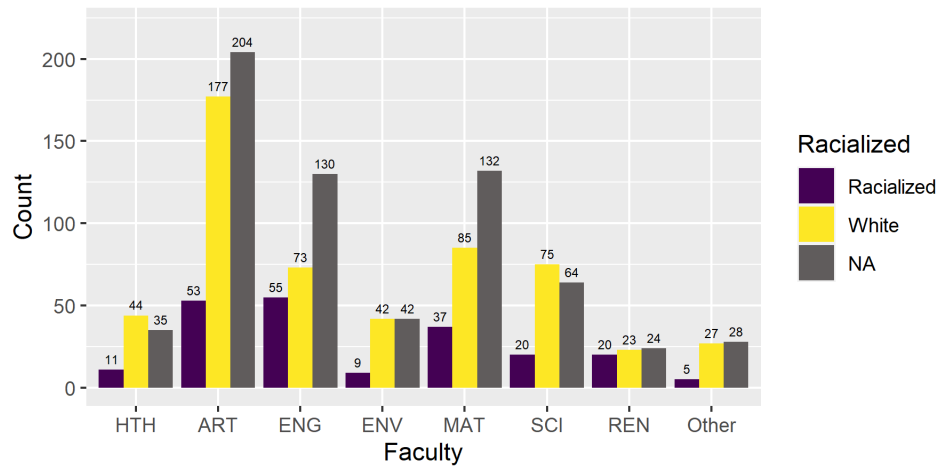
Figure 3: Number of Winter 2022 instructors by faculty and racial identity – Note: NA represents Winter 2022 instructors who did not complete the Equity Survey or preferred not to answer the racial identity question. "Other" includes University Colleges (except for Renison) and those that do not belong to any faculty (e.g., GRAD).

We examined the number of racialized instructors across each Faculty with SCP surveys in Winter 2022. As shown in Figure 3, setting missing cases aside, there were substantially more white instructors than racialized instructors across all Faculties and Renison (although Renison had only 3 fewer racialized than white instructors). Across four of the six Faculties (Health, Arts, Environment, & Science) almost four-times more white instructors taught courses in the Winter 2022 term. In Engineering, 73 white instructors and 55 racialized instructors were included in the dataset.

Instructor Time in Canada
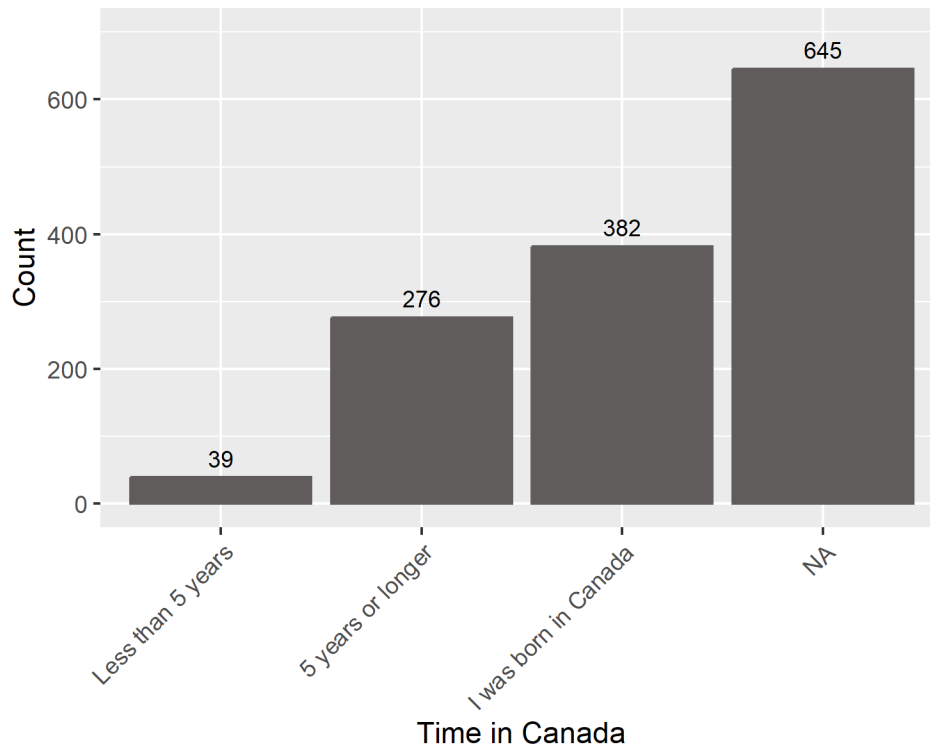
Figure 4: Number of Winter 2022 instructors by self-reported time in Canada. Note: NA represents Winter 2022 instructors who did not complete the Equity Survey or preferred not to answer the time in Canada question.

Figure 4 outlines the length of time in Canada for Winter 2022 instructors who responded to the Equity Survey. Overall, 382 instructors in the sample were born in Canada, 276 had been in Canada 5+ years and only 39 instructors reported being in Canada for <5 years. Again, it is important to be mindful of the proportion of missing cases (645) for whom no information on time spent in Canada is available.

Instructor Racial Identity by Time in Canada

Figure 5 provides a further breakdown of the equity sample characteristics. As indicated by the yellow bars we can see that of the respondents who are white, two-thirds (69%) were born in Canada, while for the racialized respondents in this sample, less than one-quarter (16%) were born in Canada. On the other hand, nearly three-quarters of the racialized instructors (74%) report being in Canada 5+ years compared to only one quarter (27%) of the white respondents who have been in Canada 5+ years.
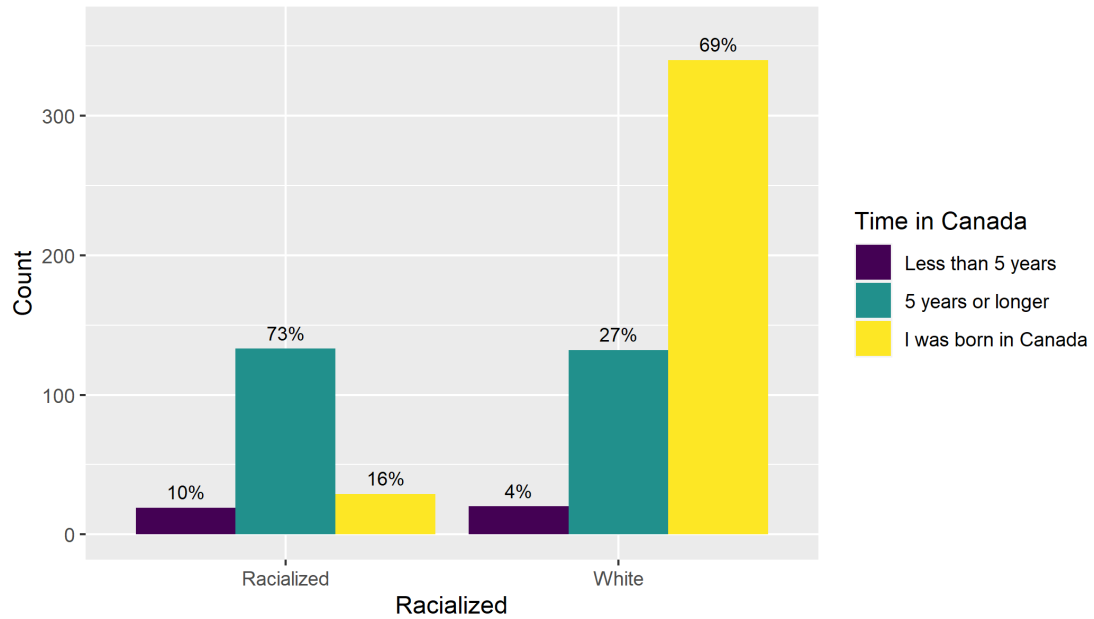
Figure 5: Number of white and racialized Winter 2022 instructors by time spent in Canada.

Instructor-level Variable: Sex

Instructor Sex

As Figure 6 illustrates, the total sample of instructors included in this analysis is 1342. This included 505 female instructors and 837 male instructors (based on Human Resources' data for sex at birth and does not necessarily reflect instructor gender identity at the present time).
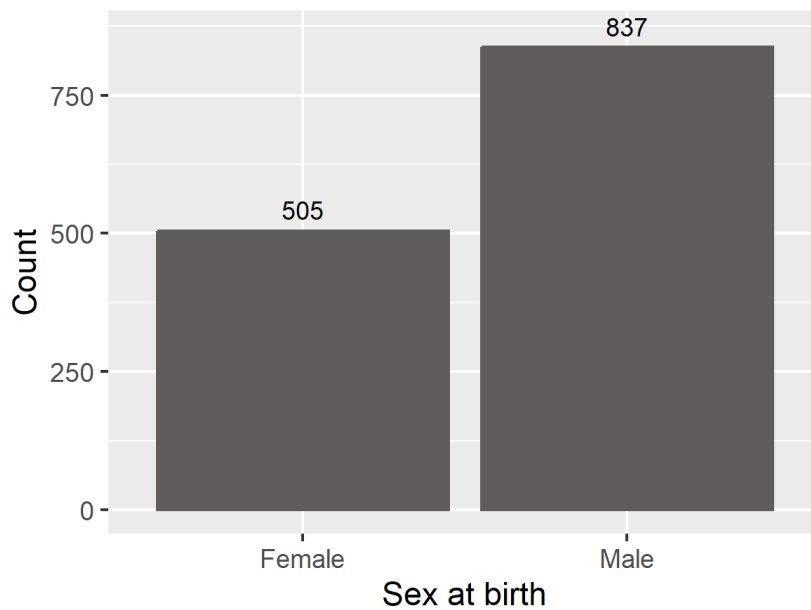
Figure 6: Number of Winter 2022 instructors by instructor sex.

Instructor Sex by Faculty

Figure 7 shows the sex breakdown of Winter 2022 instructors across each faculty. Male instructors outnumber female instructors across the STEM Faculties: Engineering, Math and Science. In Engineering, there are nearly four times more male instructors, (203 males, 55 females). In Math, there are over three-times more male instructors, (194 males, 60 females). And in Science we see over three times as many male instructors as female (194 males and 60 females).
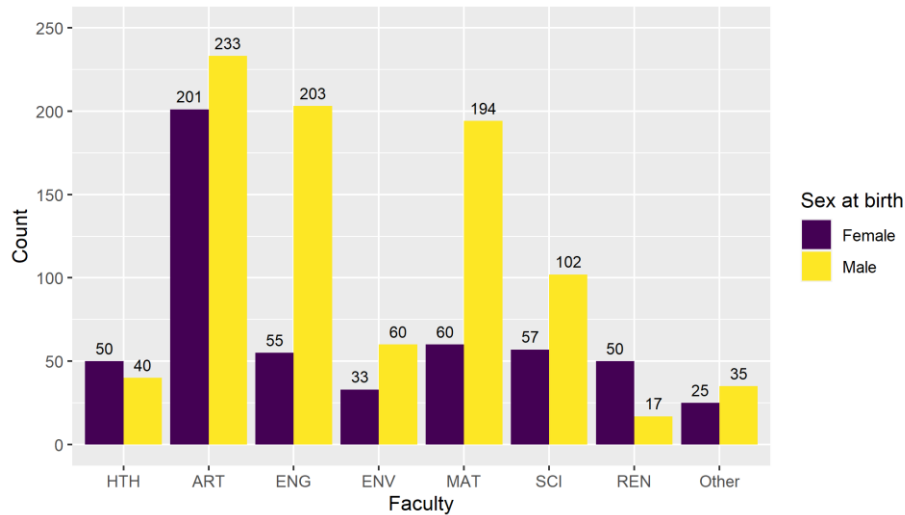
Figure 7: Number of Winter 2022 instructors by Faculty and Sex.

Instructor Sex by Appointment Type
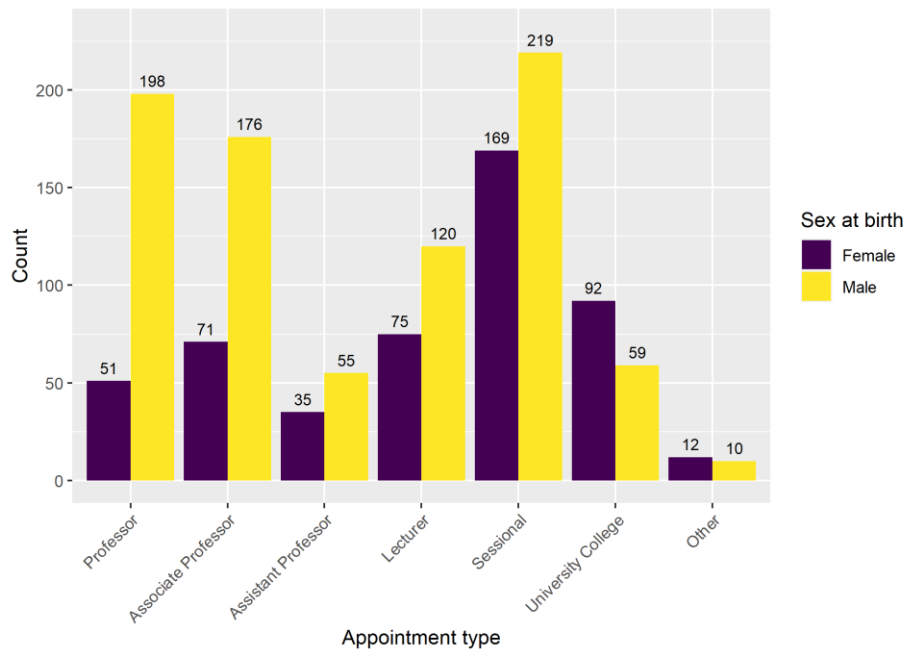
Figure 8: Number of Winter 2022 instructors by appointment type and sex.

Figure 8 further breaks down our sample characteristics outlining the instructors' appointment type by sex for Winter 2022. In general, male instructors outnumber female instructors across all ranks. For example, when it comes to the Professor or Associate Professor ranks, there are nearly four times more males who are Professors (51 females and 198 males), and almost 2.5 times more males are Associate Professors (71 females and 176 males). Thirty-five females in our sample are Assistant Professors compared to 55 males who hold this rank. When it comes to the appointment type of Lecturer, males continue to outnumber females (120 males and 75 females) and for Sessional appointments we see about 1.2 times more males (219) compared to females (169).

Course-level Variable: Class Size

Figure 9 outlines the number of courses by class size. The bulk of courses in Winter 2022 had 100 or fewer students with 793 courses that had 1-25 students, 563 courses with 26-50 students and 500 with 51-100 students in their courses. There were only 59 courses offered with 201+ students.
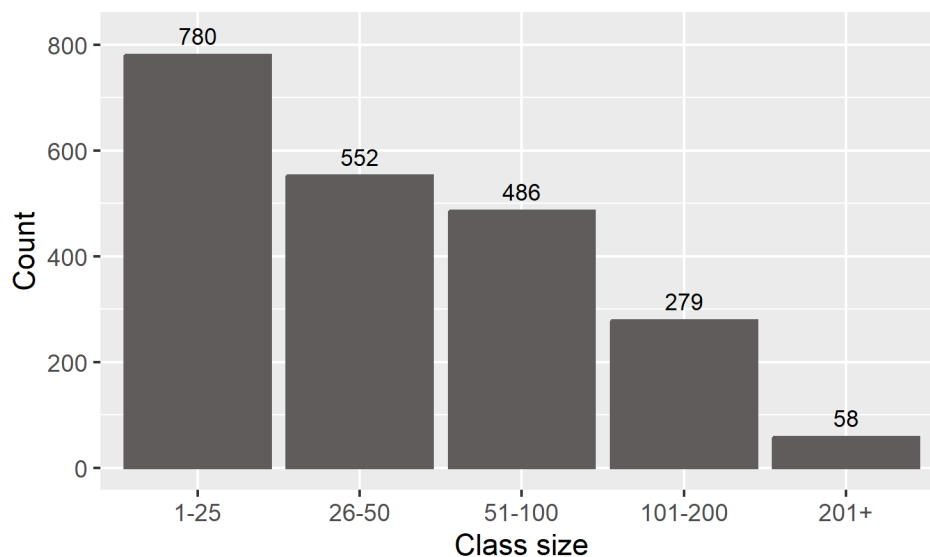
Figure 9: Number of Winter 2022 courses by class size.

Course-level Variable: Course Type by Faculty
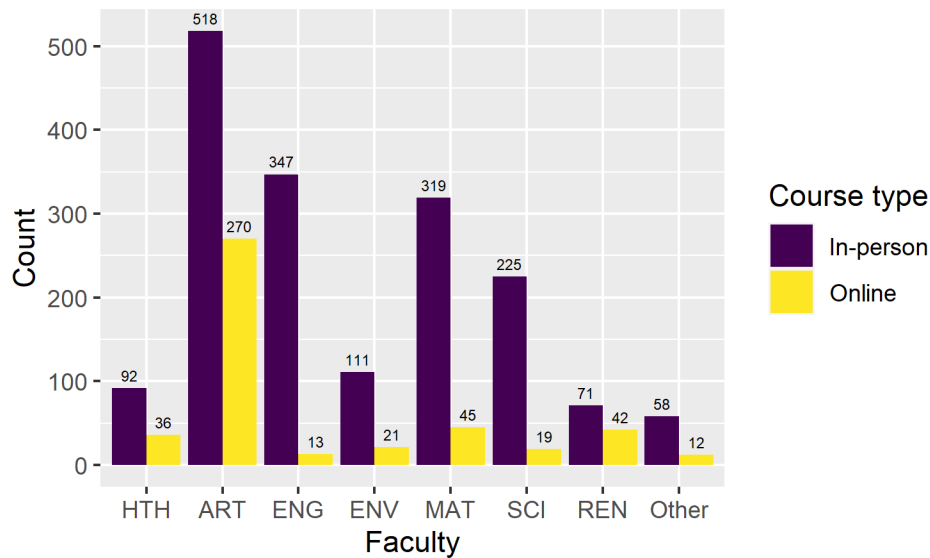
Figure 10: Number of Winter 2022 course sections by faculty and course type.

The majority of Winter 2022 courses surveyed were held in person, as reflected in Figure 10. The Faculty of Arts had the highest number of online course offerings (270) but still held almost two-times that number of courses in person (518), while Engineering held the second highest number of in-person courses (347) with only 13 courses offered online.

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 in only 1% of the cases. As is often the case with course evaluations, 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 82% of the time across all six items. Conversely, respondents selected 'Disagree' and 'Strongly Disagree' between 6% and 12% of the time across all six items. With respect to selecting the mid-point ('Neutral') on the five-point scale, we can see that student respondents selected this option between 10% and 17% of the time across all six items.
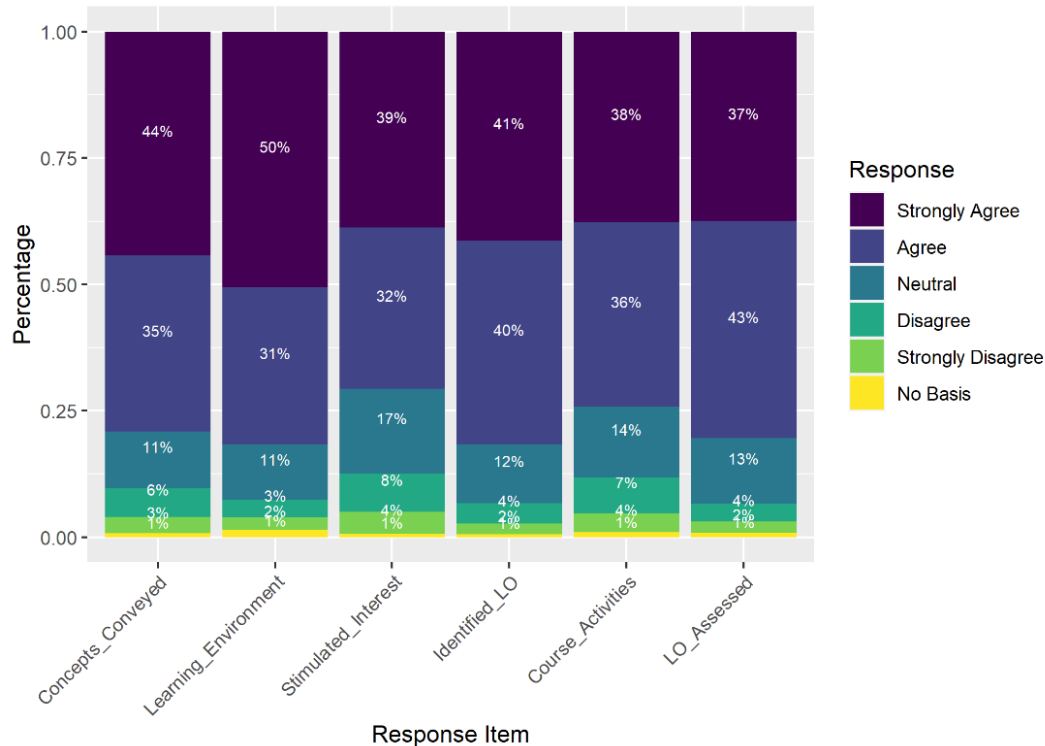
Figure 11: Proportion of Winter 2022 responses per response item.

A Note on No Basis for Rating

As informed by the 2018 pilot test, the SCP survey includes the response choice 'No Basis for Rating.' This option is useful to include for several reasons. Without this option, respondents will often select the '3' on a five-point Likert scale as a 'catchall' response for perceptions such as 'don't know', 'not applicable', 'no opinion' etc. Or, in cases where student respondents do not feel they can assess the question, they may elect to leave the question blank contributing to missing data. Perhaps most importantly, the 'No Basis for Rating' category provides a proxy measure for whether the survey items are in fact measuring things that resonate with students' learning experiences. For example, if this category was selected with high frequency, it could indicate that respondents were having difficulty understanding the survey items and/or that the items do not resonate with students' learning experiences.

SCP Mean Ratings

The following sections contain analyses of instructors' mean rating (MR) of SCP scores for each response item and for various combinations of the 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's) and 95% confidence intervals (CI's) 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. In this case, this trend remains, with most response items receiving scores between 3.9 and 4.5).

As the literature and our 2018 pilot test report find, smaller courses (from 1-100 students) tend to receive higher scores across all six items. It also is interesting to observe very similar trends in the plotting of the data across all course sizes (the dotted line patterns). We see a peak for Learning_Environment across all course sizes (ranging from 3.9 in 200+ courses to close to 4.5 in courses with 25 or fewer students). Conversely, Stimulated_Interest (ranging from 3.8 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 course sizes.

Figure 12: Mean Winter 2022 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, if we look at the red line (Science) the mean rating for Learning_Environment was about 4.15 compared to a mean rating of 3.9 for Course_Activities which translates to a difference of 0.1 decimal-points on the 5-point scale.
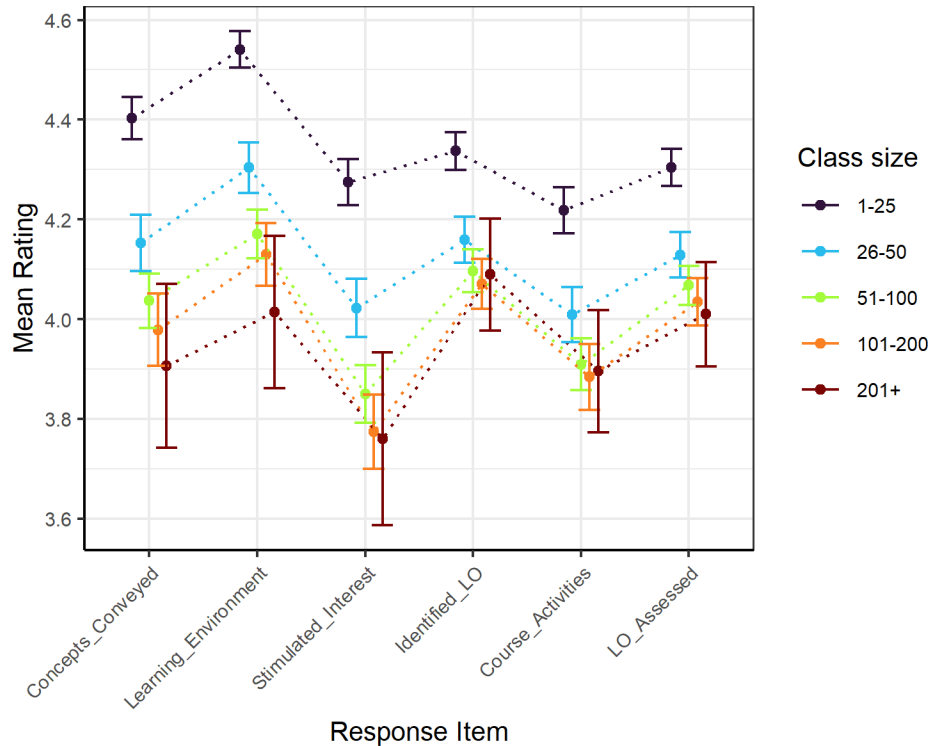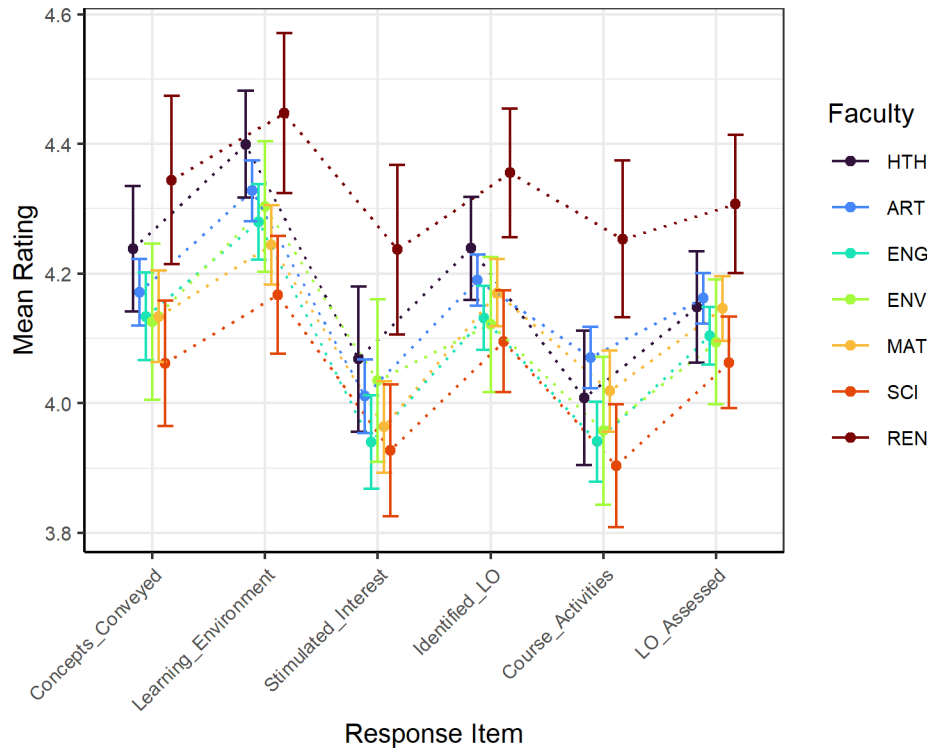
Figure 13: Mean Winter 2022 rating by response item and Faculty of course offering. Error bars correspond to 95% confidence intervals.

Mean Ratings by Response Item and Instructor Time in Canada

Figure 14 shows the mean ratings for response items taking Instructors time spent in Canada into account. The findings reveal that mean ratings are slightly higher (0.1 points) for instructors born in Canada (brown line) ranging from a mean rating of about 4.1 (Stimulated_Interest) to 4.4 (Learning_Environment) across the survey items, followed by those reporting they have been in Canada for 5+ years (green line) 4.0 (Stimulated_Interest) to 4.3 (Learning_Environment), and finally, with those reporting <5 years in Canada (black line) receiving the lowest mean ratings across all six items (3.95-4.25). The difference in scores is relatively small ranging from about 0.1-0.2 points with large confidence intervals for <5 years in Canada (black line) this reflects the small sample size for this variable (n=39). 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 for time spent in Canada, while Stimulated_Interest and Course_Activities scored the lowest.
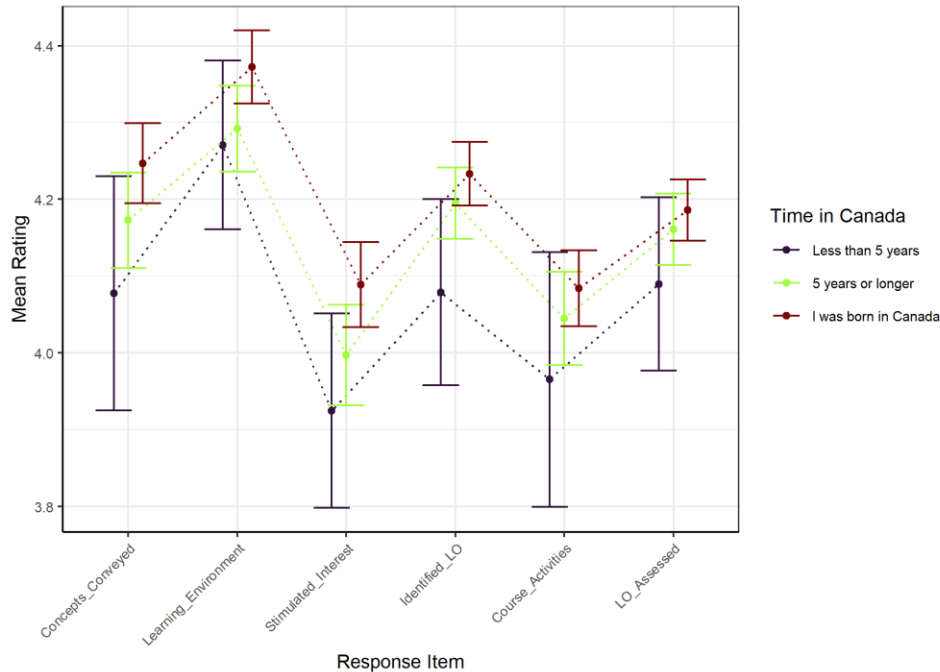
Figure 14: Mean Winter 2022 rating by response item and instructor's self-reported time in Canada. Error bars correspond to 95% confidence intervals.

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, instructor rank, student enrolments, and courses type (e.g., online or in person). **It is important to note that the following analyses are exploratory and cannot be used to make causal claims about the effect of racial or sex on SCP ratings.** Nonetheless, they are 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 confidence intervals to the zero line:
    - When the confidence interval <u>doesn't cross</u> the zero line, it indicates a <u>statistically significant difference</u> in MCA using a 95% confidence interval (p≤0.05)
    - When the confidence interval <u>crosses</u> the zero line, it indicates a difference in MCA that is <u>not statistically significant</u> 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)
    - When the plot point is <u>above</u> the zero line, the difference benefits those who experience systemic privilege.
    - When the plot point is <u>below</u> 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 teaching a course section size of 201+ Environment Faculty) no data will be displayed in the accompanying figure.
- To protect instructor anonymity, in cases where a certain interaction exists but results in a small number of instructor-section pairs (n=<5: e.g., associate professor teaching a course section of 1-25 students in the Environment faculty), 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 course 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 2022 instructors who did not complete the Equity Survey (42%). As a result, it is difficult to attribute significant weight to the 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 has never been explored in this context at UW. An additional word of caution is worth repeating: **the analyses conducted are exploratory only and do not, under any circumstances claim to make causal inferences.**

Difference in MR for White and Racialized Instructors

Figure 15 displays the difference in mean ratings for white and racialized instructors. The difference in MR for white and racialized instructors for the following two items: Concepts_Conveyed and Stimulated_Interest is statistically significant (error bars correspond to 95% confidence intervals). Table 2 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 0.06-0.12 points on the five-point Likert Scale. For Concepts_Conveyed, white instructors received a mean rating of 4.24 compared to a mean rating of 4.13 for racialized instructors (difference = 0.12 points). Similarly, for Stimulated_Interest, white instructors received a mean rating of 4.08 compared to 3.96 for racialized instructors (difference = 0.12). Importantly, what we observe is that for all six response items, rounding to one decimal place results in a difference of 0.1, meaning that white instructors receive mean ratings that are 0.1 higher across all six survey items compared to the mean ratings of racialized instructors. **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.**
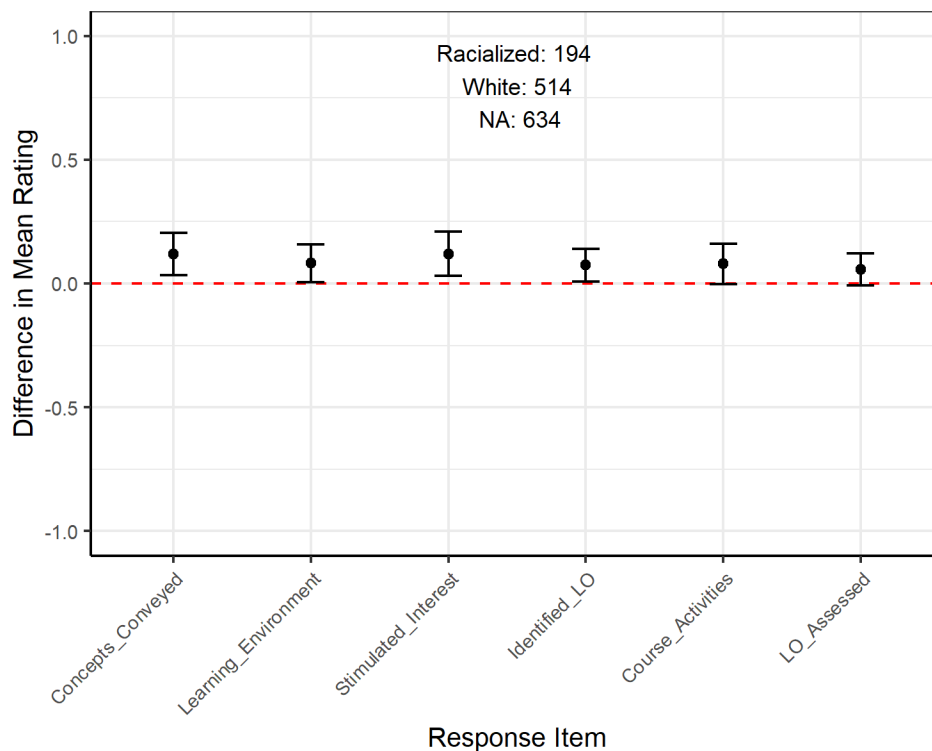

Figure 15: Difference in mean Winter 2022 rating for white and racialized instructors by response item.

Table 1: Difference in mean Winter 2022 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	4.13	0.12	(4.19, 4.28)	(4.05, 4.2)	(0.03, 0.2)
Learning_Environment	4.35	4.28	0.08	(4.31, 4.4)	(4.21, 4.34)	(0.01, 0.16)
Stimulated_Interest	4.08	3.96	0.12	(4.03, 4.12)	(3.88, 4.04)	(0.03, 0.21)
Identified_LO	4.23	4.16	0.07	(4.19, 4.26)	(4.1, 4.21)	(0.01, 0.14)
Course_Activities	4.08	4.01	0.08	(4.04, 4.13)	(3.93, 4.08)	(0, 0.16)
LO_Assessed	4.19	4.13	0.06	(4.15, 4.22)	(4.08, 4.19)	(-0.01, 0.12)

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

In Figure 16 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. However, despite the finding of non-significance, across all six Faculties, white instructors receive slightly higher mean ratings across all six items (the average difference is at most 0.25 points for any item and generally less than 0.1 across all Faculties and all response items). The only exception we see here is in the Faculty of Engineering where students assign higher mean ratings (<0.1 points) to racialized instructors as compared to white instructors for the following three response items: Concepts_Conveyed, Learning_Environment and Stimulated_Interest.
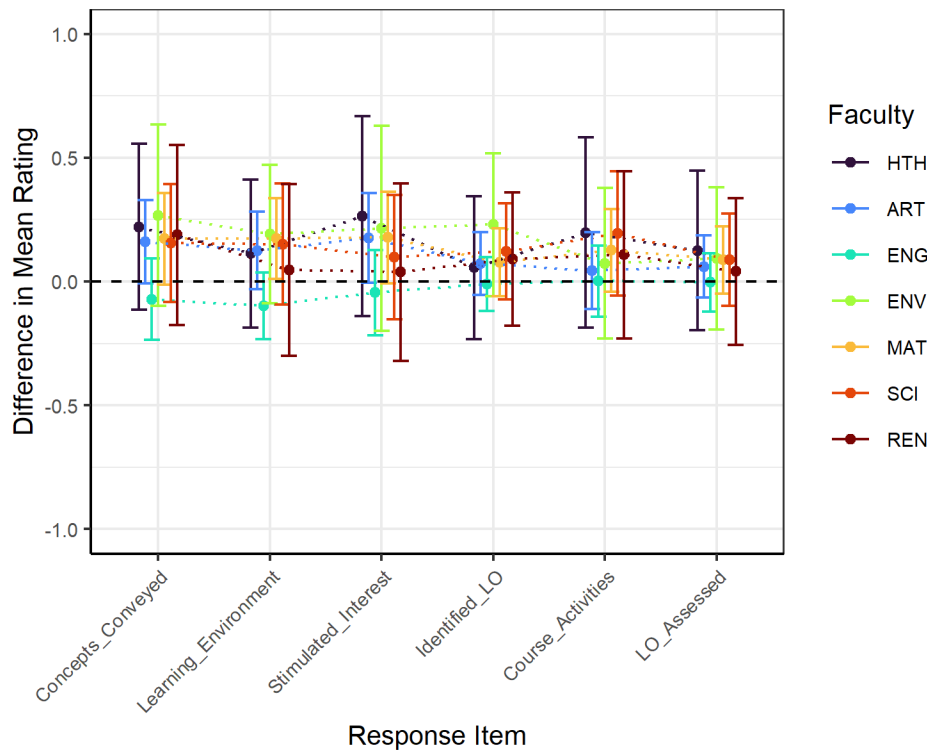

Figure 16: Difference in mean Winter 2022 ratings for white and racialized instructors by response item and Faculty of course offering. Error bars correspond to 95% confidence intervals.

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

Figure 17 displays the difference in mean ratings given by student respondents for white and racialized instructors while also taking instructor appointment and course enrolment into consideration.

In general, accounting for instructor appointment type and course enrollment, the difference between ratings given by students for white instructors and racialized instructors, is not statistically significant. In most cases the difference in mean ratings is close to 0, (as evidenced by the data points falling close to or on the 0-point).
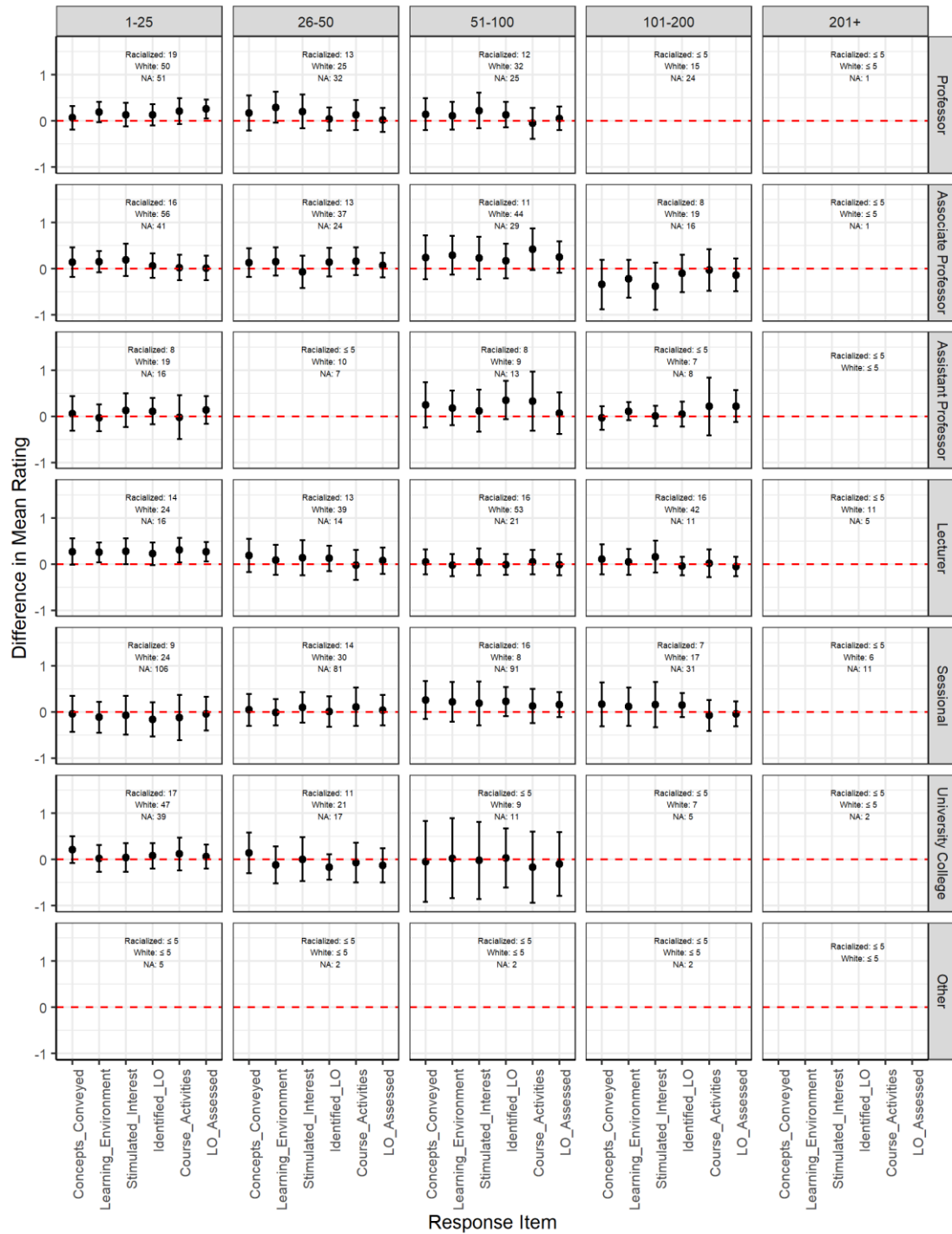
Figure 17: Difference in mean Winter 2022 ratings for white and racialized instructors by response item, instructor appointment type, and class size. Error bars correspond to 95% confidence intervals.

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

Figure 18 shows the difference in mean ratings for racialized and white instructors accounting for course type (online/in-class) and instructor appointment type. Again, we observe the

differences in scores are not statistically significant when accounting for instructor appointment or course type.
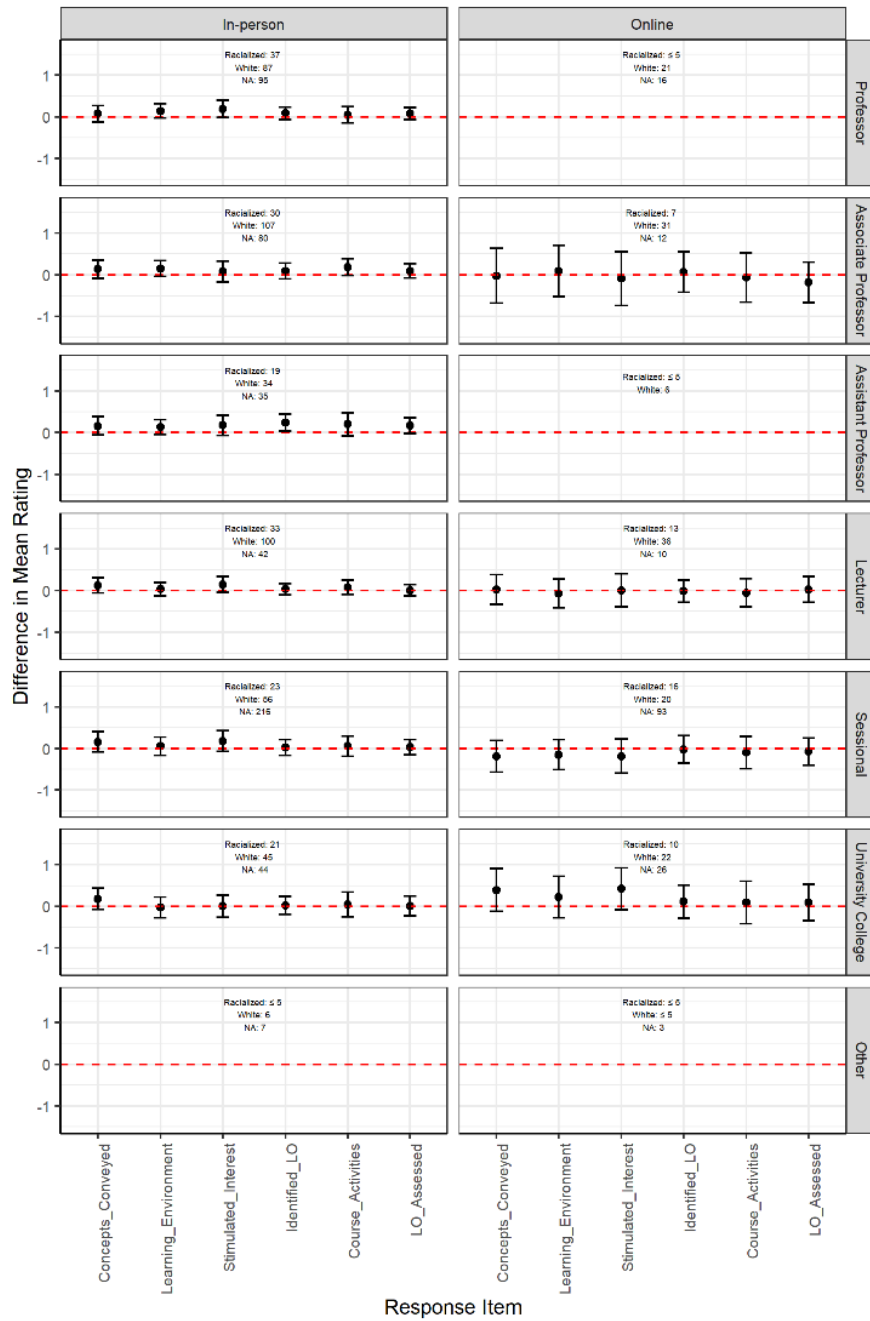

Figure 18: Difference in mean Winter 2022 ratings for white and racialized instructors by response item, instructor appointment type, and course type. Error bars correspond to 95% confidence intervals.

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

Figure 19 shows the difference in mean ratings for racialized and white instructors accounting for course type (online/in-class) and course size. Again, we observe that the differences in scores are not statistically significant accounting for course type or class size.
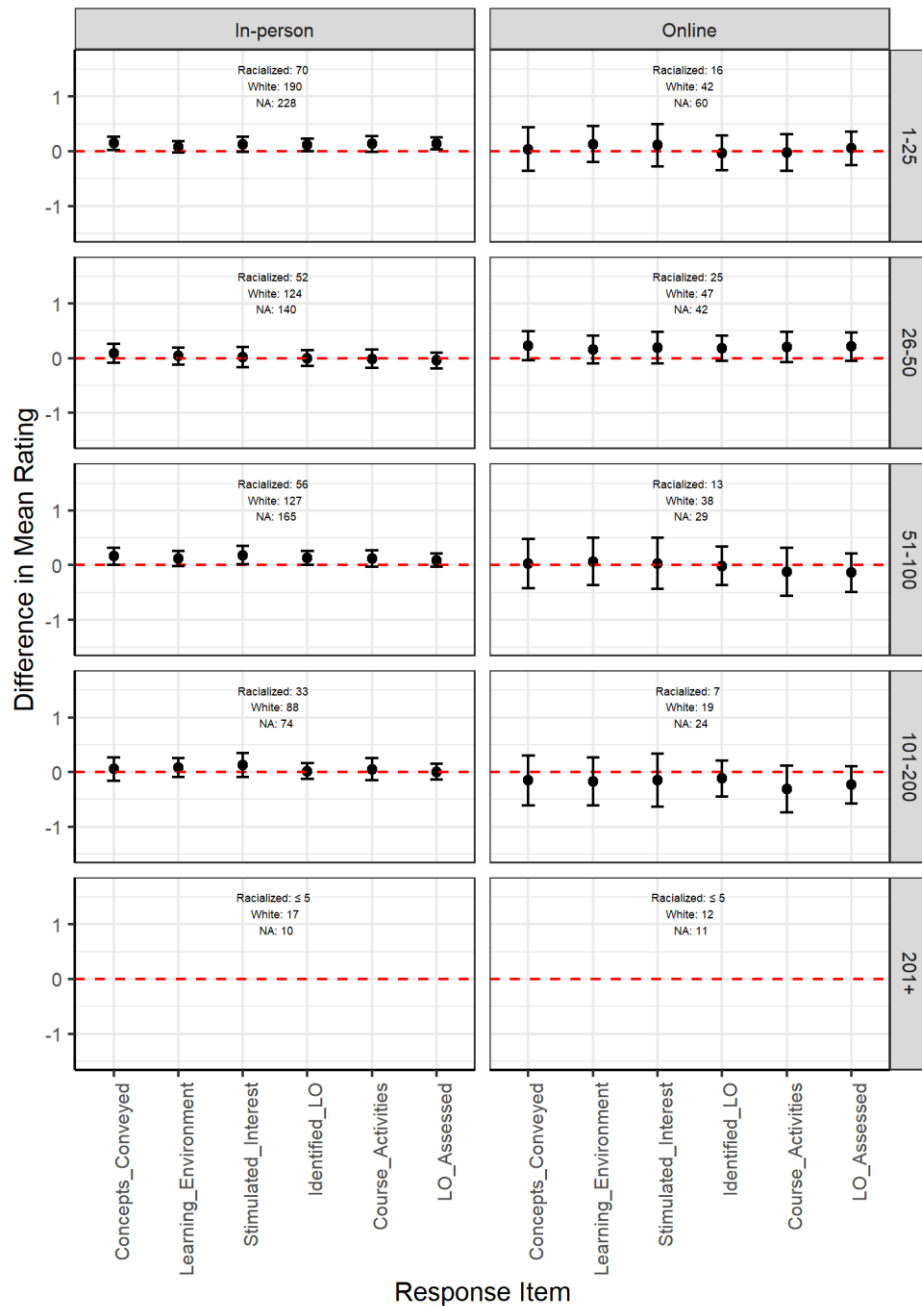

Figure 19: Difference in mean Winter 2022 ratings for white and racialized instructors by response item, class size, and course type. Error bars correspond to 95% confidence intervals.

Difference in MR for White and Racialized Instructors by Class Size

Figure 20 displays the difference in mean ratings assigned to white and racialized instructors by course size. This plot reveals that when accounting for course size, the difference in mean rating for white and racialized instructors is not statistically significant – with one specific exception. Among courses with 1-25 students (black line), white instructors receive a mean rating that is 0.15 points higher than racialized instructors on two of six response items (Concepts_Conveyed and LO_Assessed), and the difference is statistically significant. The difference in mean ratings for courses with enrollments of 26-50 (blue line) and 51-100 (green line) are slightly higher for white instructors (ranging from 0.15 or less), but are not statistically significant.
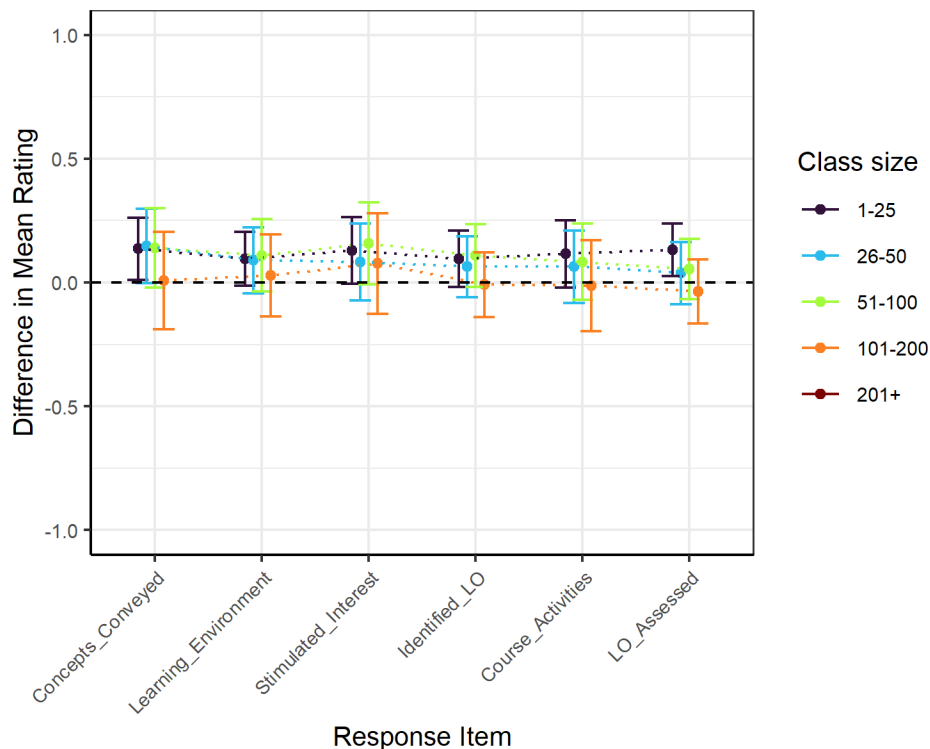

Figure 20: Difference in mean Winter 2022 ratings for white and racialized instructors by response item and class size. Error bars correspond to 95% confidence intervals.

Overall, this plot indicates that across all class sizes the difference in scores for white and racialized instructors is not statistically significant. Practically speaking (in reference to the 5-point scale), the difference observed in scores is relatively small (<0.15 points) in most cases (and ranges from -0.2 to +0.3 in the extreme cases: top ends of the error bars in both directions), **this illuminates how important it is for academic leaders to make decisions that are not based on decimal-point differences in scores for instructors.** Moreover, these findings emphasize the importance of continued investigation of how race and SCP scores interact.

Figure 21 provides a visual of the difference in mean ratings for all six items accounting for instructor race and time in Canada. The results of this analyses show that the difference in mean ratings for white and racialized instructors is not statistically significant accounting for time spent in Canada. For those reporting <5 years in Canada (black line) we see the largest difference in scores (about an average of 0.2 points higher across all items) for white instructors with <5years in Canada, compared to racialized instructors. The confidence intervals in this case (black line) are quite large, reflective of the small sample size. The second largest difference in mean ratings is found for those reporting 5+ years (green line) in Canada. Again, we observe mean ratings for white instructors are an average of about 0.1 points higher compared to racialized instructors. Finally, the smallest observable difference in scores is found among instructors born in Canada (brown line) but we see the opposite pattern emerging, where racialized instructors score slightly higher (0.1 points or less).
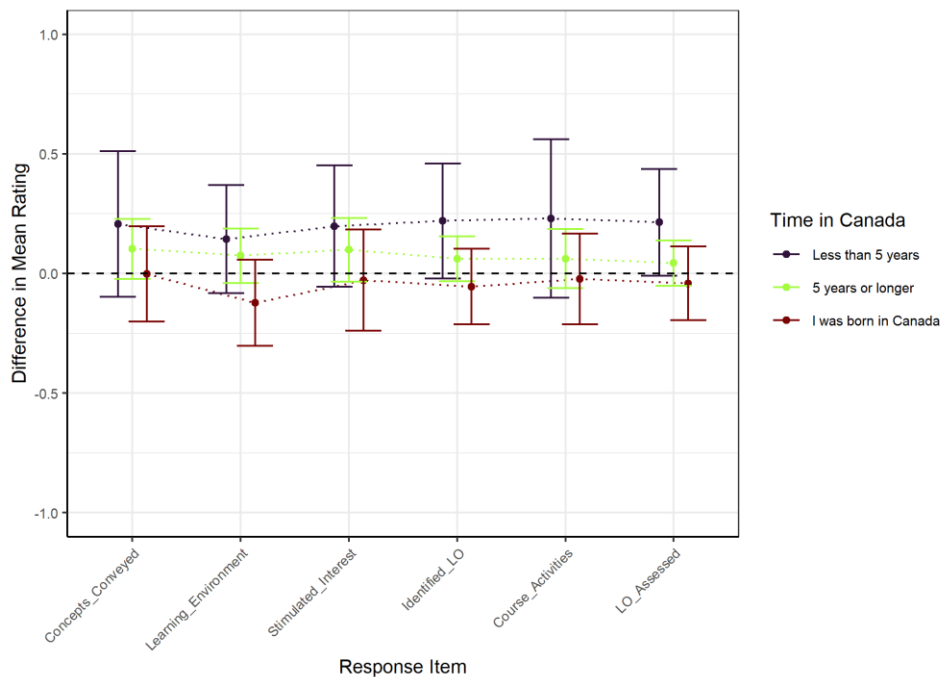

Figure 21: Difference in mean Winter 2022 ratings for white and racialized instructors by response item and instructor's time spent in Canada. Error bars correspond to 95% confidence intervals.

These findings are interesting given they might point to a possible link between instructor time spent in Canada and mean ratings. In the literature on SCPs 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, 2022). The consensus on whether this is a bias in and of itself, or if it is indeed an impediment to student learning is contentious. As such, we wanted to explore the possibility that respondent bias against instructors who were not born in Canada may result in lower mean ratings. We did not have data on instructor's first language, or a method to measure

'cultural familiarity' in the equity survey, so we relied on "Time spent in Canada" (<5 years, 5+years, or born in Canada), an imprecise proxy measure. We created a dichotomous variable to measure whether the instructor was born inside or outside of Canada, collapsing two categories for time spent in Canada: <5 years & 5+ years into one category.

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 22 shows the difference in mean ratings for instructors born in or outside of Canada by instructor racial identity. The large confidence intervals reflect the small sample size generated from this parsing of the data. With the exception of Learning _Environment, where we see a statistically significant difference in mean ratings for racialized instructors born in Canada as compared to racialized instructors born outside of Canada (a difference in score of <0.2 points on the 5-point scale), the remaining differences in ratings were not statistically significant. The findings show some evidence that mean ratings for racialized instructors born in Canada (black line) are slightly higher than
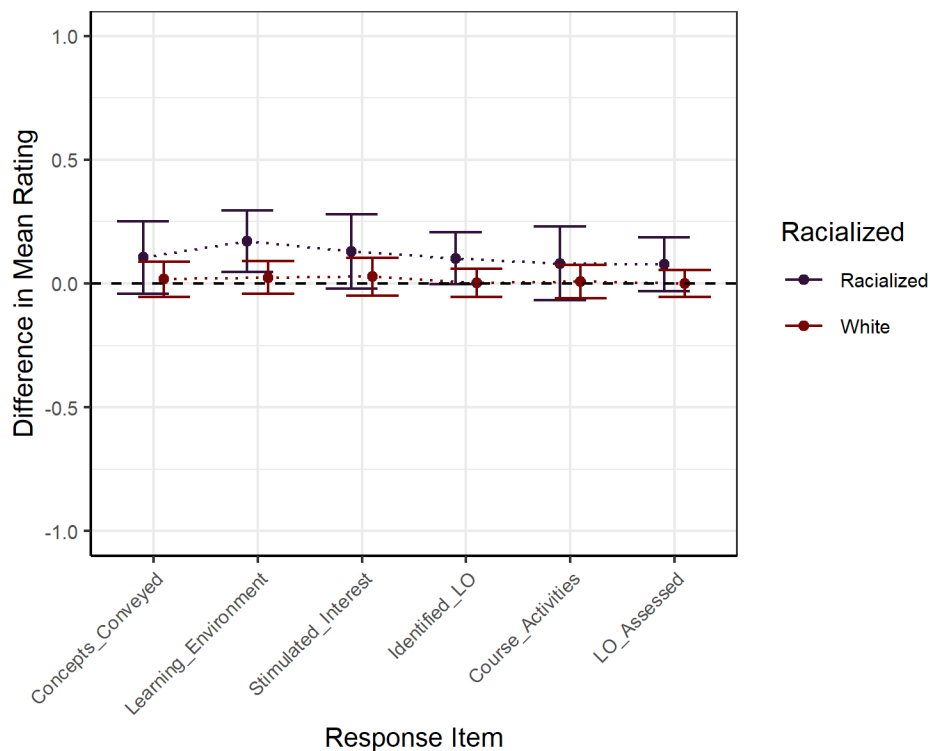

Figure 22: Difference in mean Winter 2022 ratings for instructors born in Canada and instructors born outside Canada by response item and instructor racial identity. Error bars correspond to 95% confidence intervals.

mean ratings for racialized instructors born outside of Canada (the difference is <0.2 points on average). Conversely, this does not hold true for white instructors born in Canada (brown line)

compared to white instructors born elsewhere as we see virtually no difference in these scores (the average difference across all six response items falls close to 0). In this sense, place of birth does not seem to be associated with mean ratings for white instructors to the same extent that place of birth seems to be associated with mean ratings for racialized instructors. These results provide some initial support for the hypothesis that respondent bias against instructors not born in Canada may result in lower mean ratings. Unfortunately, the proxy measure for a possible language barrier (or cultural familiarity etc.) give some pause, in attributing too much weight to these findings. To better understand these differences, we ran a series of additional statistical tests. The next section summarizes these findings and further examines the dichotomized 'born in Canada' measure.

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

In our analysis of the equity survey data at the institutional level we observed a significant overall difference in mean ratings for racialized and white instructors for two survey items: Concepts_conveyed and Simulated_interest. We also observed that the difference in mean ratings assigned to racialized and white instructors appeared most salient in the subpopulation of lecturers teaching sections with 1-25 students (see Figure 17).

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 23). Interestingly, Figure 23 very closely resembles Figure 15, which depicts the difference in mean ratings for white and racialized instructors. In both plots, we observe the difference in mean ratings for Concepts_Conveyed and Stimulated_Interest is statistically significant. This provides some support for the hypothesis that differences in scores might be explained by language and/or cultural barriers experienced by racialized and immigrant instructors and perceived by student respondents.
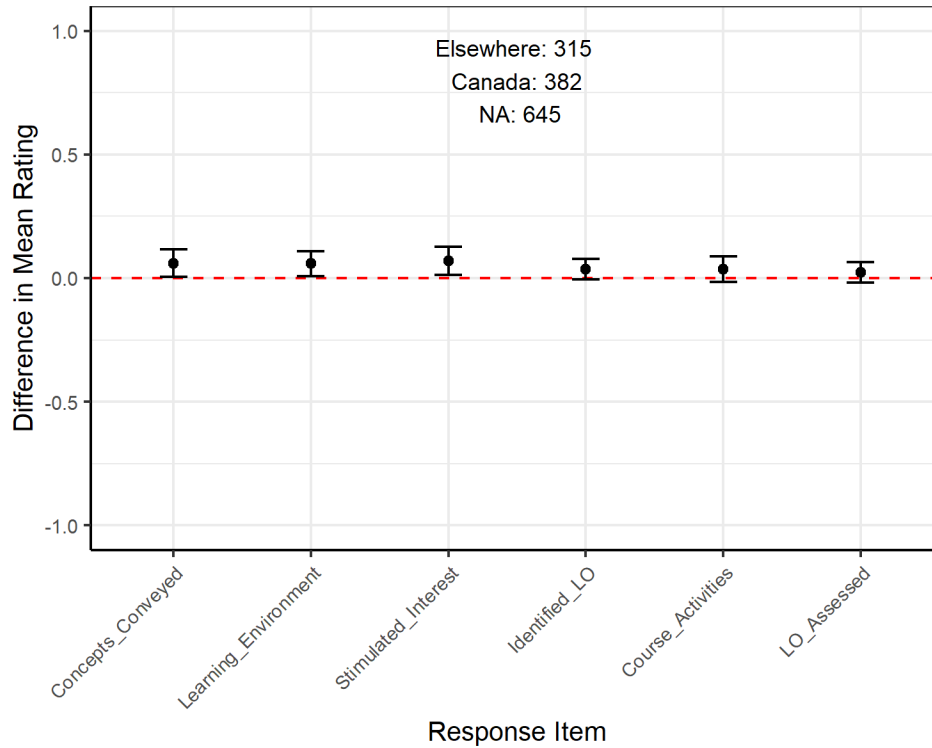
Figure 23: Difference in mean Winter 2022 ratings for instructors born in Canada and instructors born outside Canada by response item. Error bars correspond to 95% confidence intervals.

To further explore the relationships observed for racialized lecturers teaching courses with 1-25 students (see Figure 15) we broke the sample down further. The bar graph in Figure 24 shows that of the white lecturers teaching courses with 1-25 students 67% report being born in Canada. Comparatively, of the racialized lecturers teaching courses of the same size 86% report being born outside of Canada. Further research is necessary to examine the observed differences for instructors not born in Canada, particularly those teaching smaller courses.
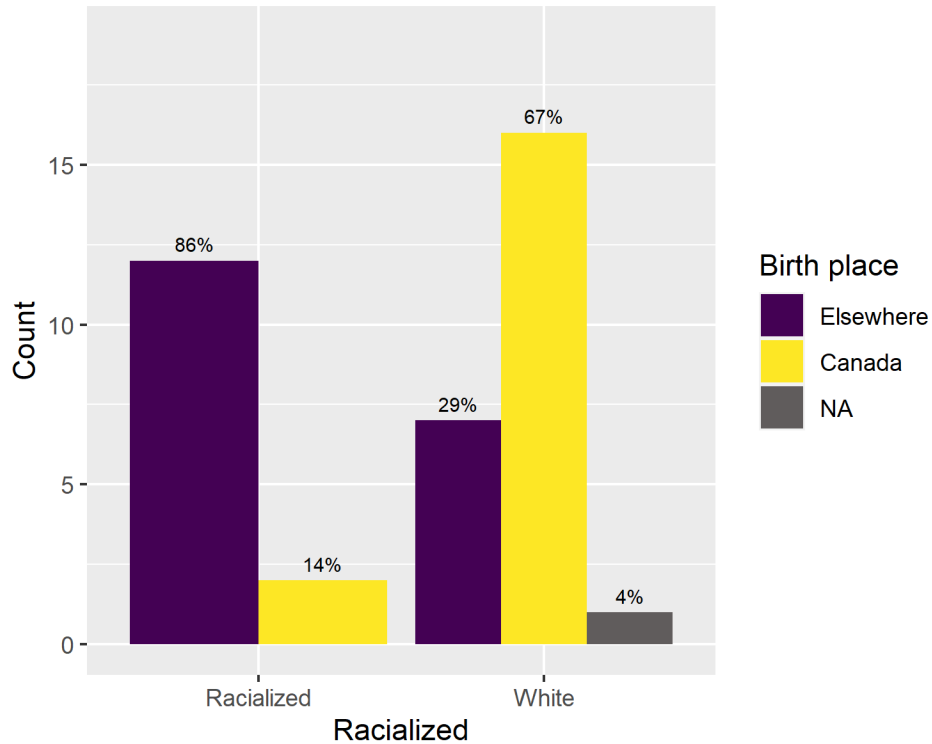
Figure 24: Percentage of white and racialized Winter 2022 lecturers teaching classes with 1-25 students by birthplace.

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 25 shows the difference in mean ratings assigned by student respondents to male and female instructors, across all response items. The confidence intervals cross the zero-line for all response items, signifying that the difference in ratings for male and female instructors is not statistically significant at the 95% confidence interval.
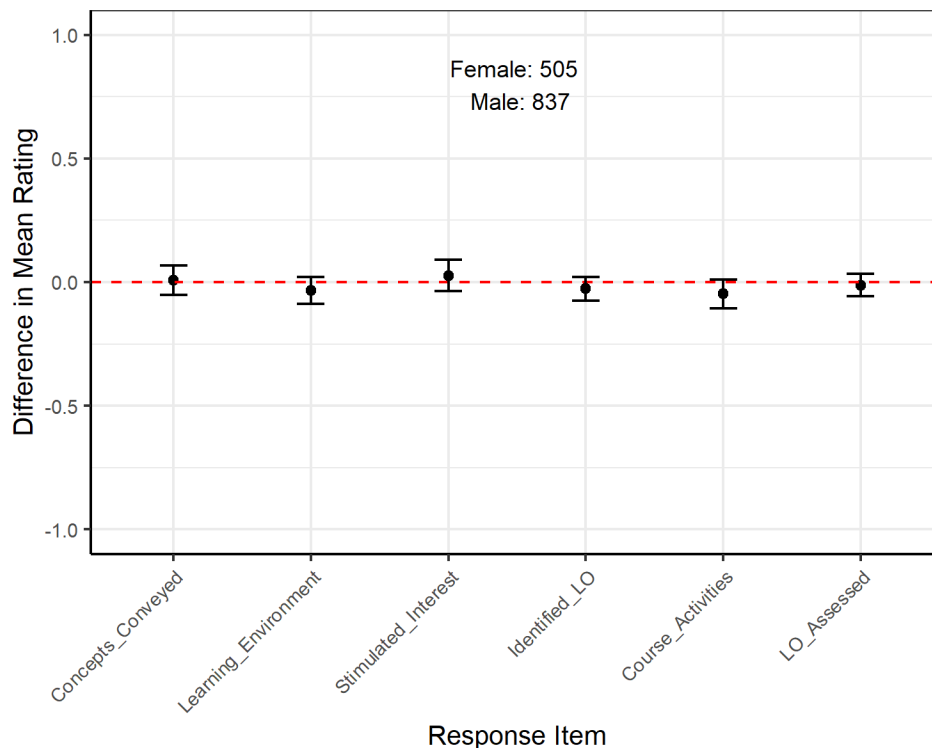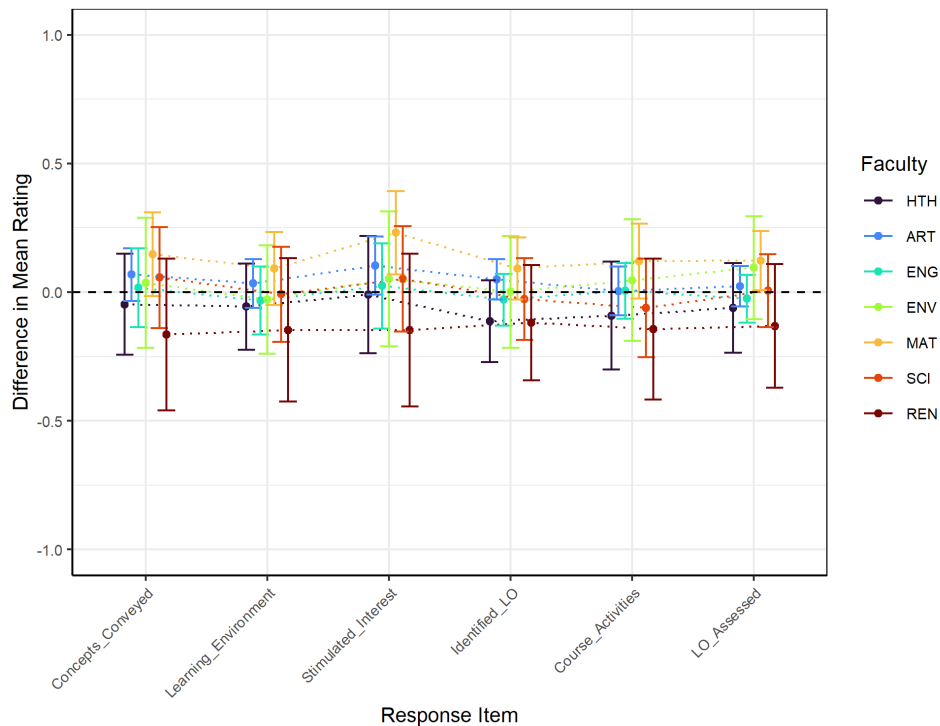
Figure 25: Difference in mean Winter 2022 ratings for male and female instructors per response item. Error bars correspond to 95% confidence intervals.

A clearer picture of the difference between mean ratings for male and female instructors is evident in Table 3, which also displays the confidence intervals for each response item.

Table 2: Difference in mean Winter 2022 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	4.16	4.16	0.01	(4.12, 4.2)	(4.11, 4.21)	(-0.05, 0.07)
Learning_Environment	4.28	4.32	-0.03	(4.25, 4.32)	(4.28, 4.37)	(-0.09, 0.02)
Stimulated_Interest	4.01	3.99	0.03	(3.97, 4.05)	(3.94, 4.04)	(-0.04, 0.09)
Identified_LO	4.16	4.19	-0.03	(4.13, 4.19)	(4.15, 4.23)	(-0.07, 0.02)
Course_Activities	4.00	4.05	-0.05	(3.96, 4.03)	(4, 4.09)	(-0.11, 0.01)
LO_Assessed	4.13	4.15	-0.01	(4.11, 4.16)	(4.11, 4.18)	(-0.06, 0.03)

We can see the largest difference in mean ratings for Course_Activities, with student respondents rating female instructors an average of 4.05 points and male instructors and average of 4.00 points. For the Learning_Environment item, student respondents rated female instructors an average of 4.32 and male instructors and average of 4.28 points. While for Stimulated_Interest, student respondents rated male instructors 0.02 points higher on average than female instructors. Again, across the six core items we can see how the difference in mean ratings on the five-point scale is very small and not statistically significant, ranging from -0.01 to -0.05.

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

We also explored mean ratings for male and female instructors accounting for the Faculty of course offering. Figure 26 reveals the difference in scores for female and male instructors, accounting for Faculty, are not statistically significant. In general, we see a close clustering of scores on the 0 line except for Stimulated_Interest in Math, where mean ratings for male instructors were 0.25 points higher.

Figure 26: Difference in mean Winter 2022 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 sought to examine mean ratings at the Faculty level for male and female instructors accounting for class size. As evidenced by the clustering of mean ratings around the zero-line in Figure 27, the difference in mean ratings for male and female instructors are not statistically significant when we account for class size and Faculty. Although the orange plotted lines for Health and Environment appear to show more extreme differences in scores, the large confidence intervals indicate the small sample sizes for classes with 100+ students in these two Faculties.
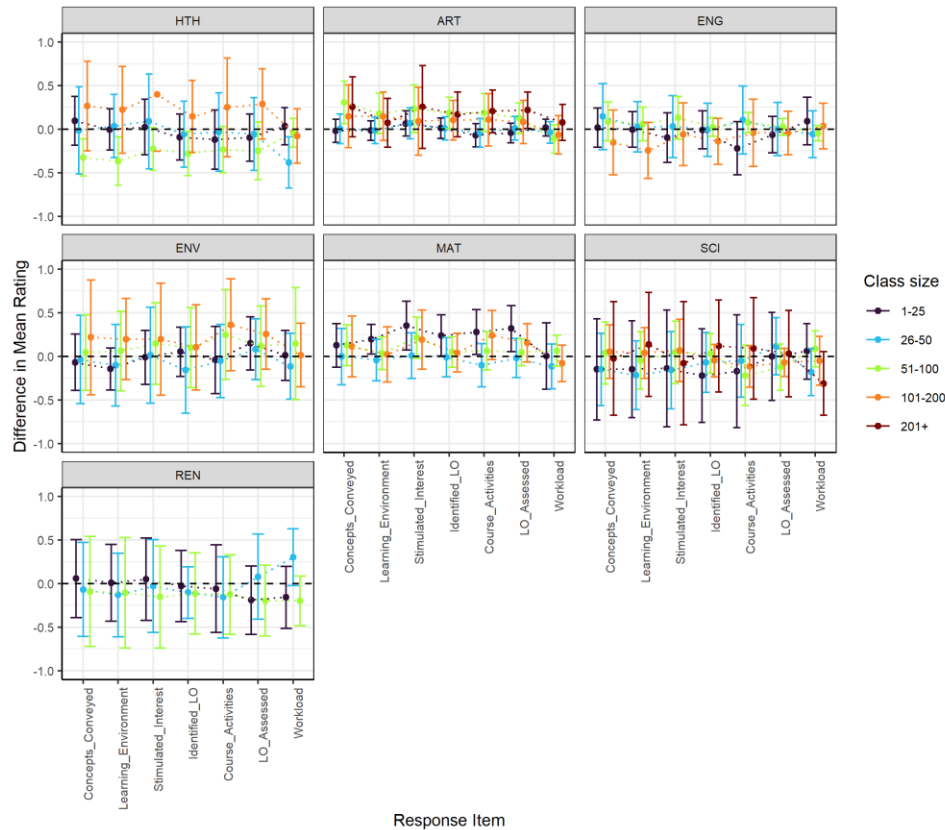
Figure 27: Difference in mean Winter 2022 ratings for male and female instructors by response item, class size, and Faculty of course offering. Error bars correspond to 95% confidence intervals.

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

Figure 28 depicts the difference in mean ratings for male and female instructors at the Faculty level accounting for class type (online/in-person). It is necessary to interpret these results within context of the sample sizes. If we refer back to Figure 10, we can see that the majority of courses offered in Winter 2022 were in-person, across all Faculties. In general, again we see a large clustering of scores around the zero-line across all Faculties indicating the difference in mean ratings for male and female instructors is not statistically significant accounting for the course type and the Faculty of course offering. In Health, the brown line indicates online courses taught by female instructors receive higher mean ratings than online courses taught by male instructors; however, this sample is small (n=36) (n_male=10 and n_female=21). A similar pattern emerges in Science, and a similar issue presents itself insofar as for Winter 2022 only 19 courses (n_male=10 and n_female=10 – there are two instructors teaching the same section) were online in Science. Finally, in Environment we observe that online courses taught by male instructors seem to receive higher mean ratings than online courses taught by female instructors, but again the small sample size (n=21) (n_male=12 and n_female=7) undermines the reliability of the findings.
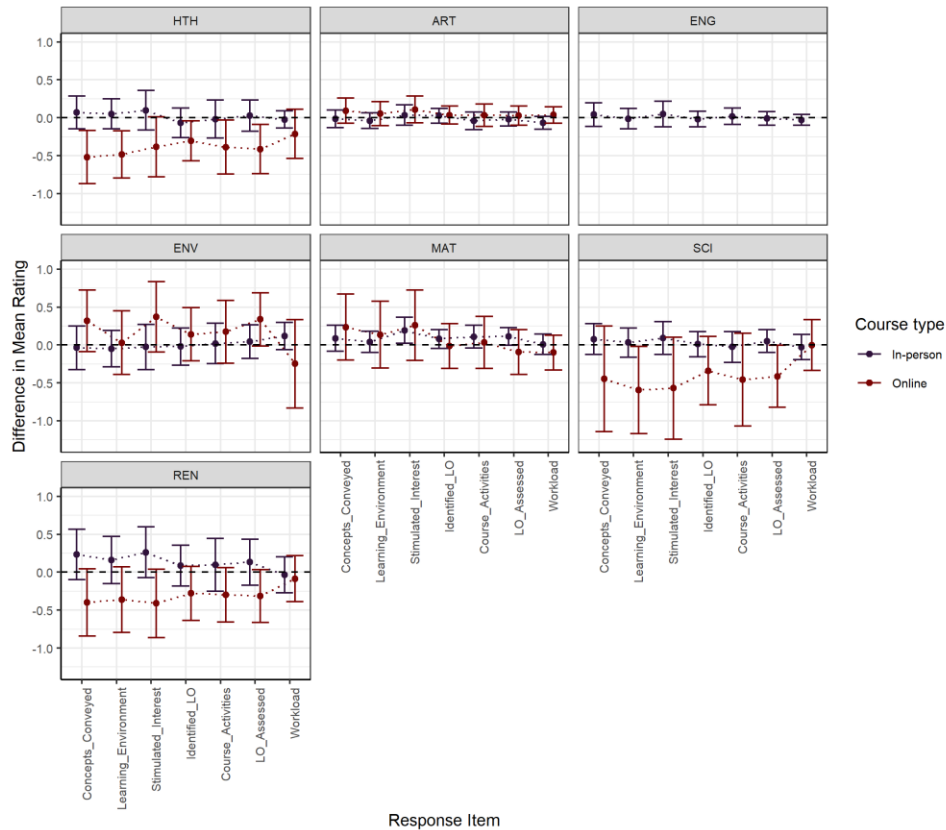
Figure 28: Difference in mean Winter 2022 ratings for male and female instructors by response item, course type, and Faculty of course offering. Error bars correspond to 95% confidence intervals.

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

Another relationship of interest at the Faculty level was the difference in mean ratings for male and female instructors accounting for instructor appointment type. Overall, Figure 29 shows close clustering of mean ratings at the zero-line indicating the differences are not statistically significant. In Health, ratings assigned to female associate and sessional instructors are slightly higher (up to 0.05) across all response items, but again it is important to be mindful of the sample size before drawing firm conclusions (refer to Figure 7: nF=50 & nM=40 overall in Health). In Environment, it appears that male sessional and lecturer instructors receive higher ratings than their female counterparts, however the difference is not significant (sample size nF=33 and nm=60 see figure 7).
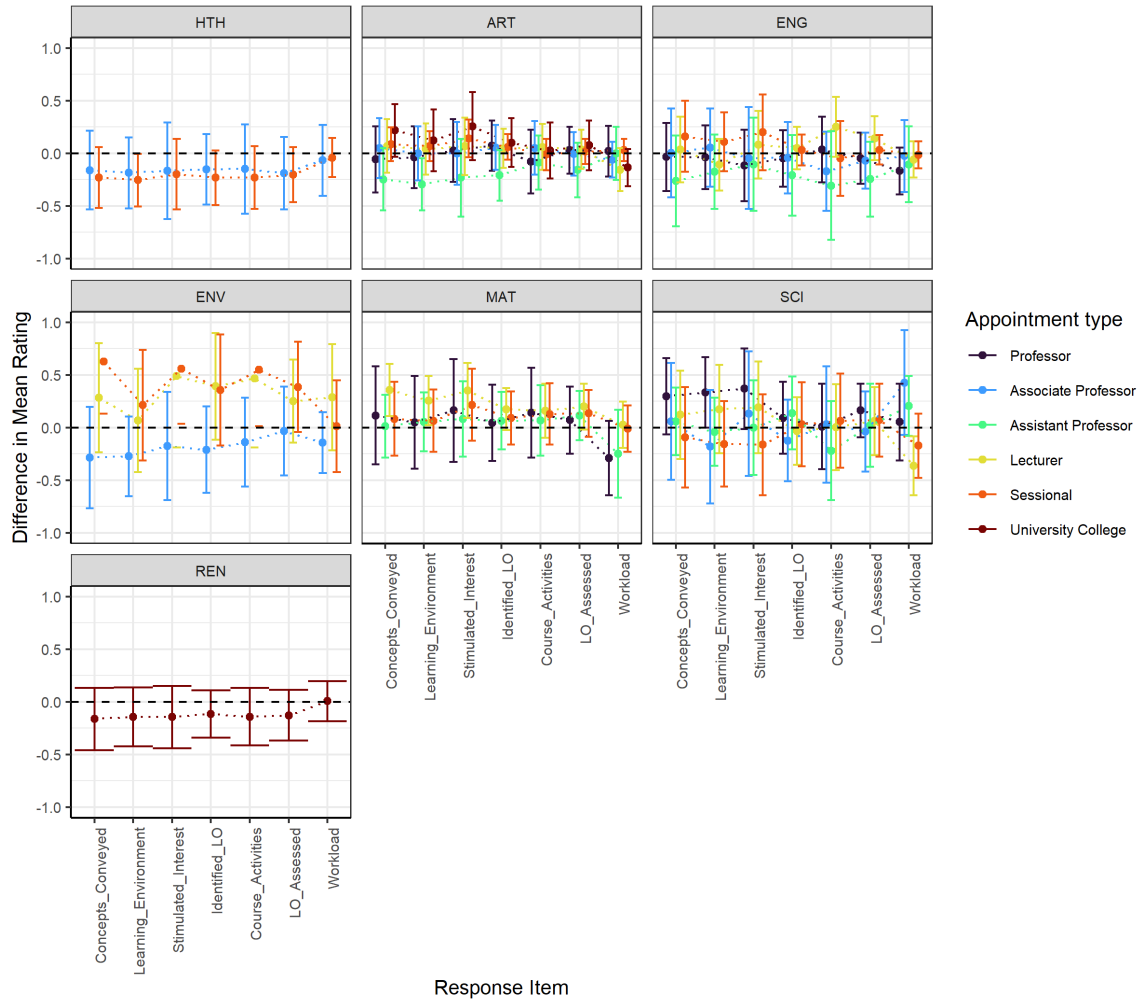
Figure 29: Difference in mean Winter 2022 ratings assigned for male and female instructors by response item, instructor appointment type, and Faculty of course offering. Error bars correspond to 95% confidence intervals.

Difference in MR for Male and Female Instructors by Class Size

In Figure 30 the difference in mean ratings for male and female instructors across all six survey items is plotted, accounting for class size. The larger confidence intervals (brown and orange) reflect smaller sample sizes and therefore should be interpreted with extra caution. Overall, we can see that the difference in mean ratings for male and female instructors accounting for class size is close to 0, ranging from -0.1 to +0.1. To put this in the context of the five-point Likert scale, the largest difference (seen in Course_Activities in classes with fewer than 50 students) would translate into a difference between 4.2 and 4.1 for this single response item, with female instructors getting the higher score. All other differences are smaller.
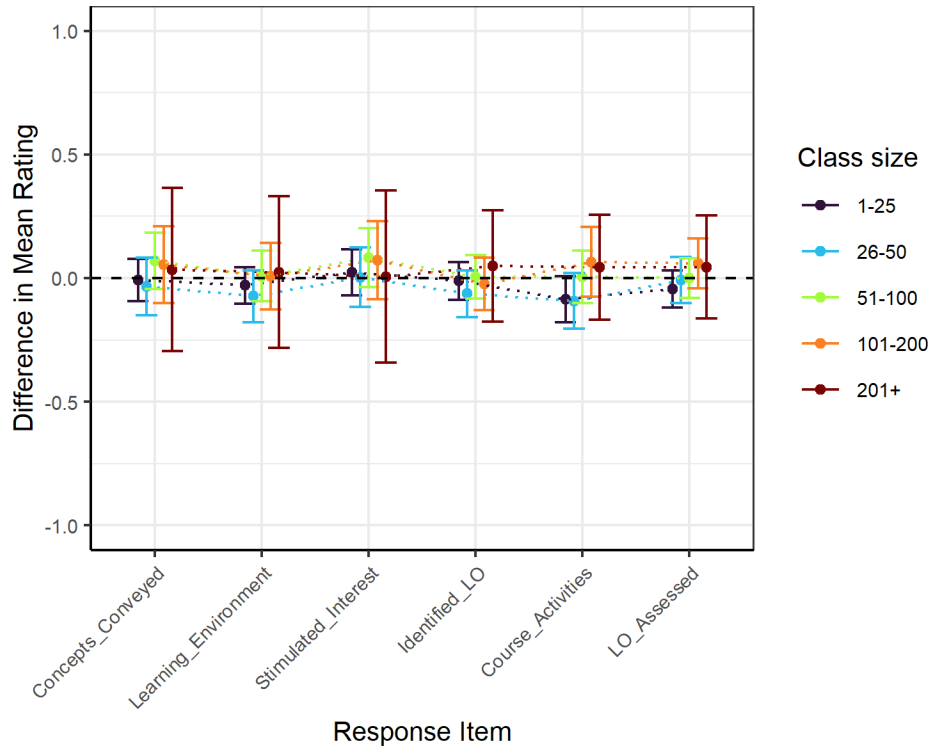
Figure 30: Difference in mean Winter 2022 ratings for male and female instructors by response item and class size. Error bars correspond to 95% confidence intervals.

In general, female instructors received slightly higher mean ratings across most items in smaller classes (1-50 students) as evidenced by the black and blue dotted lines. But this difference is incredibly small; in all cases it is less than 0.1. On the other hand, this plot indicates that male instructors seem to receive slightly higher mean ratings in larger classes (51-100 students and 101-200 students) as evidenced by the orange and green lines. But again, the difference in mean ratings is relatively small when considered on the five-point scale (<0.1) and is not statistically significant.

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

Figure 31 displays the difference in mean ratings for male and female instructors by instructor's time in Canada. The difference in mean ratings is not statistically significant accounting for time spent in Canada or instructor sex. In general, the difference in mean ratings for male and female instructors is less than 0.3 accounting for time spent in Canada. We see that female instructors who report being in Canada <5 years receive a mean rating that is 0.1-0.2 points higher across most response items as compared to male instructors who also report being in Canada for <5 years. A similar pattern is observed for female instructors who report being in Canada for 5+ years; receiving a mean rating that is about 0.1 points higher as compared to male instructors who also report being in Canada 5+ years. The difference in mean ratings for male and female instructors who are born in Canada (brown line) is closer to 0.
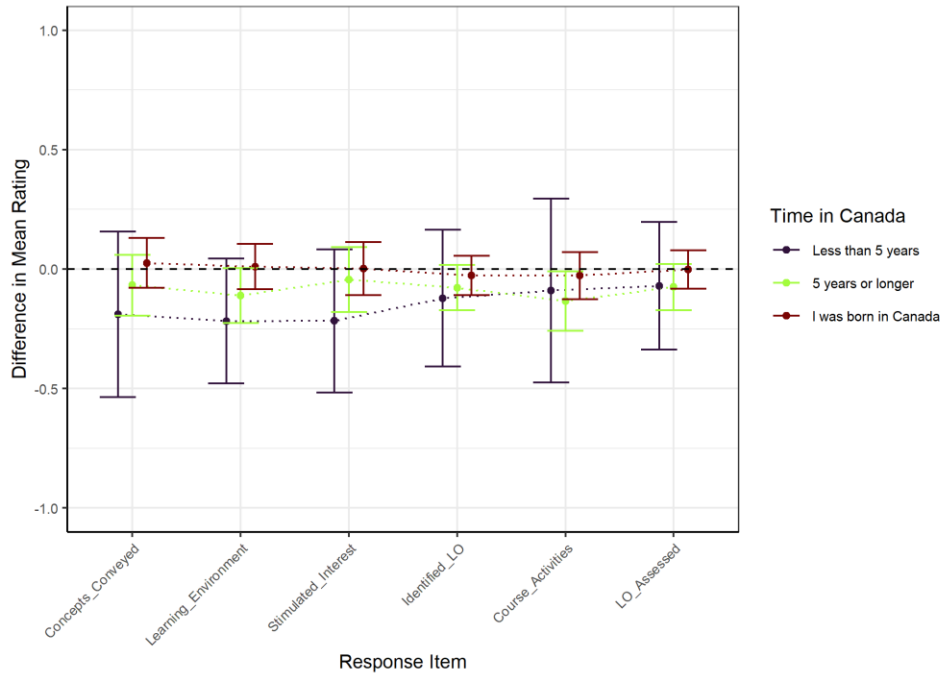
Figure 31: Difference in mean Winter 2022 ratings for male and female instructors by response item and instructor's time spent in Canada. Error bars correspond to 95% confidence intervals.

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

Figure 32 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, the differences in mean ratings for male and female instructors accounting for their appointment type and the class size, is not statistically significant (as evidenced by the confidence intervals crossing the 0 line for all plots). In most cases, the difference in mean ratings is very close to 0, and across nearly all cases, we observe that differences in ratings are <0.5 points on the 5-point Likert scale.

Though not statistically significant, it is worth noting that the plot in column 4 row 4 reveals that classes with 101-200 students taught by male lecturers received higher mean ratings across all six response items except Identified_LO (up to 0.5 points higher). This is consistent with 2018 pilot test findings in which classes with 101-200 students taught by female lecturers received lower mean ratings (up to 0.5 points lower). This is a finding that we will continue to examine in future analyses.
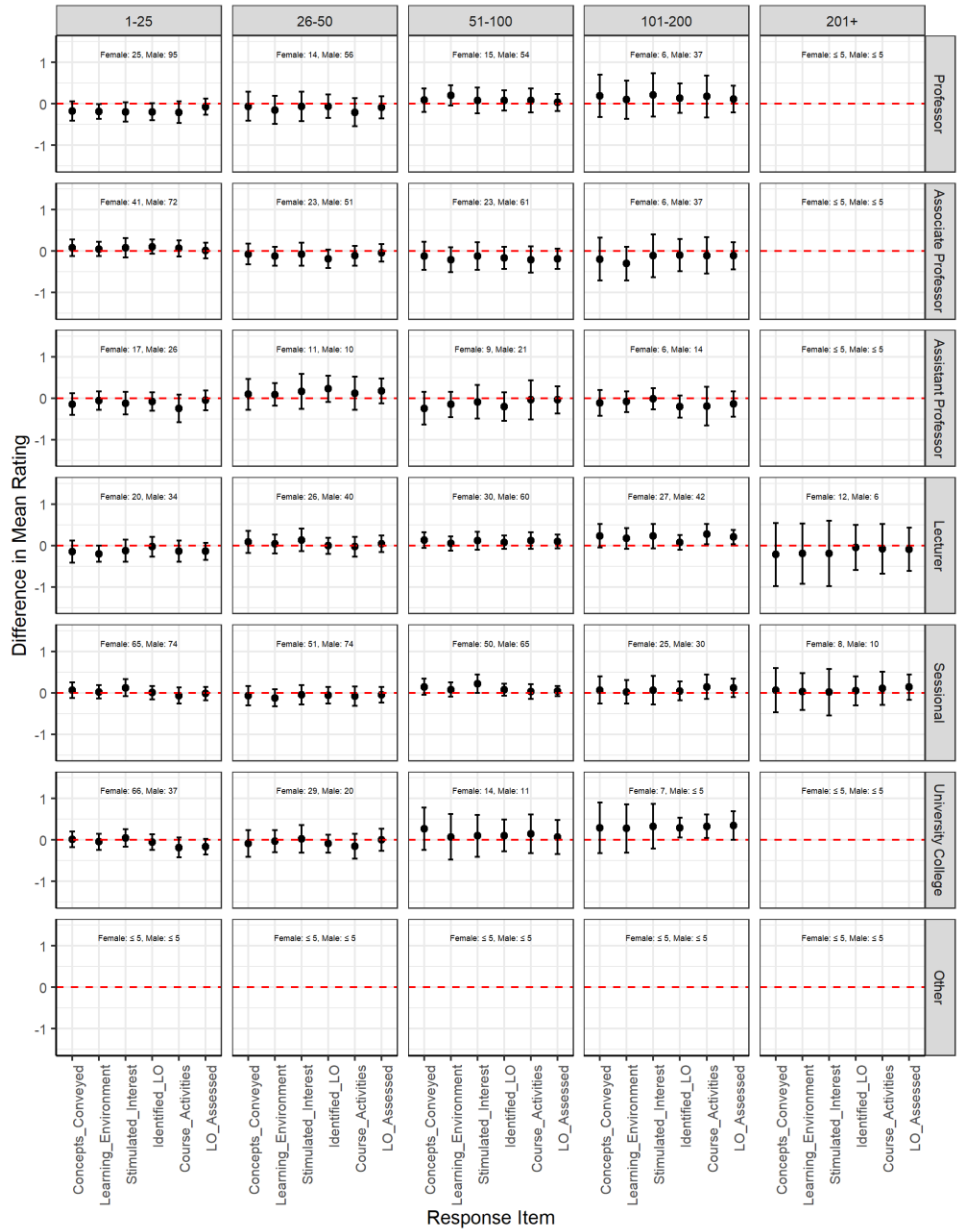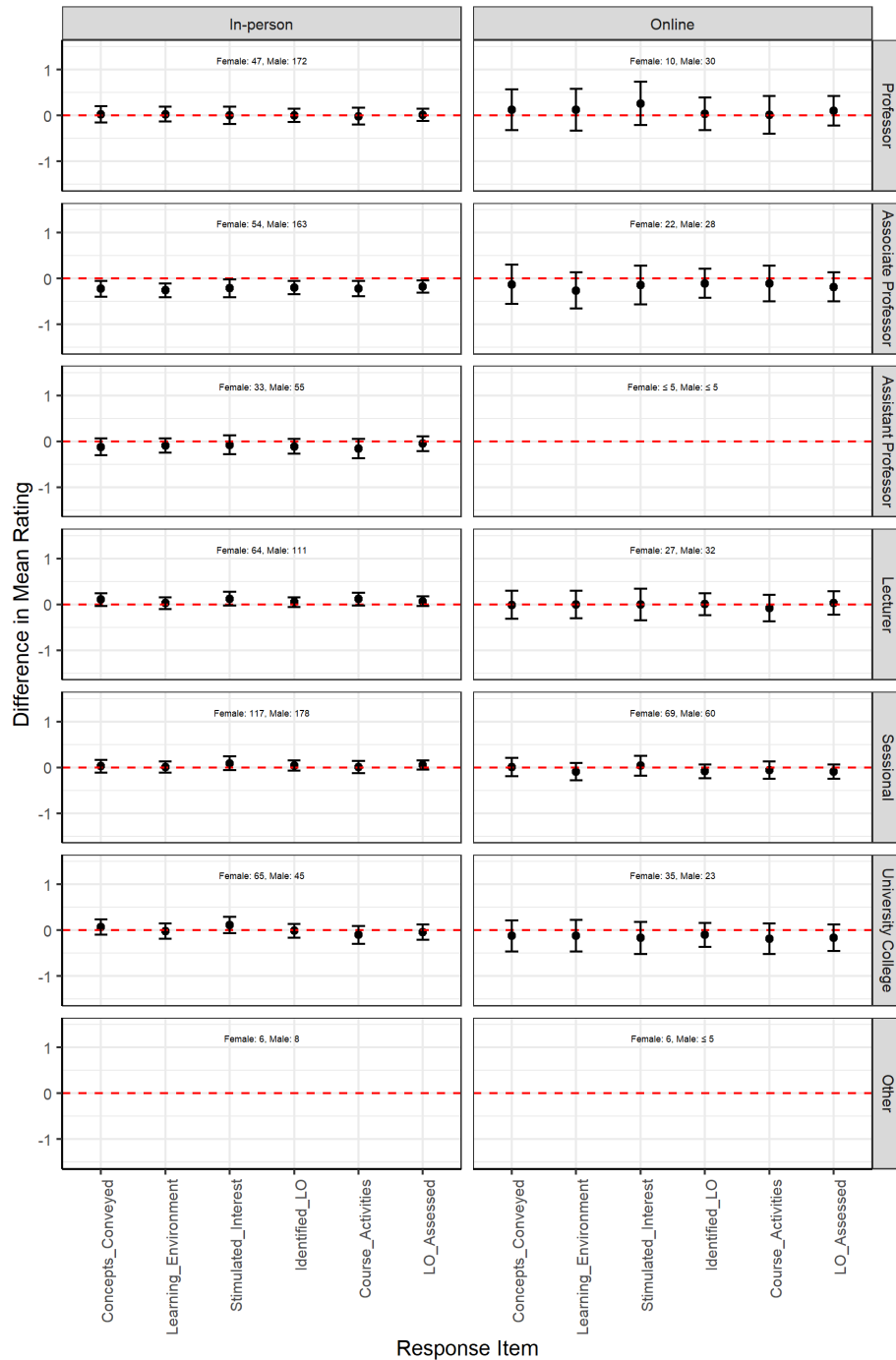
Figure 32: Difference in mean Winter 2022 ratings for male and female instructors by response item, instructor appointment type, and class size. Error bars correspond to 95% confidence intervals.

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

Figure 33 shows that the difference in mean ratings for female and male instructors, accounting for course type (online versus in-class) or instructor appointment type, is not statistically significant and close to 0 across all six response items. Most confidence intervals are large, again reflecting small differences in average scores and accounting for whether the course is in-person or online and professor appointment type and sex.

Figure 33: Difference in mean Winter 2022 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 34 shows the differences in mean ratings for female and male instructors accounting for course type (online versus in-class) and class size. Again, we observe a similar pattern to Figure 28 and Figure 29: the difference in mean ratings across all 6 items, accounting for class size and

course type, is not statistically significant, as the confidence interval crosses zero in all cases. Figure 34 also indicates extremely small differences in mean ratings (<0.5 points difference, even at the extreme ends of the error bars).
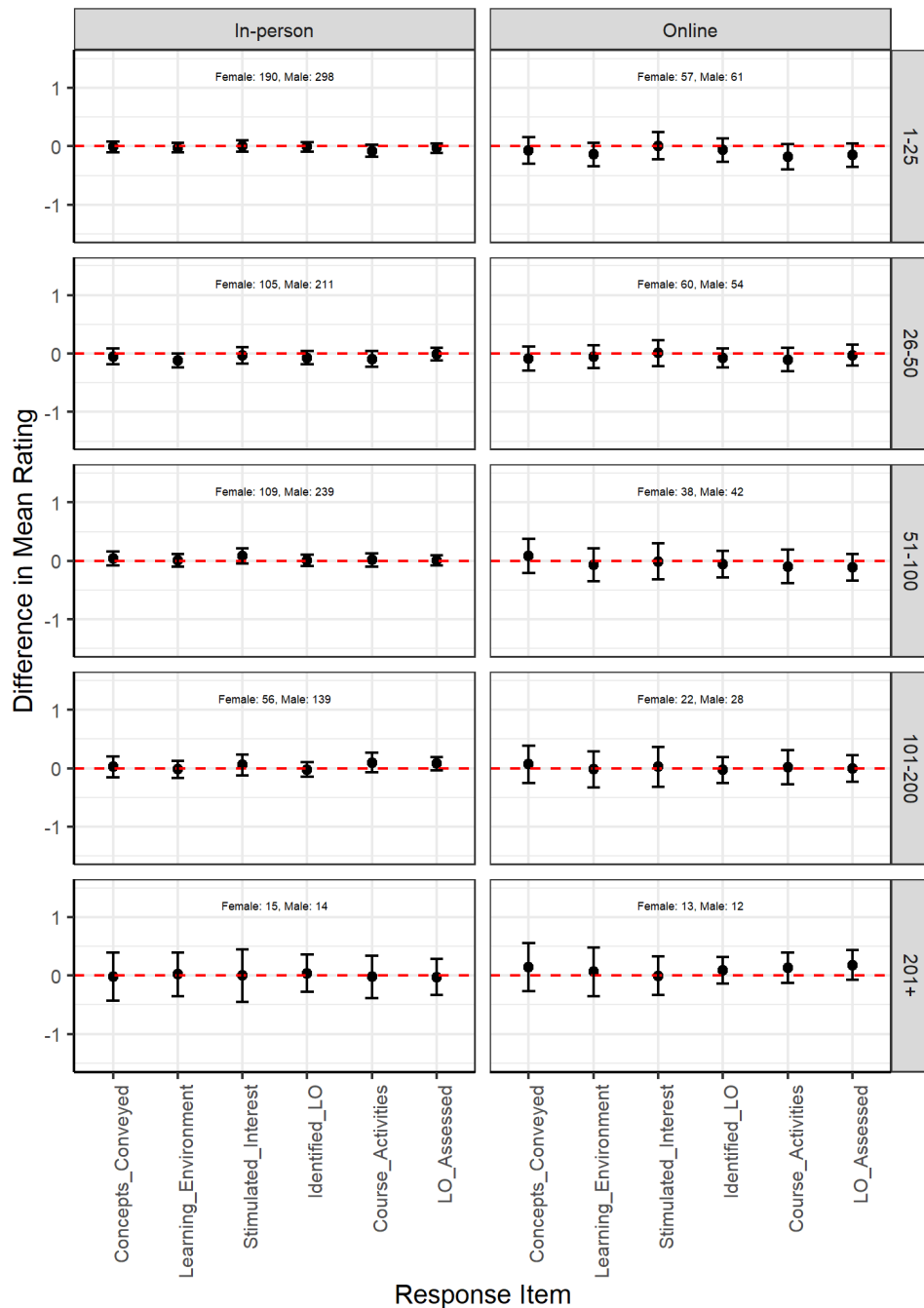
Figure 34: Difference in mean Winter 2022 ratings for male and female instructors by response item, class size, and course type. Error bars correspond to 95% confidence intervals.

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 2022 student responses to SCP surveys. The following key findings emerged from this analysis:

- The finding that many Winter 2022 instructors (42.6%) did not complete the Equity Survey means that any analyses based on racial or Indigenous identity are based on incomplete data, and therefore must be interpreted with great caution.
- The dataset included in this analysis contained substantially more white instructors than racialized instructors across all Faculties, and only 39 of the instructors who responded to the Equity Survey reported being in Canada <5 years.
- In general, male instructors outnumber female instructors across all ranks.
- The bulk of Winter 2022 courses included in this analysis (84%) had 100 or fewer students.
- Most of the Winter 2022 courses included in this analysis (79%) were delivered in person, as opposed to online.
- 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 2022 students tended to assign scores between 4-5 on the five-point scale across all SCP survey items. This finding was also true at across each Faculty and Renison.
- The smaller Winter 2022 courses included in this analysis tended to receive higher mean ratings than the larger courses.
- We were unable to meaningfully examine the impact of instructor Indigenous identity on mean ratings due to the small size of the dataset (n=8).
- Overall, we found very small but statistically significant differences in mean ratings (0.1-0.2 points) between white and racialized instructors in two of six response items: Concepts_Conveyed and Stimulated_Interest.
- We found some very preliminary support for a possible hypothesis for the differences in mean ratings we observed; there was some evidence to suggest the differences may be explained by time spent in Canada, a proxy variable we used to examine possible language and/or cultural biases being associated with lower scores assigned by students. This suggest that time spent in Canada may be a factor contributing to slightly lower mean ratings (differences of <0.2 points on the 5-point scale were observed).
- Deeper analyses of specific instructor groups and mean ratings did not reveal evidence of statistically significant differences in scores for racialized and white instructors. This may be due to the 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.

- The small differences observed in Winter 2022 mean ratings for male and female instructors was found to be statistically non-significant accounting for class size, course type, or instructor appointment.
- Although the overall difference in mean ratings assigned by students to male and female instructors was not statistically significant accounting for class type, course size, or instructor appointment type, we did observe that students in classes of 101-200 students assigned slightly higher ratings (0.5 points or less) in five of six response items to male instructors appointed as lecturers as compared to female instructors appointed as lecturers. However, it's worth noting that the sample size was small.

Concluding Remarks

The Winter 2022 analysis of the SCP survey data is the first to explore the impact of systemic bias on SCP ratings in the UW context. The TAP office will continue its commitment to explore these issues in-depth so that we can continue to understand how SCP ratings are impacted by bias; a second study is planned for Winter 2023 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 2022 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 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. As Wang & Gonzalez (2020) explain:

  > *Most faculty positions require a doctoral degree. Acquisition of this degree demands around 20 years of education and enormous financial cost. Historically, in the USA, White Americans have had access to more financial resources and students from White American families have had more opportunities to receive education and follow the academic path to a doctoral degree. As a result, more White Americans (especially, White men) have pursued careers requiring higher education levels and complex skills such as managers, engineers, lawyers and college professors. This White American image in those professions has been enhanced and spread to society through the media and literature (e.g. books, TV shows and movies). The academic profession at the university level has long been dominated by White Americans because of high educational requirements.* (p.846-847)

- Mirroring what we observed in the 2018 pilot test, males included in this study outnumbered females across all ranks. 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).

- This is the first analysis to use Equity Survey data. Our work was impacted by the large number of Winter 2022 instructors (42.6%) who did not complete the Equity Survey. It is our hope that future such analyses will benefit from higher Equity Survey response rates.

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 (2022). 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

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. (2022). 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 2022 instructors who selected "Yes" for the Equity Survey question, "Do you identify as an Indigenous person?"	
	Not indigenous	Winter 2022 instructors who selected "No" for the Equity Survey question, "Do you identify as an Indigenous person?"	
	NA	Winter 2022 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 2022 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 2022 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 2022 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 2022 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 2022 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 2022 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 2022 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 2022 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 2022 instructors for whom Human Resources data on sex indicated female	
Instructor group	Professor	Winter 2022 instructors for whom	In the case

		Human Resources data on primary job profile or additional job profile indicated "Professor"	where a Winter 2022 instructor belonged to more than one instructor group, the order of overriding is "Professor, Associate Professor, Assistant Professor, Lecturer, Sessional, University College".
	Associate Professor	Winter 2022 instructors for whom Human Resources data on primary job profile or additional job profile indicated "Associate Professor"	
	Assistant Professor	Winter 2022 instructors for whom Human Resources data on primary job profile or additional job profile indicated "Assistant Professor"	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.
	Lecturer	Winter 2022 instructors for whom Human Resources data on primary job profile or additional job profile indicated "Lecturer" or "Clinical Lecturer"	
	Sessional	Winter 2022 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 2022 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 2022 courses offered by the Faculty of Health	
	ART	Winter 2022 courses offered by the Faculty of Arts	
	ENG	Winter 2022 courses offered by the Faculty of Engineering	
	MAT	Winter 2022 courses offered by the Faculty of Mathematics	
	SCI	Winter 2022 courses offered by the Faculty of Science	
	REN	Winter 2022 courses offered by Renison University College	
	Other	Winter 2022 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
$$\mu_i \overset{iid}{\sim} Normal(0, \sigma_\mu^2)$$
- 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_r 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 ω_c 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:

$$y_{icjk} = \beta_0 + \beta_1 X_i + \mu_i + \omega_c + \alpha_{cj} + \epsilon_{icjk}$$

Where α_{cj} 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:

$$y_{ijk} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_1 X_2 + \mu_i + \alpha_j + \epsilon_{ijk}$$

Where X_1 and X_2 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|instructor_id) + (1|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