REPORT TO UBC SENATES: FINAL REPORT ON STUDENT EXPERIENCE OF INSTRUCTION RECOMMENDATIONS

Report to Okanagan Senate Learning and Research Committee and Vancouver Teaching and Learning Committee – September 2022

Report to Okanagan and Vancouver Senates – October 2022

Table of Contents

Introduction and background

In February 2019, a joint Student Evaluation of Teaching (SEoT) working group formed with membership across both UBC Okanagan (UBCO) and UBC Vancouver (UBCV) campuses. Working under the auspices of the UBCO Senate Learning and Research and the UBCV Senate Teaching and Learning committees, the group had the following remit:

1. Interrogate anonymized UBC data, to determine if there is evidence of potential biases.
2. Review and assess the recent literature on the effectiveness of SEoT, with particular reference to potential sources of bias in evaluations.
3. Review the University questions (University Module Items (UMIs) used in SEoT in light of the data and available literature, recommending changes where appropriate.
4. Propose recommendations for appropriate metrics, effective analysis and presentation of data to support SEoT as a component of teaching evaluation.
5. Consider the implications any proposed changes may have on other components of teaching evaluation.

After robust analysis and consultations conducted between March 2019 and April 2020, the SEoT working group presented a report to both the Okanagan and Vancouver Senates in May 2020. Included in the report was information about the working group's membership and consultation process, an annotated bibliography of research on bias in student evaluations of teaching, studies done at UBC on bias based on binary sex data[1], and information about a new set of metrics used in reporting SEoT results.

In addition, and most pertinent to the present purpose, the report included sixteen recommendations about student evaluations of teaching, which were endorsed by both Senates; see Appendix 1. In the Fall of 2020, two new committees were formed to oversee the process of implementing these recommendations: a Steering Committee and an Implementation Committee. Since one of the recommendations in the original working group's report was to change the name of the process from "student evaluations of teaching" to "Student Experience of Instruction" (SEI), these new committees are called the SEI Steering and SEI Implementation committees.

The SEI Steering Committee is made up of senior leaders, faculty and students on both campuses, and provides strategic guidance and oversight for the Implementation Committee, which is tasked with operationalizing the implementation of the recommendations. Please see Appendix 2 for membership of these groups.

The Implementation and Steering Committees were put in place in order to implement the recommendations from the previous SEoT working group. They are completing their work as of

[1] This variable was pulled from administrative data, which only recorded responses as binary, M or F, at that time.

early Fall 2022, and this report presents a summary of all implementation work, with a particular focus on what has been done since the Report to Senates on Progress on the SEI Recommendations in May, 2021.

Summary of implementation work

Since early Fall 2020, the Implementation Committee has worked with multiple individuals and units on the recommendations put forth from the SEoT working group. In addition, the Implementation Committee created a number of resources and events to communicate changes to the student evaluation surveys and work to date across both campuses, including a new website on student experience of instruction (seoi.ubc.ca), and two cross-campus open forums held on March 10th and September 28th 2021.

Over the course of the project there has been a strong focus on changes to the UMI questions, which were completed and launched in September 2021. This committee has also undertaken work on recommendations related to the need for additional data and analyses to address questions related to bias in SEI data at UBC, as well as exploring how UBC could adopt a more integrative approach in the evaluation of teaching.

We focus in particular below on the process for revising the UMIs on both campuses, and data analyses on SEI results that have been completed so far. We then discuss the status of each of the sixteen recommendations from the Student Evaluations of Teaching working group, endorsed by both Senates in May 2020.

Engagement and pilot process for revised University Module Items

The SEoT working group recommended that the questions on end-of-course student surveys be focused on the student experience rather than the evaluation of teaching, as students are in the best position to offer feedback on the former. The working group proposed six new core university questions, based on the six questions used in the Vancouver survey, to solicit feedback from students on their experiences in courses. In addition, the working group recommended that further data collection and analysis be undertaken for a proposed new question on feedback that would replace a previous question from the Vancouver survey on the fairness of assessment of student learning (see details on the proposed questions below, under *Updates on Recommendations*).

In taking this work forward, PAIR, in consultation with the SEI Implementation Committee, developed a plan to evaluate and test the proposed core university questions within the UBC

community, from January-July 2021.[2] This process began with focus groups with participants across both campuses, some with students (16 groups, 116 students total) and some with faculty (8 groups, 40 faculty total). The focus groups introduced a set of revised UMI questions and asked the participants how they interpreted the questions, how students would respond, and any suggestions they had for revision.

The next step was to conduct 29 one-on-one interviews with students who had not participated in the previous focus group sessions, in which students were asked to speak aloud to verbalize how they interpreted each of the six questions, what types of examples about the course they recall when responding to the question, and what information they recall and consider when responding to each question.

Data from the focus groups and interviews were coded, and revised UMI questions were then pilot-tested in a survey in which 333 students participated, across both campuses. PAIR then used Item Response Theory and Differential Item Functioning (DIF) to evaluate the performance of the questions on the pilot survey. The results from the quantitative analysis suggested that the revised UMIs were functioning better than the previous ones in that each of the questions seemed to be contributing more equally to the overall information from the surveys (whereas previously item 6 contributed to most of the statistical information). In addition, though on some methods of performing DIF analysis there were indications of some differences in how students responded to the questions based on class size and binary student gender, there was not consistency across these measures of DIF, and overall, the results were inconclusive.

The SEI Implementation Committee proposed a new set of UMI questions developed from this testing procedure for approval by Senate Committees at both UBCO and UBCV. These were approved in the summer of 2021 and implemented in SEI surveys starting in the Fall term 2021.

The UMIs currently in use at both UBCO and UBCV are:

1. Throughout the term, the instructor explained course requirements so it was clear to me what I was expected to learn.
2. The instructor conducted this course in such a way that I was motivated to learn.
3. The instructor presented the course material in a way that I could understand.
4. Considering the type of class (e.g., large lecture, seminar, studio), the instructor provided useful feedback that helped me understand how my learning progressed during this course.
5. The instructor showed genuine interest in supporting my learning throughout this course.
6. Overall, I learned a great deal from this instructor.

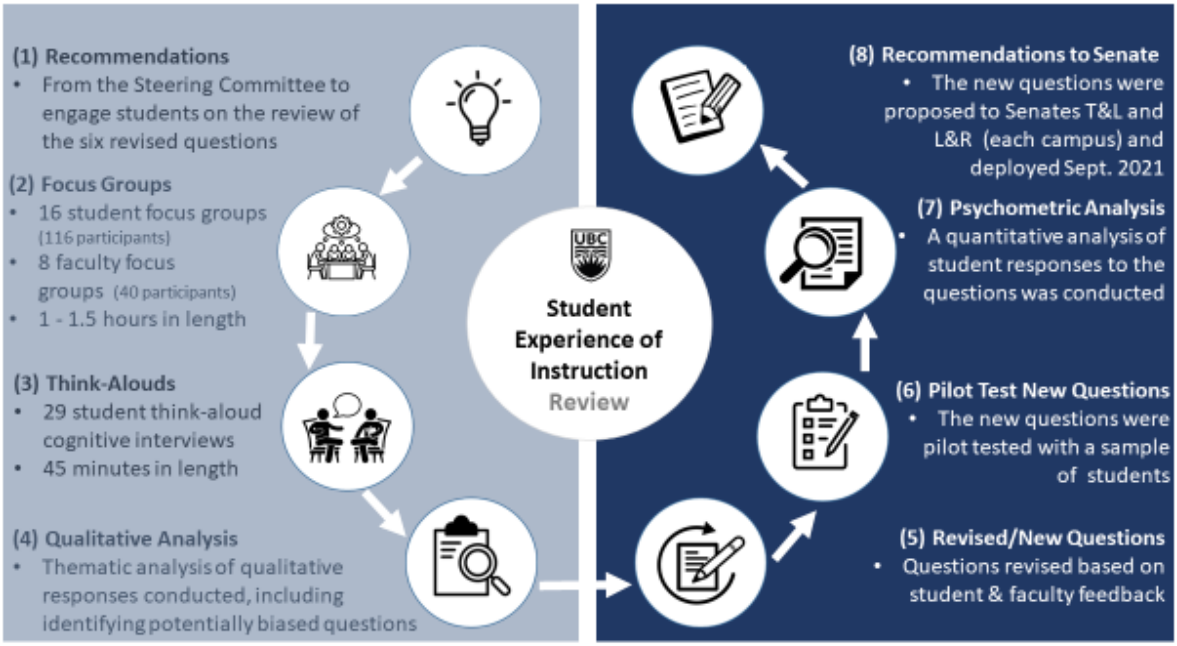
[2] See the following website for information on the process of testing the university module questions, and a detailed report on the results of the testing: https://seoi.ubc.ca/upcoming-changes/revised-university-module-questions/

Response options for all questions above are: *strongly agree, agree, neutral, disagree, and strongly disagree.*

In addition, a set of open-ended, text-based questions are included on surveys on both campuses:

1. Do you have any suggestions for what the instructor could have done differently to further support your learning?
2. Please identify what you consider to be the strengths of this course.
3. Please provide suggestions on how this course might be improved.

Visual representation of the process and timeline for revising the UMIs are provided below. Please see Appendix 3 for a comparison of previous UMIs and new UMIs for each campus.



Timeline for process (2021)	Jan	Feb	March	April	May	June	July	Aug	Sep
Focus groups: students & faculty									
Conduct Think aloud Sessions									
Qualitative Thematic analysis									
Revised/New Questions									
Pilot new questions									
Psychometric Analysis									
Senate committees for review									
Deploy the final questions									

Data analyses of SEI results 2021/22

With the approval of the Okanagan and Vancouver Senates, the new six UMI questions were implemented in the SEI surveys across both UBC campuses starting in the Fall of 2021; an outline of the previous and updated questions is available in Appendix 3. The following sections highlight the results of the analyses conducted using these data. For the full data report please see Appendix 4.

1.0 Methods

To conduct the analyses, a sample SEI data set was created by randomly selecting 100 course/sections surveys from each of five fields of study (Sciences, Humanities, Social Sciences, Engineering and Health Sciences). Stratified random sampling by field of study is key to ensure adequate representation across fields of study. A list of academic units/programs within each field of study is shown in Appendix 4. The SEI data were linked with administrative data to obtain additional variables of interest, e.g., class meeting time, instructor gender, class size.

We attempted to use the Employment Equity Survey data to obtain other variables of interest e.g., gender identity, ethnicity, disability, etc. However, about half of the instructors who taught in 2021 W1 were missing employment equity data. Furthermore, for those instructors with Equity Survey data, available gender data was not different than what is in the SEI data (binary), with sparse data on other gender categories. Because we could not ascertain the randomness of missing equity data, which could potentially affect how different groups were represented in the dataset, employment equity data were excluded from further consideration.

The final SEI sample dataset comprised of 11,032 student responses to the six UMI questions. Tables 1.a, 1.b, 1.c show the distribution of the dataset, used in the final analysis, by course, instructor, and students' attributes.

Table 1.a: Distribution the 2021W1 SEI Responses by Field of Study & Year Level

Field of Study	Number of responses
Engineering	1,892
Health Sciences	1,520
Humanities	1,784
Sciences	3,090
Social Sciences	2,746
Total	11,032

Year Level	Number of responses
1st	3,181
2nd	3,086
3rd	2,637
4th	969
5th	1,159

Table 1.b: Distribution the 2021W1 SEI Responses by Student Demographics

Campus	Number of responses
UBCO	2,134
UBCV	8,898

Student Gender	Number of responses
Female	6,542
Male	4,490

Table 1.c: Distribution of the 2021W1 SEI Responses by Instructor Attributes

Instructor Rank	Number of responses
Assoc. Prof	1,845
Asst. Prof	2,917
Lecturer	1,754
Professor	1,933
Sessional	2,583

Instructor Gender	Number of responses
Female	4,211
Male	6,821

Table 1.d: Distribution the 2021W1 SEI Responses by Course Attributes

Class Meeting Time	Number of responses
Before 11:00 AM	3,635
After 11:00 AM	7,397

Class Size	Number of responses
< 100	4,519
>= 100	6,513
1 - 49	2,427
200+	2,891

We examined the data using three approaches, Differential Item Functioning (DIF), Item Response Theory (IRT), and Generalized Linear Mixed Models (GLMM). DIF is used to determine if students respond to the SEI questions differently across groups, such as class size, meeting time, campus, year level, and student or instructor gender. The IRT approach enables us to determine how students are interacting with the new SEI questions, how well these questions function across different attitudinal levels among students, and how well the response options work for each question. Finally, GLMM can be useful for examining data that are not continuous, such as categorical responses (e.g., strongly agree to strongly disagree) or binary responses (positive/negative). GLMM is also appropriate for examining data that are clustered in some way, e.g., students nested in courses or fields of study (Rabe-Hesketh and Skrondal, 2010).

2.0 Results

2.1 Differential Item Functioning

We used multiple DIF analysis approaches to examine how students respond to UMI questions, based on attributes in Table 1.a-d: the Mantel-Haenszel (M-H), logistic regression (binary), generalized linear model (ordinal) and IRT-based Lord's Chi-square test. If multiple tests indicate DIF is present, then the findings are more robust. Results are reported in Table 2.

Table 2: Differential Item Functioning (DIF) between different student, instructor and course attributes.

DIF Method	Campus	Student Gender	Class Size < 100 vs > 100	Class Size 1 – 49) vs 200+	Class Meeting Time Before 11 vs After 11	Year Level 1st, 2nd & 3rd vs 4th & 5th	Instructor Gender
Mantel-Haenszel*	Negligible	UMI 6 moderate	UMI 1 moderate	**UMI 1**, 4 (large) UMI 5, 6 moderate	Negligible	Negligible	UMI 3 moderate F
Lord's Chi-square Test	None	None	UMI 1	**UMI 1**, 2 & 6	None	None	UMI 3

Logistic (Binary)**	None	UMI 6 uniform F ***	UMI 1 uniform >100	**UMI 1**, 4, 5, 6 uniform >50	None	None	UMI 3 uniform F
GLM (ordinal)**	---	UMI 6 uniform F	UMI 1 uniform >100	All uniform >50	None	----	UMI 3 uniform F

* To determine the effect size (magnitude) of DIF we used delta MH and the following criteria: a) none or negligible DIF detected with absolute values of delta MH less than 1; b) moderate DIF detected with absolute values of delta MH between 1 to 1.5; and c) large DIF detected with absolute values of delta MH larger than 1.5.
** Logistic & GLM methods used to indicate direction and type of DIF, if moderate or large DIF detected by Lord's & M-H methods.
*** Type and direction of DIF, e.g., "uniform F" indicates uniform DIF favouring female students.

As shown in Table 2, DIF was either not detected or was negligible, for grouping by campus, class meeting time or year level. Moderate uniform DIF was detected for student gender by only one procedure, the M-H method (delta MH of 1.05 and p-value < 0.0001), but not by the IRT-based Lord's method. The M-H method detected that female student responses tended to be more positive to UMI question 6, "Overall, I learned a great deal from this instructor." However, because DIF was detected with only one method the results were inconclusive.

Across all four methods, UMI question 1, "Throughout the term, the instructor explained course requirements so it was clear to me what I was expected to learn" showed large DIF between the smallest and largest class sizes (enrolments of 1-49 compared with classes with 200+ enrolments). The direction of DIF indicated that responses were more positive for the largest class size over the smallest (delta MH of 1.73 and p-values of < 0.001 for the four methods). Similarly, UMI question 6 showed moderate uniform DIF between the smallest and largest class sizes, across all four methods (delta MH of 1.2 and p-values of 0.0354, 0.003, < 0.0001 and < 0.0001, for the four methods, respectively). The results for the other UMIs, comparing the smallest and largest class sizes, were different across the test methods and were therefore inconclusive.

There was moderate DIF detected (delta MH of 1.37 and p-values of < 0.0001 for all 4 methods) for question UMI 1 comparing class sizes over 100 to those below 100 (again favoring the larger class sizes). Finally, UMI 3, "The instructor presented the course material in a way that I could understand," showed moderate (bordering on negligible) uniform DIF (delta MH of 1.01 and p-values of 0.0004, < 0.0001, <.0001, and 0.0038, for the four methods, respectively) for instructor gender; female instructors received slightly more positive responses on this item.

2.2 Item Response Theory

IRT analysis enables us to determine how well these questions function across different attitudinal levels among students. Prior to running an IRT model, we need to meet a few model assumptions, one of which is unidimensionality. This was determined using factor analysis and an examination of the scree and variance plots. The results of the factor analysis showed that all

six UMIs had high factor loadings, representing one underlying construct being tested, in this case the experience of instruction.

Two-parameter IRT models estimate the location and discrimination parameters of the survey items along the attitudinal scale of respondents. We used a 2-parameter multi-level IRT (MLRT) model to account for variation between fields of study and assess the effect of other variables on student SEI responses, including course attributes and instructor demographics within fields of study.

Reliability estimates were consistent across approaches; Cronbach's alpha is a measure of scale reliability, which indicates internal consistency. For the 2021 survey items, Cronbach's alpha of 0.94 suggests a high survey reliability. Furthermore, an IRT conditional reliability curve is shown in Figure 1.

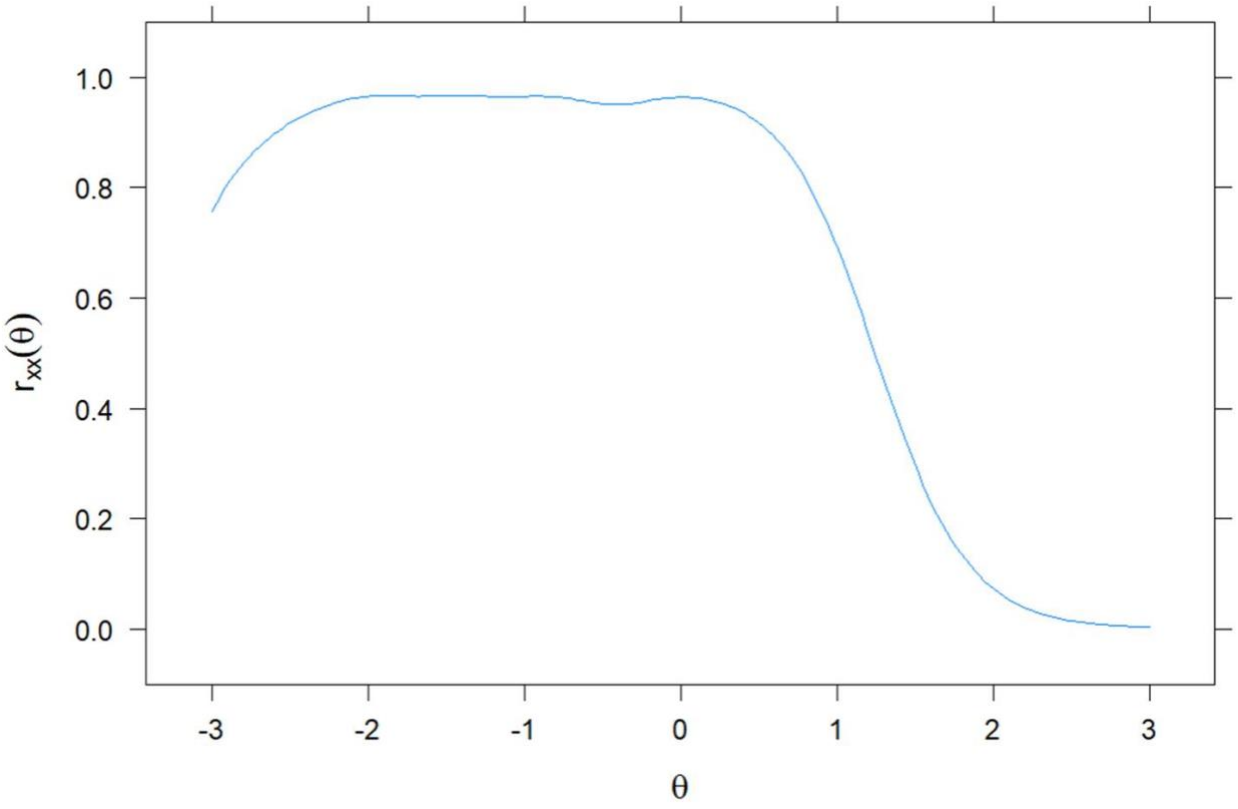


Figure 1. Conditional Reliability Curve

This is an overall reliability of a survey based on how well UMIs, overall, provide statistical information about the experience of instruction, and how precisely scores can be estimated across different values of attitudinal scale. Figure 1 indicates that score estimates are most reliable on a wide range of attitudinal scale (**θ**); with an overall IRT marginal reliability estimate of 0.84, which also suggests a high survey reliability.

The MLRT model was compared to a base IRT model (with no covariates) and to a one-level full model (with the same number of covariates as the MLRT model). The one-level full model performed better than the base model and the MLRT model (p-values < 0.0001). Based on these comparisons shown in Table 3, all references to the 2021 SEI survey IRT results are based on the 1-level full model.

Table 3: IRT Model Comparisons

Model	Criteria*					χ^2	df	p-value
	AIC	SABIC	HQ	BIC	logLik			
Base Model	112820.9	112944.8	112894.8	113040.2	-56380.46			
1-level	112617	112790.4	112720.4	112923.9	-56266.48	228	12	< 0.0001
MLRT	112883	113044.1	112979	113168	-56402.49			
1-level	112617	112790.4	112720.4	112923.9	-56266.48	272	3	< 0.0001

* AIC=Akaike Information, BIC=Bayesian Information, HQ=Hannan Quinn, logLik=Log Likelihood

The item discrimination parameter indicates the strength of the relationship between an item and the measured construct, i.e., experience of instruction. It determines the rate at which the probability of positively endorsing an item changes given the individual attitude/perception levels (Thorpe & Favia, 2012). Within the range 0.5 to 2.5 (Reeve and Fayers, 2005), the higher the discrimination parameter, the steeper the slope will be on the item characteristic curve, indicating a stronger ability to detect differences in the attitude/perception of respondents compared with less steep slopes. However, discrimination values above 2.5 don't add much to the slope of Item Characteristic Curves (ICC). Ideally, a balanced set of questions would have discrimination parameters of comparable magnitude, indicating a more balanced contribution of all questions to the survey information.

The item discrimination parameter estimates (slopes) for the 2-parameter IRT models are given in Table 4, for both the new UMI 2021 survey questions and the random sample from the pre-2021 version of the survey (the UMI questions in use prior to 2021). Typically, the larger the discrimination parameter, the steeper the slope, which implies that the item is more effective at discriminating among different attitudes along the continuum. Thus, for a given level of endorsement, UMI question 6 in the pre-2021 SEI survey with a discrimination parameter of 8.67

would have more than 5 times the contribution to the survey information compared to UMI question 1 with a discrimination parameter of 3.62.

Yet a discrimination parameter of 8.67 is quite high, which is an indication that the survey question is not working properly. A disproportionally large item slope indicates a disproportionally large contribution to the overall survey information.

Table 4: Item Discrimination Parameter Estimates

Data Source	Discrimination Parameter Estimates					
	UMI 1	UMI 2	UMI 3	UMI 4	UMI 5	UMI 6
UMI from the pre-2021 SEI Survey	3.62	5.38	4.15	2.02	3.28	8.67
UMI from the new 2021 SEI Survey	3.26	4.80	3.83	3.15	3.00	5.85

In Table 4, UMI question 4 in the pre-2021 survey that asks if *the evaluation of student learning was fair* (2.02), has the least relative discrimination. However, the new UMI 4 question asking about *useful feedback* has a discrimination parameter that is comparable to other items (3.15), indicating that this item discriminates as much as the other items, among different attitude/perception levels.

Overall, the parameter estimates in the new UMI questions (2021 SEI survey) have been improved compared to those reported for the pre-2021 survey, and they are now more consistent across the items.

Figures 2 and 3 display the Item Information Curves (IIC) for each of the new 2021 SEI survey UMI questions, and for the pre-2021 survey UMI questions, respectively. The IICs measure the statistical information an individual item contributes to the overall survey. The x-axis is the individual's level of endorsement; a person with an endorsement level of 2 has a more positive attitude regarding the course than someone with a level of -0.2. The y-axis indicates the magnitude of the information provided by each of the survey items. Higher information signifies higher precision (or reliability) in differentiating among respondents (Reeve & Fayers, 2005). In addition, items should be well spaced across the continuum (x-axis).

There are notable differences evident when comparing the item information curves in Figures 2 and 3. Figure 2 indicates improvement in the relative contributions of all new UMI questions to the overall survey information compared with the pre-2021 survey sample, notably for UMI questions 2, 3 and 4. Furthermore, the newly-worded 2021 UMI items shown in Figure 2 appear to differentiate across a broader range on the x-axis than the pre-2021 survey UMI items shown

in Figure 3. The y-axis scales differ between Figures 2 and 3 as a result of the disproportionately large UMI 6 discrimination parameter (8.67) in Figure 3. Although UMI 6 has a relatively large discrimination parameter estimate in the new 2021 survey and it appears to discriminate across a similar range on the x-axis, it displays sharp peaks on the information curve, which implies that the item is not functioning as well as it could. However, the new UMI 6 peaks (Figure 4) were less jagged and show improvement compared to that of the pre-2021 UMI 6 (Figure 5).

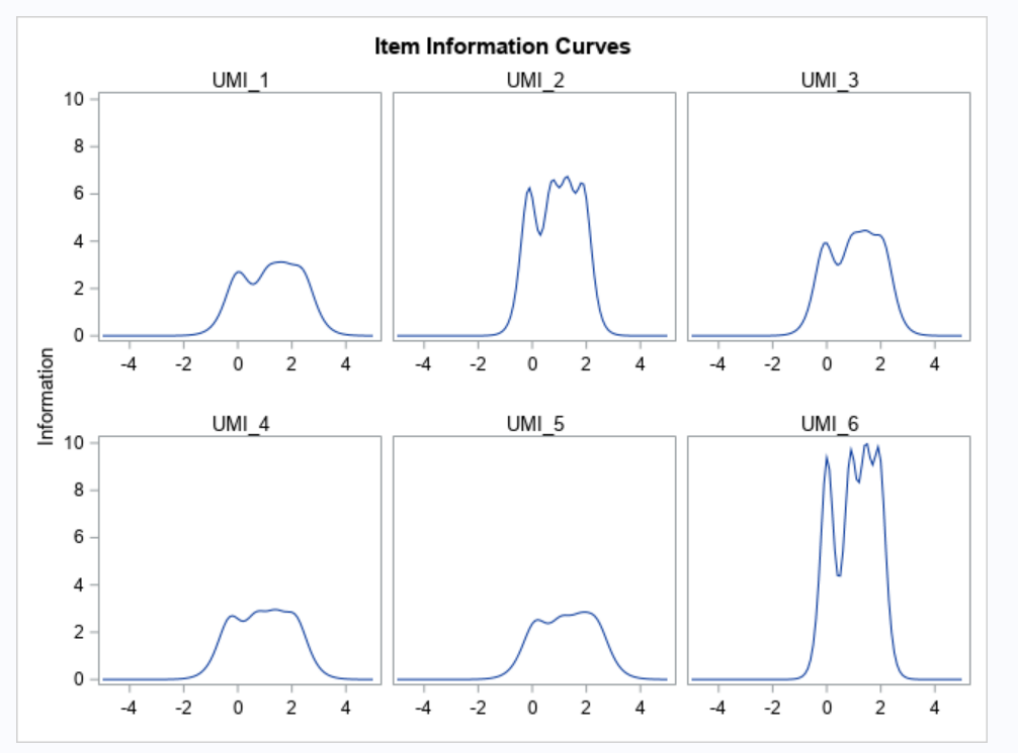


Figure 2: Item Information Curves for the new 2021 SEI Survey UMI questions

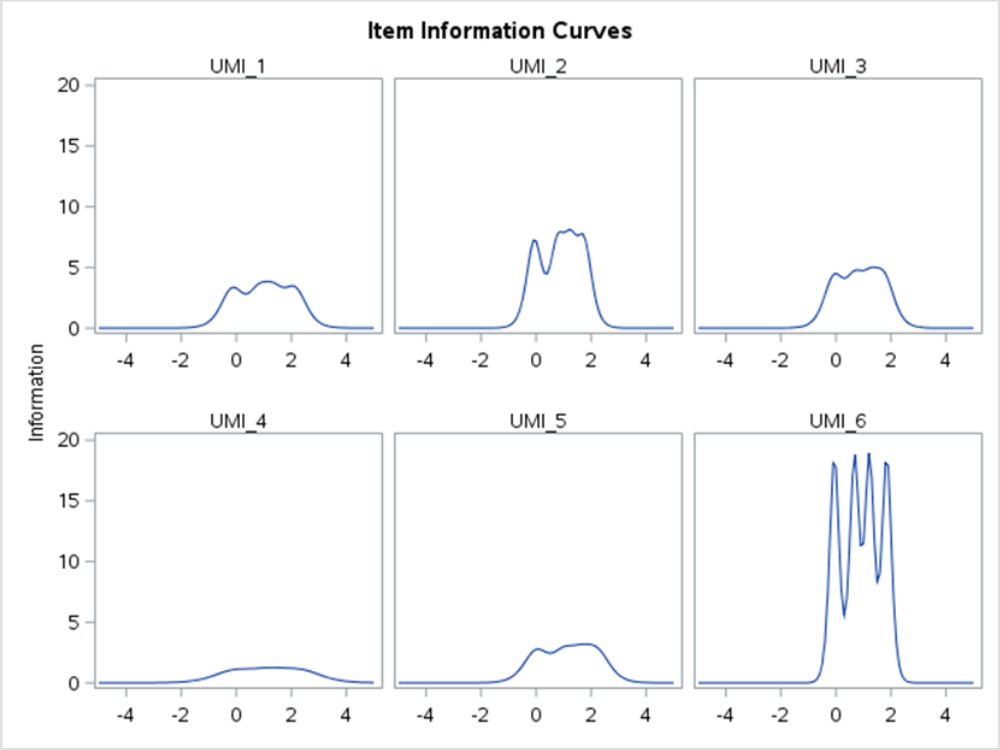


Figure 3: Item Information Curves for the pre-2021 SEI Survey UMI questions

Looking at Figure 3, the IICs for the pre-2021 UMI questions show that UMI 6 disproportionally contributes to the overall survey information; however, for the new set of UMI questions, the contribution of each item seems to be more consistent. Overall, the proposed changes to the UMI questions appear to have improved their relative discrimination among students with varying levels of endorsements for most items.

2.3 Generalized Linear Mixed Model

We used a Generalized Linear Mixed Model (GLMM) approach to model variation in SEI scores within 5 fields of study (Sciences, Humanities, Health Sciences, Engineering and Social Sciences; see Appendix 4 for a list of units/programs included). In this approach, respondents to SEI surveys are considered to be clustered within fields of study (grouping variable the GLMM with a random intercept). Proc GLIMMIX in the SAS statistical software was used to fit the cumulative logit of the probability of higher SEI ratings in the response profile (corresponding to the 5-point Likert scale) as a function of course attributes (year level and meeting time), instructor demographics (rank and gender) and student gender; and with the field of study as a grouping variable.

The estimated covariance parameters for the six UMI questions, which measure the variation in Field of Study effects, are shown in Table 5. For each UMI question, the estimated variance of

the Field of Study random intercepts is given along with standard error and p-value for testing if the variance is significantly different from zero.

Table 5: Estimated variance of the Field of Study random intercepts in the GLMM

Question	Covariate Estimate	Standard Error	Z value	p-value
UMI 1	0.0092	0.0081	1.13	0.1282
UMI 2	0.0302	0.0230	1.32	0.094
UMI 3	0.0314	0.0239	1.31	0.0943
UMI 4	0.0355	0.0266	1.33	0.0911
UMI 5	0.0315	0.0239	1.32	0.0936
UMI 6	0.0301	0.0230	1.31	0.095

The estimated values for all UMI questions in Table 5 are not significantly larger than 0 (p-values > 0.05) which indicates that there is no significant variation in the Field of Study effect on SEI ratings (no significant random effect). A Generalized Linear Model (GLM) across all fields of study (no field of study random intercept) was also fitted to the data. There are minor differences between the GLM and GLMM model. However, all subsequent data was reported for the GLMM – even though not significant for any of the UMIs (Table 5), it was used as it did explain some of the variance across other variables in the model. Tests of the model fixed effects are shown in Table 6.

Table 6: P-values for the model fixed effects

Question	Instructor Rank	Instructor Gender	Student Gender	Year Level	Meeting Time
UMI 1	**< 0.001**	0.050	**0.025**	**0.002**	0.055
UMI 2	**< 0.001**	0.142	**0.025**	**< 0.001**	0.105
UMI 3	**< 0.001**	**0.004**	**0.023**	**< 0.001**	0.643
UMI 4	**< 0.001**	0.080	0.071	**< 0.001**	0.154
UMI 5	**< 0.001**	**0.012**	0.148	**< 0.001**	0.109
UMI 6	**< 0.001**	0.266	**0.007**	**< 0.001**	0.225

Model parameter estimates and associated statistics for fixed effects are shown in Appendix 4. For all UMI questions, there were no significant differences in SEI ratings between course sections that met before or after 11:00 AM.

SEI ratings for 1st, 2nd and 3rd year courses were consistently significantly lower compared to 4th and 5th year courses. It is important to note that these differences are not due to Differential Item Functioning (see Table 2 for DIF results). Recall that DIF is conceptualized as occurring when survey respondents who have similar attitudes/perceptions on a measured trait respond differently due to construct-irrelevant factors; i.e., DIF analysis takes into consideration the sum of scores for all UMI questions as a measure of respondent attitude/perception.

Female instructors received relatively higher ratings compared to their male counterparts in UMI questions 3 ("The Instructor presented the course material in a way that I could understand") and 5 ("The instructor showed genuine interest in supporting my learning throughout this course"). However, the odds ratio for the two questions were relatively small (1.3 and 1.2, respectively). Chen, Patricia Cohen & Sophie Chen (2010) showed that odd ratios < 1.5 translate to small effect size. There were no instructor gender differences in the other 4 UMI questions.

Female students rated their experience of instruction significantly higher compared to male students in UMI questions 1, 2, 3 and 6. Again, though statistically significant, odds ratios were close to 1.0 (1.1 for UMI questions 1, 2, and 3 and 1.2 for UMI 6).

There were also differences in ratings depending on instructor rank for all UMI questions. However, differences between instructor ranks and their magnitudes vary across questions, but odds ratios were relatively small (< 1.4), with slightly higher ratings for assistant professors and lecturers. Also, it is important to note that instructor rank was based on SEI survey data which reports "Standard Job Title" and does not consider tenure or other relevant appointment information.

Finally, there were consistent and significant differences in SEI ratings between fields of study with Humanities rated higher compared to the overall average, but with odd ratios not exceeding 1.2 for all UMI questions.

3.0 Conclusion

The Item Response Theory (IRT) results indicated that the new UMI questions implemented in 2021 seem to function better than the old version of UMI questions. In the old version, UMI question 6 provided most of the statistical information for the overall survey, but did not differentiate broadly among respondents' attitudes/perceptions. Furthermore, the presence of sharp peaks in the item information curve indicates the item was not functioning well. The Item Information results were similar to those obtained in a 2021 pilot study (McKeown, Zumrawi & Pena, 2021) and provide further evidence that the new UMI questions are more consistent in

their contribution to the overall survey, and are more widespread across the attitudinal continuum (x-axis).

While most of the new 2021 survey UMI questions showed no DIF among different groupings by student, instructor or class attributes, UMI 1 exhibited moderate to large DIF, and UMI 6 exhibited moderate DIF between class sizes. Moderate DIF between genders was also detected for UMI 6, with female students positively endorsing that question more than male students (recall that only binary data were used for gender based on challenges with using Employment Equity Survey data in these analyses). However, this result was not consistent across test methods and thus was not conclusive. Negligible/moderate DIF in instructor gender was also detected for UMI 3, with female instructors receiving slightly more positive endorsement on this item, however, the direction (favouring female instructors) was consistent with previous studies at UBC (CTLT, 2010).

GLMM results showed that SEI ratings for 1st, 2nd and 3rd year courses were consistently significantly lower compared to 4th and 5th year courses. Also, female instructors received slightly higher ratings (on UMI 3 and 5) and female students rated their instructors slightly higher (on UMI 1, 2, 3 & 6) compared to their male counterparts. However, in both cases the effect sizes were small. Finally, there were also significant differences in ratings depending on instructor rank for all UMI questions. Differences between instructor ranks and their magnitudes vary across questions, but odds ratios were relatively small (< 1.4), mostly favouring assistant professors and lecturers.

Due to the lack of sufficient Employment Equity Survey data, we were not able to test how the new UMI questions function across other variables of interest, e.g., gender identity, ethnicity, disability, and more. Thus, and based on these results, we recommend that further IRT and DIF analysis be carried out on the new UMI questions. Furthermore, we will continue to monitor the Employment Equity Survey response rate and examine the randomness of missing data.

Status of all 2020 report recommendations

As noted above, in May of 2020 sixteen recommendations about Student Evaluations of Teaching were endorsed by both the UBCO and UBCV Senates. Most of the work to implement these recommendations has been completed. Some recommendations need to be addressed in an ongoing fashion, while one requires further review, consultation, and financial commitment beyond the scope of the implementation project.

Student Involvement – Recommendations 1 – 4

The first set of recommendations focused on the role and contributions of students to the process of the evaluation of teaching. Under each of the recommendations below is an update on work to date.

1. **Evaluation of teaching should include student feedback.**

   Complete

   This recommendation reaffirmed the important role that student feedback plays in the evaluation of teaching. End-of-course student surveys are one source of data for the process of evaluating teaching, among others (see recommendations 10 and 15 for further information about evaluation of teaching processes and policies).

2. **The name of the process by which student feedback is gathered should be changed from 'Student Evaluation of Teaching' to 'Student Experience of Instruction'.**

   Complete

   Communications about the end-of-course student surveys all now use "Student Experience of Instruction" for the name of the process. The new website with information about the process (seoi.ubc.ca) replaces the previous website (teacheval.ubc.ca), which used the old terminology.

3. **Questions asked of students should focus on elements of instruction based on their experience with instructor(s) in specific contexts and relationships.**

   Complete

   The wording changes to UMIs in SEI surveys on both campuses are a result of this recommendation. Throughout the process of piloting and reframing the questions, students reflected on their perceptions of what the questions were asking and how they might be interpreted in different course contexts. They also made suggestions for improving the questions to ensure they capture various student experiences in courses.

4. **Student leadership on both campuses should be actively engaged in raising the profile of student feedback on instruction.**

   Ongoing

   Students have an important voice and perspective in work to improve the process of gathering student feedback on instruction and how it is used to evaluate and improve teaching at the university. Students have been invited and have participated in this initiative, including participation as members of the Steering and Senate committees, as well as in the work to refine the questions, as outlined above. The Implementation Committee also consulted with student groups and developed information for students about how results from the surveys are used at the university and advice for providing effective, constructive

feedback. Partnering with students on this work was very helpful and productive, ensuring the information will be useful to students. This included the development of a video resource with the UBCV Provost's office and AMS leadership featured on the website; the AMS also ran a campaign in the 2021-2022 academic year to encourage constructive feedback on the SEI surveys. It is helpful to continue to have student involvement in any further creation of resources aimed at a student audience, as well as discussions and activities to support significant student response rates to the surveys

University Module Items – Recommendations 5-9

5. **UMI-6 (Overall the instructor was an effective teacher) should be retained in the core question set, but modified.**
6. **Minor changes in the wording of other UMI questions are suggested to better reflect the focus on each student's experience of instruction.**
7. **UMI-4 (Overall, evaluation of student learning was fair) should be removed from the common set**
8. **A new UMI item, pertaining to the usefulness of feedback, should be trialed.**
9. **There should be a common set of UMI questions asked across both campuses**

Complete

As discussed above, a set of proposed UMIs was developed based on the recommendations from the SEOT working group, and the wording of these was refined after pilot testing. The resultant revised UMIs were implemented into all SEI surveys, using the same questions across both campuses. This reflected a change on the Okanagan campus from 19 questions to 6 and will support future alignment of analyses of data from the surveys across the institution. The previous and updated questions are outlined in Appendix 3.

PAIR will continue to conduct ongoing testing of the functioning of the questions, as well as for bias based on faculty demographic data from the UBC Employment Equity Survey and from a student demographic data project currently underway.

Data and Reporting – Recommendations 10-12

10. **Units should be supported to adopt a scholarly and integrative approach to evaluation of teaching.**

In progress

Members of the SEI Implementation Committee, along with others, completed a discussion paper on an Integrative Approach to Evaluation of Teaching in October of 2021 (see Appendix 5). This paper was created to contribute to the process of developing broader

Senate policies on the evaluation of teaching writ large, through a working group made up of members from both UBCO and UBCV. The paper provides a brief overview of integrative approaches to the evaluation of teaching in other institutions, a summary of some of the teaching evaluation practices at UBC, and a set of recommendations.

Support for units for a scholarly and integrative approach to evaluation of teaching will be further considered and implemented by a new cross-campus working group to develop a draft cross campus policy on Integrative Evaluation of Teaching (see further details under recommendation 15).

11. **Reporting of quantitative data should include an appropriate measure of centrality, distributions, response rates and sample sizes, explained in a way that is accessible to all stakeholders, regardless of quantitative expertise.**

Complete

Individual instructor reports of results have included the interpolated median (instead of the mean), the dispersion index (instead of the standard deviation), and the percent favorable (percentage of respondents who chose Agree or Strongly Agree on each question) since 2018 Winter Term 1.[3] These reports also include the response rate as well as a table with the recommended response rates according to the number of students in the course, based on research by Zumrawi, Bates, and Schroeder (2014).[4]

The interpolated median, dispersion index, and percent favorable are explained on the new Student Experience of Instruction website, under “Metrics.” In addition, workshops explaining these metrics have been held several times at CTLT Institutes over the past few years. PAIR will continue to hold such workshops from time to time.

Finally, a set of videos explaining these metrics and how to interpret them is in the process of being created, and these will be posted on the SEI website, under "Metrics.”

Faculty preparing dossiers for reappointment, tenure and promotion, as well as heads or directors, can request conversion of past results using previous metrics into the new metrics. In addition, unit heads, program directors, and dean’s offices can request aggregate reports.

---

[3] Individual reports included both the previous and new metrics beginning in 2018 Winter Term 1, and only the new metrics beginning in 2020 Winter Term 1.

[4] Zumrawi, A.A., Bates, S.P. & Schroeder, M. (2014). What response rates are needed to make reliable inferences from student evaluations of teaching? *Educational Research and Evaluation*, *20*(7-8), 557-563. DOI: 10.1080/13803611.2014.997915

Please see information about how to request aggregate data reports on the Student Experience of Instruction website.

12. **UBC should prioritize work to extract information from text/open comments submitted as part of the feedback process.**

In progress

In addition to the quantitative information from the Likert-style questions on student surveys, text comments from students may provide more in-depth information about students' experiences in courses. It is important to recognize that the comments sometimes include harmful and abusive language, including racist, sexist, ableist and other discriminatory statements.

Recommendation 12 from the SEoT working group's May 2020 report suggested that a pilot process be undertaken to "investigate the potential of automated approaches to extract useful information from large volumes of text submissions," for formative purposes, so that instructors may more easily understand patterns in the comments. In time, this may also contribute to ways to address harmful comments on the surveys.

The Implementation Committee has reviewed a few such systems, and a summary is included in Appendix 6, explaining investigations undertaken so far and suggestions for possible next steps. The committee reviewed two UBC-developed systems (one from Computer Science and one from Arts ISIT), and two systems from Explorance, the vendor that provides the software system UBC uses for SEI surveys and reporting, Blue. Each has benefits and drawbacks, and none are ready for broad implementation at this time.

Next steps suggested by the committee are pilot testing of one or more systems, as well as further investigation of other emerging tools and platforms. Both of these would require commitments of time and possibly funding to pursue.

The Implementation Committee did not find a tool that could be easily implemented at UBC for locating and removing harmful comments in surveys, though it could be possible to use dictionary-based or machine-learning models to do so, combined with manual removal of harmful comments before reports are provided to faculty. Further investigation is warranted, and commitments of time and resources would be needed before such options could be widely implemented at the institution.

Dealing with Bias – Recommendations 13-14

13. **UBC needs additional and regularized analysis of our own data to answer questions related to potential bias, starting with instructor ethnicity, as it is frequently highlighted as a potential source of bias in the literature on student evaluation of teaching.**

Ongoing

The Implementation Committee has worked with the EIO and PAIR on analyses of SEI data for bias. Before 2022, only analyses on binary sex data for faculty and students had been done using administrative data (see Appendix 3 of the May 2020 SEoT working group report to Senates); this is because there was not enough other demographic data available to yield valid results if analyzed for bias.

A new Employment Equity Survey (EES) has been rolled out for newly-hired UBC employees, and was launched to existing employees starting in September 2021. The questions better address and reflect how the members of the UBC community self-identify. The Implementation Committee was planning to do analyses for bias with data from the new EES and the new UMI questions, but unfortunately, there was not a high enough participation rate in the EES, and we were not able to ascertain if the missing data was missing at random. We were therefore not able to test how the new UMI questions function across other variables of interest e.g., gender identity, ethnicity, disability, and more. We recommend that further IRT and DIF analysis be carried out on the new UMI questions as well as continuing to monitor the Employment Equity Survey response rate and examine the randomness of missing data.

14. **The work of collecting, integrating, interpreting and using feedback on teaching should mitigate against bias, but should not presume the complete removal of bias.**

    Ongoing

    As noted in response to Recommendation 13, regular analyses of SEI data for bias should continue to be conducted, and we recommend below that the Provost's Offices on both campuses, along with Senate Committees, hold the responsibility to ensure this happens. It will then be possible to recommend actions to be taken to mitigate bias, if found, even if complete elimination may not be possible.

Broader Issues – Recommendation 15 – 16

15. **The Vancouver Senate should review the policy on Student Evaluations of Teaching and consider a broader policy on the evaluation of teaching writ large. The Okanagan Senate should develop a similar policy for the Okanagan campus.**

    In progress

As noted above, over the Summer and Fall of 2021 the Implementation Committee wrote a discussion paper with recommendations for a broader, integrative approach to evaluation of teaching, Appendix 5, that has fed into work to develop a policy on evaluation of teaching. Since that time a dual-campus working group and a dual-campus review group have been formed, with faculty co-chairs from both UBCO and UBCV, to work on this recommendation. Initial work from these groups has focused on identifying what the main components of the policy should be:

1. A clear definition of what we are evaluating (e.g., good teaching, quality teaching, effective teaching, teaching excellence) with careful attention to the language used in this definition

2. The identification of principles (or values, dimensions, competencies) that form the foundation of good/effective/excellent teaching at UBC

3. Elements of a new policy such as clearly-stated practices of good evaluation along with accountability processes

4. High level framework to guide implementation of the new policy

Broad consultation is taking place over the Summer and Fall of 2022, and a summary of the feedback provided during the consultation will be taken to the two senate committees in the Fall of 2022. The working group will then develop a draft policy from September to December 2022.

16. **Senate should commit to support the ongoing work of implementing policies related to the evaluation of teaching.**

    In Senate purview

    This recommendation is focused on the need to ensure there is support for broad implementation of policies developed through the above recommendation, and thus this work will need to happen alongside the development of the policies.

Additional areas of work

The SEI Implementation committee also completed or is in the process of completing the following:

- A new website, seoi.ubc.ca, that includes, among other things, information about the changes to the UMI questions, the metrics used in reporting quantitative data, advice for faculty and students, and various reports related to Student Experience of Instruction at

UBC. This website is meant to be a resource for people at both UBC Vancouver and UBC Okanagan, and it will be maintained on an ongoing basis by PAIR.

- Suggestions for faculty members on ways they could report and reflect on their SEI results in dossiers for reappointment, tenure and promotion (these will be posted on the seoi.ubc.ca website in Fall 2022).

- Revisions to the *SAC Guide to Promotion and Tenure*, to reflect a broader approach that addresses all UMI questions and the three metrics for each. The committee will be working with Faculty Relations and the Senior Appointments Committee on these revisions in Fall 2022.
-
- Consultations and presentations with various parts of the UBC community, including open forums in both March and September 2021, as well as several workshops through the Centre for Teaching, Learning, and Technology (see Appendix 2).

Summary of ongoing work

As noted above, the following work is continuing in 2022 and beyond.

- Next steps for investigating and testing automated systems for analyzing text comments for formative purposes

- Dual-campus working group, working with committees in both Senates, to develop Senate policies for evaluation of teaching

- Regular analyses of SEI data done by PAIR, including for bias

In addition, PAIR is working on an online, interactive reporting system that unit heads and dean's offices can use to generate reports of SEI data for their units. The initial release of this system is expected for June 2023. During the initial rollout, a few UBC-wide reports will be made available to heads and administrators. More reports will be developed over time to support other reporting needs. In time, this may be available to individual faculty as well.

Recommendation: Ongoing Governance

With the completion of this report, the work of the SEI Implementation and Steering Committees has largely come to an end. That said, there continues to be a need for ongoing governance of SEI practices at the institution beyond the end of this project that was focused on implementing the SEI recommendations. For example, it would be helpful to clarify responsibility for activities such as: ensuring regular data analyses occur, reviewing the results, and recommending

revisions to questions or processes as needed; providing advice on further supports that may be helpful for faculty, students, or academic leaders; continuing to investigate language processing options for text comments; and advising on the development of interactive reporting dashboards.

Since the UMIs are now the same across both campuses, and the work done on SEI over the past few years has been undertaken collaboratively by people from UBCO and UBCV, we recommend that governance of SEI activities continue to be shared across both campuses. After discussion with the SEI Steering Committee, we recommend that responsibility lie with the Provosts' offices at UBCO and UBCV, with regular connections to the Senate Learning and Research Committee (UBCO) and the Senate Teaching and Learning Committee (UBCV) for updates and feedback.

Appendices

Appendix 1 – Recommendations from May 2020 Senate report

Appendix 2 – Steering & Implementation Committees Membership and Consultations

Appendix 3 – Comparison of previous UMIs and new UMIs for each campus

Appendix 4 – Data analyses of SEI results completed

Appendix 5 – Discussion paper on an integrative approach to evaluation of teaching

Appendix 6 – Report on investigation of options for automated text analysis

Appendix 1 – Recommendations from May 2020 Senate report

Student Involvement

1. Evaluation of teaching should include student feedback.

Students have a unique and valuable perspective from which to provide feedback on teaching at UBC. Student feedback on teaching is one of several sources of data that should be used for making personnel decisions and for the improvement of teaching.

2. The name of the process by which student feedback is gathered should be changed from 'Student Evaluation of Teaching' to 'Student Experience of Instruction'.

Evaluation of teaching is a complex process, whether for formative or summative purposes. To do it effectively requires input from multiple perspectives and sources (students, peers, self) integrated across time. As noted in (1) above, students have an important perspective that should be part of that. However, students should be asked to focus on their experience, rather than to 'evaluate' teaching writ large.

3. Questions asked of students should focus on elements of instruction based on their experience with instructor(s) in specific contexts and relationships.

In line with a recent statement from the American Sociological Association (Article, Sept 2019) questions for students should focus on their experiences and be framed as an opportunity for students to provide feedback, rather than positioning the request as a formal and global evaluation of the teacher.

4. Student leadership on both campuses should be actively engaged in raising the profile of student feedback on instruction.

Gathering and considering feedback on teaching and learning from students is a responsibility shared between faculty and students. Student leadership should play an active and visible role in raising awareness of the purposes for, and ways in which, this feedback can improve instruction. Student leadership should also be part of efforts to raise awareness of comments that are not appropriate and/or counter-productive in the context of an anonymous survey.

UMI Questions

5. UMI-6 (*Overall the instructor was an effective teacher*) should be retained in the core question set, but modified.

The working group had extensive discussions about the inclusion or deletion of this item. Analysis of UBC data indicates that UMI-6 scores are able to be predicted to a high degree of confidence based on a weighted linear combination of other UMI questions (except UMI-4). However, in its current form, UMI-6 asks students to directly evaluate the 'overall effectiveness of the teacher'. As we have argued above, students are not in a position to be able to make sweeping, all-inclusive judgments about the effectiveness of instruction. On balance, the working group recommends retaining UMI-6, but rewording it as '*Overall, this instructor was effective in helping me learn'*. This centers the question on the individual experience of the student.

6. Minor changes in wording of other UMI questions are suggested to better reflect the focus on each student's experience of instruction.

The instructor made it clear what students were expected to learn, to be changed to
The instructor made it clear what I was expected to learn
The instructor helped inspire interest in learning the subject matter, to be changed to
The instructor engaged me in the subject matter
The instructor communicated the subject matter effectively to be changed to
I think that the instructor communicated the subject matter effectively.
The instructor showed concern for student learning to be changed to
I think that the instructor showed concern for student learning
The latter two questions are phrased so as to balance first person perceptions with overall cohort experience and classroom climate.

7. UMI-4 (*Overall, evaluation of student learning was fair*) should be removed from the common set

UMI-4 is something of an outlier in the current UMI set used in Vancouver campus surveys. It is consistently answered by fewer students. It is also problematic because the concept of 'fairness' is highly ambiguous. Student consultations have indicated they are often unsure how to interpret what 'fairness' means.

8. A new UMI item, pertaining to the usefulness of feedback, should be trailed.

Whilst the working group recommends removal of the previous UMI-4 item, on fairness of assessment (see recommendation 4), there was a strong sense that, given the importance of timely and effective feedback in the learning process, this should be reflected in the core UMI questions.

We recommend a question worded as follows: *"I have received feedback that supported my learning"*. However, this question should be piloted in a limited set of courses in 2020/21 to ensure that we understand how responses might be influenced by variables such as class size, etc. It is certainly the case that the opportunity to provide feedback, and indeed the nature of that feedback (e.g., written and / or numerical), will look very different in a seminar class of 20 compared to a large introductory lecture of 200. We should collect data from a pilot to better set.

The results of the pilot could be included in the 2020/21 Report to Senates and a decision taken on how to proceed.

9. There should be a common set of UMI questions asked across both campuses

There should be a commonly-used core set of five or six questions across both campuses. Modular approaches to constructing feedback surveys may be appropriate (university-wide items plus Faculty, Department and course-specific items). However, units should be mindful that most students complete several surveys per semester, potentially causing 'feedback fatigue' and reducing rates of participation. Therefore, units should be mindful of the overall length of feedback surveys students are being asked to complete. Units should also explore other ways to gather specific feedback as the course progresses.

Data and Reporting

10. Units should be supported to adopt a scholarly and integrative approach to evaluation of teaching.

Because teaching is complex and contextually dependent, departments and units should be supported to adopt an integrative and scholarly approach to evaluation that synthesizes multiple data sources (e.g., students, peers, historical patterns, and self-reflection documentation) for a holistic picture, without over-reliance on any single data source. This approach will necessarily look different in different units but should include both in-kind support from units such as CTLT/CTL and funding for department leaders to accomplish the work proposed. When used for personnel decisions, the unit's approach, strategy, and norms can then be communicated to all levels of review, along with the file. The VPAs on both campuses should work with the Senior Appointments Committee (SAC) to identify and disseminate anonymous examples of effective ways to integrate, synthesize and reconcile multiple perspectives on teaching effectiveness.

11. Reporting of quantitative data should include an appropriate measure of centrality, distributions, response rates and sample sizes, explained in a way that is accessible to all stakeholders, regardless of quantitative expertise.

The interpolated median should be used as the measure of centrality, with the dispersion index as a measure of spread. Reports should include distributions of responses, response rates and sample sizes, clearly flagging where response rates do not meet minimum requirements for validity and accuracy. Visualizations of comparative (anonymous) data should be developed, along with an on-going program of consultation and dissemination to different groups (faculty, staff and administrators).

12. UBC should prioritize work to extract information from text/open comments submitted as part of the feedback process.

Many faculty members report the free-text student comments as sources of rich data to support reflection and enhancement of their course and teaching. It is recommended that a pilot investigation be undertaken, with one or more Faculties, to investigate the potential of automated approaches to extract useful information from large volumes of text submissions. The pilot should engage with appropriate research expertise in Faculties in these areas, and aim initially for formative purposes. There is an opportunity for UBC to take a lead among institutions in providing balance and insight when combining quantitative and qualitative data. Failing to do this continues to privilege quantitative over qualitative data about teaching.

Dealing with Bias

13. UBC needs additional and regularized analysis of our own data to answer questions related to potential bias, starting with instructor ethnicity, as it is frequently highlighted as a potential source of bias in the literature on student evaluation of teaching.

An analysis of UBC-V data with respect to instructor and student gender over the last decade reveals no systematic differences in aggregate data of ratings received by female vs. male instructors. Variables tested for (including instructor and student gender) indicate aggregate differences at the level of approximately +/- 0.1 on a 5-point scale, in other words, very small effects. Course-specific effects (e.g., subject discipline, course level) demonstrate larger effects (typically +/- 0.3 on the same scale). An analysis of UBC-O data across 2015-16 and 2018 academic year revealed mixed results, as are detailed in Appendix 4.

For both campuses, it is important to note that this is an analysis of aggregate data and, as such, will mask variation on an individual level. The lived experience of individual instructors may be quite different from this aggregate view. However, holistic evaluations of a person's teaching (see: Recommendation 15) can be used to contextualize individual instructors' experience. We cannot stress enough the importance of a holistic evaluation that allows individual lived experiences to be heard, particularly if their lived experience runs counter to the aggregate data.

Given that studies have presented evidence of bias on the basis of instructor ethnicity, it would seem both appropriate and timely that the same analysis be brought to bear in checking the UBC data for bias. This work comes with privacy and ethical implications. We recommend developing a process that would allow instructor ethnicity data to be accessed confidentially for regular investigation of bias. We have not been able to address this analysis during the timescale of this working group and thus recommend a follow-on activity to investigate this, reporting back to Senates during the 2020-2021 academic year. The follow-on report would also be in a position to recommend regularized analysis and mitigation strategies to address any systematic biases found, particularly related to gender and/or ethnicity.

14. The work of collecting, integrating, interpreting and using feedback on teaching should mitigate against bias, but should not presume the complete removal of bias.

As with most other forms of surveys, student feedback on instruction cannot be completely free from bias. Bias can be explicitly discriminatory and perpetuating of stereotypes. But bias can also be implicit, where respondents are not consciously aware of how their attitudes influence their responses. Implicit biases have been shown to occur in many domains and the general approach at UBC (e.g., on hiring committees) has been one of mitigation through education and awareness raising.

This recommendation is supported by an analysis of the voluminous literature on the topic of student evaluations of teaching, and interrogation of the UBC dataset at multiple points in the last 10 years. Research literature reports studies on a wide variety of instruments and processes, with considerable variation in the scope of data collected. Individual studies are often reported in the mainstream academic press, sometimes with extrapolation beyond the context and the effects found in the initial study. Studies investigating a variety of instructor effects (e.g., age, gender, ethnicity) vary in whether they show bias, no bias or bias toward (rather than against) female instructors. In the subset of published studies where biases are found, and enough detail is provided to be able to discern the effect size, those effect sizes on aggregate are small.

Broader Issues

15. The Vancouver Senate should review the policy on Student Evaluations of Teaching and consider a broader policy on the evaluation of teaching writ large. The Okanagan Senate should develop a similar policy for the Okanagan campus.

Student feedback, both quantitative and qualitative, should be integrated with other forms of data to estimate the effectiveness of a faculty member's teaching. The current policy (2007) says little about how student feedback should be integrated with other forms of data before making judgments about the effectiveness of teaching. Therefore, it is appropriate to revisit the UBC-V Senate Policy on Student Evaluation of Teaching and consider adding or replacing it with a policy that sets forth a broader and teaching. Similar processes should be applied and governed by either a joint Senate policy, or aligned policies for each campus.

16. Senate should commit to support the ongoing work of implementing policies related to the evaluation of teaching.

Career advancement decisions are made on the recommendation of Departmental, Faculty and a system-wide Senior Appointments Committee, each of whom is tasked to evaluate teaching effectiveness as a component of every case. It is imperative that UBC commit to providing the necessary resources and training, including administrative and technological support, to implement Senate policies on evaluating teaching (see Recommendation 15). Faculty members must be given the tools, resources, and support to effectively present a scholarly case for their teaching effectiveness. Likewise,

evaluators at all levels must be adept at appropriately interpreting and contextualizing the kinds of data offered across diverse disciplinary and teaching contexts, with due consideration to multiple sources of data and the limitations of each.

Appendix 2 – Steering & Implementation Committees membership and consultations

The Steering committee and Implementation Group began work in the Fall 2020, and smaller groups also worked on specific items.

SEI Steering Committee, 2020-2022

Name	Title
Simon Bates	Vice-Provost and Associate Vice-President, Teaching and Learning, *pro tem*, UBCV (Co-chair)
Moura Quayle	Vice-Provost and Associate Vice-President Academic Affairs, UBCV, (Co-chair)
Breeonne Baxter (Dec 2021-May 2022)	Communications Manager, VPA Communications, UBCV
Eshana Bangu (May 2021- May 2022)	Vice President Academic and University Affairs, AMS, UBCV
Stefania Burk	Associate Dean Academic, Faculty of Arts, Dean of Arts *pro tem* April 4-June 30, 2022, UBCV
Sage Cannon	Students Union Okanagan - Faculty of Creative & Critical Studies Representative, UBCO
Julia Mitchell	Director, Communications & Marketing, Office of the Provost & Vice-President Academic, UBCV
Karen Rangoonaden (Until Aug 2021)	Chair, Senate Learning and Research Committee, UBCO
Rehan Sadiq	Provost and Vice-President Academic *pro tem* as of February 1, 2022, and Professor and Executive Associate Dean, School of Engineering, UBCO
Dana Turdy (Joined June 2022)	Vice President Academic and University Affairs, AMS, UBCV
Naznin Virji-Babul	Assistant Professor, Physical Therapy Senior Advisor to the Provost on Women and Gender-Diverse Faculty, UBCV
Sally Willis-Stewart (Joined Aug 2021)	Chair, Senate Learning and Research Committee, UBCO
Georgia Yee (Sept 2020-April 2021)	Vice-President Academic and University Affairs. AMS, UBCV

SEI Implementation Committee, 2020-2022

Name	Title
Christina Hendricks	Academic Director, CTLT, Professor of Teaching, Philosophy, UBCV (Chair)
Vanessa Auld	Professor and Head, Department of Zoology, UBCV

Breeonne Baxter	Communications Manager, VPA Communications, UBCV
Brendan D'Souza	Lecturer, Department of Biology, UBCO
Tanya Forneris	Interim Academic Lead, CTL (2020-2021), Associate Professor of Teaching, School of Health & Exercise Sciences, UBCO
Andrea Han (Joined Sept 2021)	Associate Director, Curriculum and Course Services, CTLT, UBCV
Mark Lam	Lecturer, Department of Psychology, UBCV
Stephanie McKeown	Chief Institutional Research Officer, PAIR
Marianne Schroeder (Sept 2020-Feb 2021)	Sr. Associate Director, Teaching and Learning Technologies, CTLT, UBCV
Alison Wong (Joined Sept 2021)	Project Manager, PAIR
Abdel-Azim Zumrawi (Joined Feb 2021)	Statistician, PAIR

Advisory group on changes to UMI questions (2020-2021)

Name	Title
Christina Hendricks	Academic Director, CTLT, Professor of Teaching, Philosophy, UBCV
Stephanie McKeown	Chief Institutional Research Officer (PAIR)
Catherine Rawn	Professor of Teaching, Psychology, UBCV
Bruno Zumbo	Professor, Canada Research Chair in Psychometrics and Measurement, Tier 1; & Paragon UBC Professor of Psychometrics and Measurement Educational and Counselling Psychology, and Special Education, UBCV
Abdel-Azim Zumrawi	Statistician, CTLT, UBCV

Integrative approach to evaluation of teaching discussion paper working group

Name	Title
Tanya Forneris	Interim Academic Lead, CTL (2020-2021), Associate Professor of Teaching, School of Health & Exercise Sciences, UBCO (Chair)
Brendan D'Souza	Lecturer, Department of Biology, UBCO
Christina Hendricks	Academic Director, CTLT, Professor of Teaching, Philosophy, UBCV
Sajni Lacey	Learning & Curriculum Support Librarian, Library, UBCO
Jaclyn Stewart	Associate Dean Academic, Faculty of Science UBCV as of January 2022, Deputy Academic Director, CTLT (2019-2021), Associate Professor of Teaching, Chemistry, UBCV

Project Management: Debbie Hart, Senior Manager, Strategic Projects, VP Academic Office, UBCV

Project Consultation:

Starting in the Fall of 2020 the Implementation Committee consulted with several groups, which informed and provided feedback on the work of implementing the recommendations.

In addition to the work detailed above to test the new UMI, discussions have been held with and feedback collected from:

- UBC Vancouver:
  - Senate Teaching & Learning Committee
  - Associate Deans Academic, Students, and Faculty
  - Heads & Directors (at Provost's Heads & Directors meeting)
  - UBCV Student Senate Caucus
- UBC Okanagan:
  - Senate Learning & Research Committee
  - Deans Council
  - Student Academic Success Committee
- Across both campuses:
  - Senior Appointments Committee
  - Open forums: March 10 and September 28, 2021
  - Online workshops on changes to SEI questions and metrics (at CTLT Institutes, Aug 2021 and May 2022)

Appendix 3 - Comparison of previous UMIs and new UMIs for each campus

New SEI questions for both campuses from September 2021

1. Throughout the term, the instructor explained course requirements so it was clear to me what I was expected to learn.
2. The instructor conducted this course in such a way that I was motivated to learn.
3. The instructor presented the course material in a way that I could understand.
4. Considering the type of class (e.g., large lecture, seminar, studio), the instructor provided useful feedback that helped me understand how my learning progressed during this course.
5. The instructor showed genuine interest in supporting my learning throughout this course.
6. Overall, I learned a great deal from this instructor.

Response options for all questions above: *strongly agree, agree, neutral, disagree, and strongly disagree.*

A set of open-ended questions are included on surveys on both campuses as well as of Fall 2021:

7. Do you have any suggestions for what the instructor could have done differently to further support your learning?
8. Please identify what you consider to be the strengths of this course.
9. Please provide suggestions on how this course might be improved.

SEOT questions pre-Sept 2021

Okanagan Campus	Vancouver Campus
Instructor Questions The instructor set high expectations for students. The instructor showed enthusiasm for the subject matter. The instructor encouraged student participation in class. The instructor fostered my interest in the subject matter. The instructor effectively communicated the course content. The instructor responded effectively to students' questions.	The instructor made it clear what students were expected to learn. The instructor helped inspire interest in learning the subject matter. The instructor communicated the subject matter effectively. Overall, evaluation of student learning (through exams, essays, presentations, etc.) was fair. The instructor showed concern for student learning.

The instructor provided effective feedback. Given the size of the class, assignments and tests were returned within a reasonable time. The instructor was available to students outside class. The instructor used class time effectively. The instructor demonstrated a broad knowledge of the subject. Students were treated respectfully. Where appropriate, the instructor integrated research into the course material. The evaluation procedures were fair. I would rate this instructor as very good. **Course questions** Textbook and/or assigned readings contributed strongly to this course. I found the course content challenging. I consider this course an important part of my academic experience. I would rate this course as very good.	Overall, the instructor was an effective teacher

Appendix 4 - Data analyses of SEI results

EXECUTIVE SUMMARY

A set of six new/reworded University Module Items (UMI) questions were implemented in the Student Experience of Instruction (SEI) surveys across both UBC campuses starting in the Fall of 2021.

Sample data from the 2021 Winter Term 1 were used to evaluate the new questions. To determine how well the new items functioned across individuals and respondent groups, we conducted a quantitative analysis of the questions using Item Response Theory (IRT), Differential Item Functioning (DIF) and Generalized Linear Mixed Models (GLMM), using the software programs SAS and R. Results from the IRT models showed improvement in the items' contribution to the overall survey information compared with a sample drawn at random from pre-2021 SEI (2020 Winter Term 2) survey. DIF was not detected, or was negligible for grouping by campus, year level or class meeting time. Moderate uniform DIF was detected in UMI question 1 for class size (favoring larger class sizes) and for UMI questions 3 and 6 for instructor and student gender, respectively (female instructors received slightly more positive responses).

GLMM results showed differences in some UMI questions for some course attributes, instructor and student demographics, however, the effect sizes were small.

1.0 INTRODUCTION

In February 2019, a Student Evaluation of Teaching (SEoT) working group formed with membership across both UBC Okanagan and UBC Vancouver campuses. That working group produced a report to both Senates in May of 2020 with recommendations for SEI surveys and processes. To address the recommendation by the working group to revise the University questions, the SEI Implementation Committee developed an eight-step project plan (see Figure 1). This plan included a mixed-method approach that collected qualitative feedback from student and faculty participants through focus groups and interviews, revised the questions based on this feedback, then conducted pilot-tests of the new questions using an online survey, and finally conducted a quantitative analysis of the results to see how well the revised items functioned.

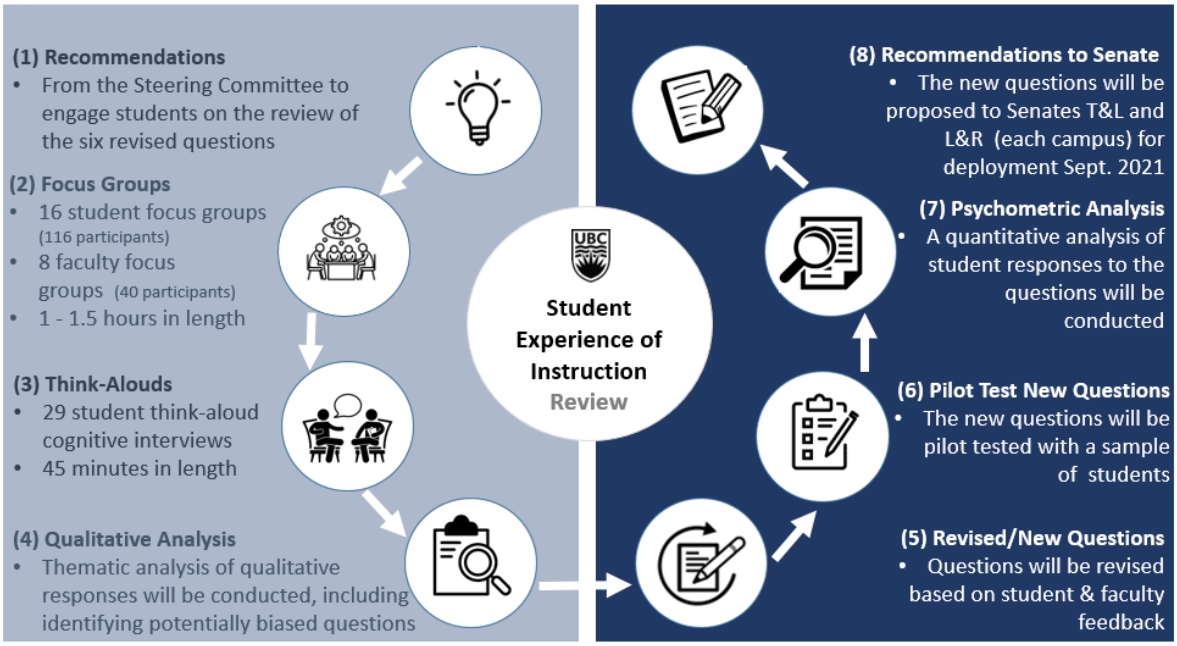


Figure 1. Eight-Step Plan used to Evaluate the Proposed SEI Questions in 2021

Based on the 8-step procedure for evaluating, revising and testing UMI questions, the following final set of six core UMI questions were recommended for implementation at both UBC Vancouver and UBC Okanagan, starting in 2021 Winter Term 1:

1. Throughout the term, the instructor explained course requirements so it was clear to me what I was expected to learn.
2. The instructor conducted this course in such a way that I was motivated to learn.
3. The instructor presented the course material in a way that I could understand.
4. Considering the type of class (e.g., large lecture, seminar, studio), the instructor provided useful feedback that helped me understand how my learning progressed during this course.
5. The instructor showed genuine interest in supporting my learning throughout this course.
6. Overall, I learned a great deal from this instructor.

Five of these questions (1, 2, 3, 5 and 6) were rewordings, however, UMI 4 is a new question based on a recommendation from the 2020 report to Senates from the SEoT working group.

Following the implementation of the new UMI questions, university-wide Student Experience of Instruction data from the 2021 Winter Term 1 was used to further test and evaluate the UMI questions. This report presents a summary of the data used, analysis, methods and findings.

2.0 DATA

SEI data from 2021 Winter Term 1 (2021W1) from both UBC campuses were used in this analysis. 100 course/section surveys were randomly selected from each of five fields of study (Sciences, Humanities, Social Sciences, Engineering and Health Sciences). Stratified sampling by field of study is key to ensure balanced representation across fields of study. Academic units/programs within each field of study are given in the Appendix to this report. The SEI data were screened and merged with enrollment data to obtain some variables of interest such as class meeting time and delivery mode. However, a significant number of course sections were missing "delivery mode" and this variable was removed from further analysis.

We attempted to use the Employment Equity Survey data to obtain other variables of interest, such a gender identity, ethnicity, disability, and more. However, about half of the instructors who taught in 2021 W1 were missing employment equity data. Furthermore, for those instructors with such data, available gender data was not different than what is in the SEI data (binary), with sparse data on other gender categories. Because we could not ascertain the randomness of missing equity data, which could potentially affect how different groups were represented in the dataset, employment equity data were excluded from further consideration.

The SEI sample dataset comprised 11,032 student responses to the six UMI questions. Tables 1.a, 1.b, 1.c and 1.d show the distribution of the dataset, used in the final analysis, by course, instructor and student attributes.

Table 1.a: Distribution the 2021W1 SEI Responses by Field of Study & Year Level

Field of Study	Number of responses
Engineering	1,892
Health Sciences	1,520
Humanities	1,784
Sciences	3,090
Social Sciences	2,746
Total	11,032

Year Level	Number of responses
1st	3,181
2nd	3,086
3rd	2,637
4th	969
5th	1,159

Table 1.b: Distribution the 2021W1 SEI Responses by Student Demographics

Campus	Number of responses

UBCO	2,134
UBCV	8,898

Student Gender	Number of responses
Female	6,542
Male	4,490

Table 1.c: Distribution of the 2021W1 SEI Responses by Instructor Attributes

Instructor Rank	Number of responses
Assoc. Prof	1,845
Asst. Prof	2,917
Lecturer	1,754
Professor	1,933
Sessional	2,583

Instructor Gender	Number of responses
Female	4,211
Male	6,821

Table 1.d: Distribution the 2021W1 SEI Responses by Course Attributes

Class Meeting Time	Number of responses
Before 11:00 AM	3,635
After 11:00 AM	7,397

Class Size	Number of responses
< 100	4,519
>= 100	6,513
1 - 49	2,427
200+	2,891

3.0 ANALYSIS AND RESULTS

Quantitative data from the SEI 2021 Winter Term 1 surveys were analyzed using Generalized Linear Mixed Models (GLMM), Item Response Theory (IRT) and Differential Item Functioning (DIF).

We used a generalized linear mixed modelling approach to model the cumulative logit of response levels, as a function of the key variables of interest, with Field of Study as a grouping

variable (random effect). This is akin to hierarchical modeling, but with some differences. The estimated model parameters and associated odds ratios were used to test for difference in ratings among groups of interest such as gender.

IRT is an approach used for test development and can be used in a similar fashion for survey item development or refinement. Through IRT, we are able to: 1) predict individual survey responses based on a respondent's attitude or perception, and 2) to establish a relationship between an individual's item response and the set of traits underlying item performance through a function called the "item characteristic curve" (Hambleton et al., 1991). This information can help the survey developer evaluate how well the questions function across different attitudinal levels, and how well the response options work for each question.

There are several assumptions of the data that need to be met before conducting and interpreting this IRT analysis: 1) unidimensionality of the measured trait; 2) local independence of the survey items; 3) monotonicity; and 4) item invariance. Unidimensionality means that all items on the survey are measuring just one underlying construct (e.g., quality of instruction as experienced by students) and that one main factor should explain most of the variance in the survey responses (Hambleton et al., 1991). When items on the survey have local independence, it means that the response to one item is independent of the other questions on the survey, except for the fact that they measure the same underlying construct. Monotonicity occurs when the probability of positively endorsing an item continuously increases as an individual's attitude/perception level increases. Finally, item invariance means that the estimated item parameters do not differ across different groups of respondents, due to misunderstanding or misinterpretation of the questions. These assumptions were met for this analysis and therefore we were able to continue with interpreting the results.

DIF analyses examined whether students responded to the UMI questions differently across groups, such as class size or meeting time, campus, year level, student or instructor gender. In surveys, DIF is conceptualized as occurring when survey respondents who have similar attitudes on a measured trait respond differently due to construct-irrelevant factors such as differential interpretation of terms used in the survey. If an item is flagged as having DIF it suggests that a survey question may indicate a different understanding across respondent groups. When DIF is detected, further review and judgement are required to determine whether refinement of the survey question is needed. We used three different methods (both non-IRT and IRT-based) to determine DIF and to see if the results corresponded across the different methods: 1) Mantel-Haenszel, 2) Regression-based methods (binary and ordinal), and 3) Lord's Chi-square test (IRT-based).

Rather than determining sample size requirements alone, researchers suggest that a combination of sample size and the number of questions on the survey should be considered together to determine if item parameters are estimated accurately in IRT models. Şahin & Anil (2017) concluded that a sample size of 250 with 30 items is viable for a 2-parameter model. Zumbo (1999) suggested that 20 test items can be successfully used to run a DIF analysis and have enough information to be able to match individuals on ability level and form meaningful

groups. We have a large enough sample size in terms of student responses (11,032). Although the number of UMIs on the SEI survey is relatively small (only six UMIs), in a pilot study, McKeown, Zumrawi and Pena (2021) found that a sample of 320 suffices to estimate a 2-parameter IRT model parameters for the six UMI questions. Additionally, for the IRT-based methods, researchers have suggested having at least 30 responses (Linacre, 1994), with valid findings demonstrated using short tests (4 to 39 items) and small sample conditions (100-300 responses) (Paek and Wilson, 2011).

Factor analysis was used to test if all six UMI questions represented a single underlying construct measuring quality of instruction from the student perspective (unidimensional assumption).

3.1 ITEM RESPONSE THEORY AND DIFFERENTIAL ITEM FUNCTIONING

The results of the factor analysis showed that all six UMI items had high factor loadings, i.e., all six UMI questions represent one underlying construct. The Scree and Variance plots in Figure 2 summarize the results of the factor analysis. The elbow in the Scree plot in Figure 2 indicates minimal contributions from subsequent factors. The first factor explained more than 80% of the variation. These findings support the unidimensionality assumption for the IRT analysis.

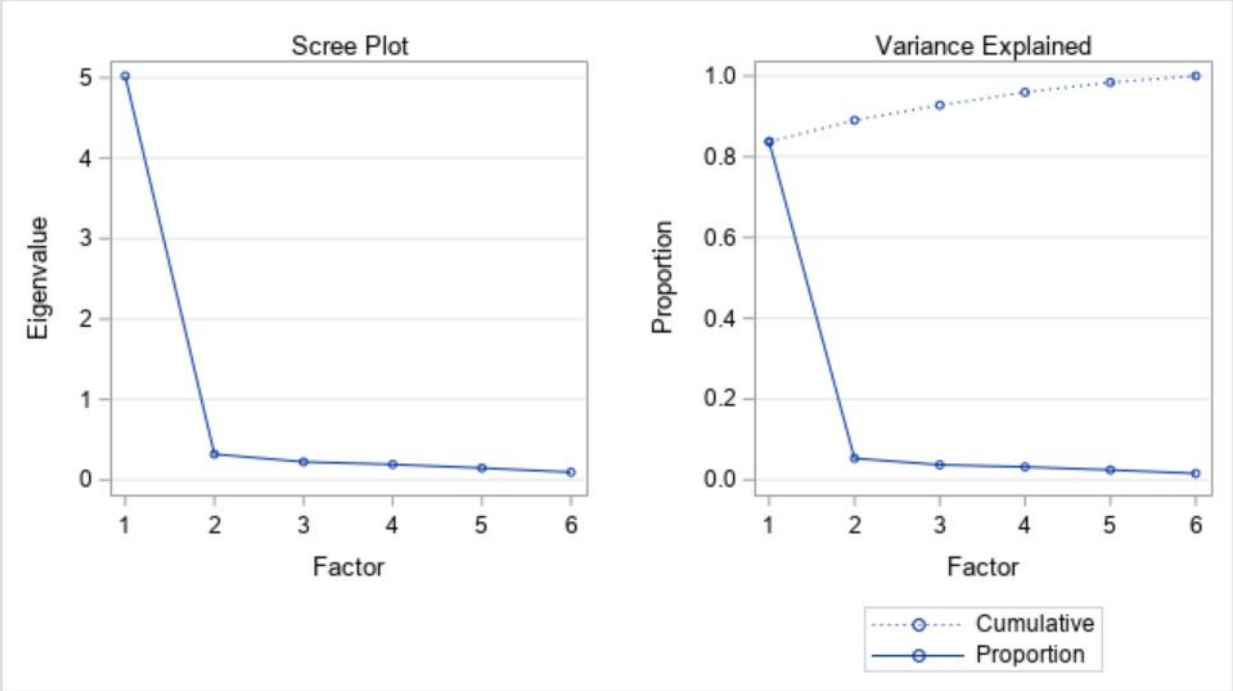


Figure 2. Scree and Variance Plots

3.2 DIFFERENTIAL ITEM FUNCTIONING (DIF)

Using DIF analysis, we examined whether students responded differently across groups, such as class size, campus, year level, or student gender. The results of the DIF analysis will flag an item if it functions differently across participant groups, will indicate the direction of the DIF, and will also indicate if an item has uniform or non-uniform DIF. Uniform DIF occurs when DIF is the same for all attitude levels across the two groups, whereas non-uniform DIF occurs when there is an interaction between attitude levels and group membership.

The R programming environment (package difR) was used to run the Mantel-Haenszel procedure and Lord's chi-square test (Lord, 1980). To interpret the effect size (magnitude) of DIF, we used ΔMH (delta MH), a transformation of the Mantel-Haenszel statistic (M-H), as proposed by Holland and Thayer (1985):

a) none or negligible DIF detected with absolute values of delta MH less than 1;

b) moderate DIF detected with absolute values of delta MH between 1 to 1.5; and

c) large DIF detected with absolute values of delta MH larger than 1.5.

We used SAS statistical software (Proc Logistic and Proc Genmod) to run a logistic regression model and a Generalized Linear Model (GLM) approaches for DIF analysis. In the logistic regression model, DIF is detected if individuals matched on attitude/perception have significantly different probabilities responding to a survey question and therefore will have differing logistic regression curves. We followed a three-model approach for the logistic regression method. The first model used a binary approach for the dependent variable, i.e. UMI survey item, where responses on the Likert scale of 4 "agree" and 5 "strongly agree" were combined and coded together as "favourable." A logistic regression model was fit to the binary data as a function of "attitude/perception," as measured by the overall survey score, in addition to predictor variables (class, student and instructor attributes) other than the grouping variable of interest. The second model included the variables in the first model and a variable representing the reference and focal groups of the variable of interest, such as student gender. Finally, the third model included the variables in the second model plus an interaction term (e.g., attitude/perception*gender).

Model 1: $\boldsymbol{Logit(P)} = \boldsymbol{\beta_0} + \sum_{i=1}^{k-1} \boldsymbol{\beta_i X_i} + \boldsymbol{\beta_k \theta}$

Model 2: $\boldsymbol{Logit(P)} = \boldsymbol{\beta_0} + \sum_{i=1}^{k-1} \boldsymbol{\beta_i X_i} + \boldsymbol{\beta_k \theta} + \boldsymbol{\beta_{k+1} Z}$

Model 3: $\boldsymbol{Logit(P)} = \boldsymbol{\beta_0} + \sum_{i=1}^{k-1} \boldsymbol{\beta_i X_i} + \boldsymbol{\beta_k \theta} + \boldsymbol{\beta_{k+1} Z} + \boldsymbol{\beta_{k+2} \theta Z}$

Where:

Logit(P) is the logit of the probability of respondent's endorsement;

$\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1}$ ... ... $\boldsymbol{\beta_{k+2}}$ are model parameters;

$\boldsymbol{\theta}$ denotes the value of the responder attitude/perception as measured by total score; and

$\boldsymbol{X_1}$, ... $\boldsymbol{X_{k-1}}$ are predictor variables (class, student and instructor attributes) other than the grouping variable of interest.

$\boldsymbol{Z}$ (K^{th} predictor variable) denotes group membership (e.g. gender, class size…etc.)

The generalized linear model method applies a similar three-model approach, except that the dependent variable uses the ordinal response scale values (Likert scale strongly agree "5" – strongly disagree "1") of the UMI survey item and fits a cumulative logit function. For both approaches, a significant difference in fit statistics between models 1 and 2, i.e., a significant $\boldsymbol{\beta_{k+1}}$ would indicate uniform DIF, whereas a significant $\boldsymbol{\beta_{k+2}}$ in model 3 would indicate non-uniform DIF.

The logistic regression and generalized linear model procedures were used to indicate the direction and type of DIF, if and only if the other two methods (Mantel-Haenszel and Lord) detected DIF.

The results of the DIF analysis between different groups of student demographics, course attributes and instructor demographics are summarized in Table 2.

Table 2: Differential Item Functioning (DIF) between different student, instructor and course attributes.

DIF Method	Campus	Student Gender	Class Size < 100 vs > 100	Class Size 1 – 49 vs 200+	Class Meeting Time Before 11 vs After 11	Year Level 1st, 2nd & 3rd vs 4th & 5th	Instructor Gender
Mantel-Haenszel*	Negligible	UMI 6 moderate	UMI 1 moderate	**UMI 1**, 4 (large) UMI 5, 6 moderate	Negligible	Negligible	UMI 3 moderate F
Logistic (Binary)**	None	UMI 6 uniform F	UMI 1 uniform >100	**UMI 1**, 4, 5, 6 uniform >50	None	None	UMI 3 uniform F

GLM (ordinal)**	---	UMI 6 uniform F	UMI 1 uniform >100	All uniform >50	None	**----**	UMI 3 uniform F
Lord's Chi-square Test	None	None	UMI 1	**UMI 1**, 2 & 6	None	None	UMI 3

* MH effect size determined using (Holland and Thayer 1985).
** Logistic & GLM methods used to indicate direction and type of DIF, if moderate or large DIF detected by Lord's & M-H methods.

Results reported in Table 2 indicate that DIF was not detected, or was negligible, for grouping by campus, class meeting time or year level.

Moderate uniform DIF was detected for student gender by the Mantel-Haenszel method (delta MH of 1.05 and p-value < 0.0001), but not by the IRT-based Lord's method. Recall that delta MH values of less than 1.0 indicate no or negligible DIF. Female students were more positive in their responses to this item, but the results were inconclusive.

Across all four methods, UMI question 1 showed large DIF between the smallest and largest class sizes (enrolments of 1-49 compared with classes with 200+ enrolments), with more positive responses given to the largest class size over the smallest (delta MH of 1.73 and p-values of < 0.001 for the four methods). Similarly, UMI question 6 showed moderate uniform DIF between the smallest and largest class sizes, across all four methods (delta MH of 1.2 and p-values of 0.0354, 0.003, < 0.0001 and < 0.0001, for the four methods, respectively). The results for the other UMIs, comparing the smallest and largest class sizes, were different across the test methods and were therefore inconclusive.

There was moderate DIF detected (delta MH of 1.37 and p-values of < 0.0001 for all 4 methods) for UMI 1 comparing class sizes over 100 to those below 100 (again favoring the larger class sizes).

Finally, UMI 3 showed moderate (bordering on negligible) uniform DIF (delta MH of 1.01 and p-values of 0.0004, < 0.0001, <.0001, and 0.0038, for the four methods, respectively) for instructor gender; female instructors received slightly more positive responses on this item.

Graphical representations of the Mantel-Haenszel and Lord's DIF statistics are shown in the Appendix to this report.

3.3 ITEM RESPONSE THEORY

A two-parameter IRT model (graded response model, using Marginal Maximum Likelihood

estimation method) was used to assess item response characteristics, item information and overall information functions, and to evaluate whether similar profiles were found between the new survey items (2021 survey) and the 2020 version of the UMI survey. Two-parameter IRT models estimate the location and discrimination parameters of the survey items along the attitudinal scale of respondents. We used a 2-parameter, MLIRT model to account for variation between fields of study and assess the effect of other variables, including course attributes and instructor demographics within fields of study. The item location parameter provides information on how difficult it is to achieve a 50% probability of a correct response for a specific item given the respondent's level on the underlying attitudinal scale. For example, if a student responds to UMI question 6, "I learned a great deal from this instructor," by answering with the most positive response option available, "strongly agree," this item would be located to the right or higher end on the attitudinal scale. A student who was very positive about their experience of instruction in the course would be more likely to have a 50% probability of endorsing the most positive response options for the UMI questions than a student with a more negative attitude about their experience of instruction in the course.

The item location parameter also provides information on how the different response options (i.e., Likert scale options) function within each item. Although the UMI questions have essentially the same response options, respondents may not use the scale in an equivalent manner across the questions. The item location parameter estimates can provide information to the survey developers about the allocation of appropriate item and response-option weightings. Item location parameter estimates (thresholds) were fairly consistent across response options for the six UMI questions (see Appendix for the all IRT model parameter estimates), which indicates that the 5-point Likert scale options function similarly within each of the six new UMI questions.

Reliability estimates were consistent across approaches; Cronbach's alpha is a measure of scale reliability which indicates internal consistency. For the 2021 survey items, Cronbach's alpha of 0.94 suggests a high survey reliability. Furthermore, an IRT conditional reliability curve is shown in Figure 3.

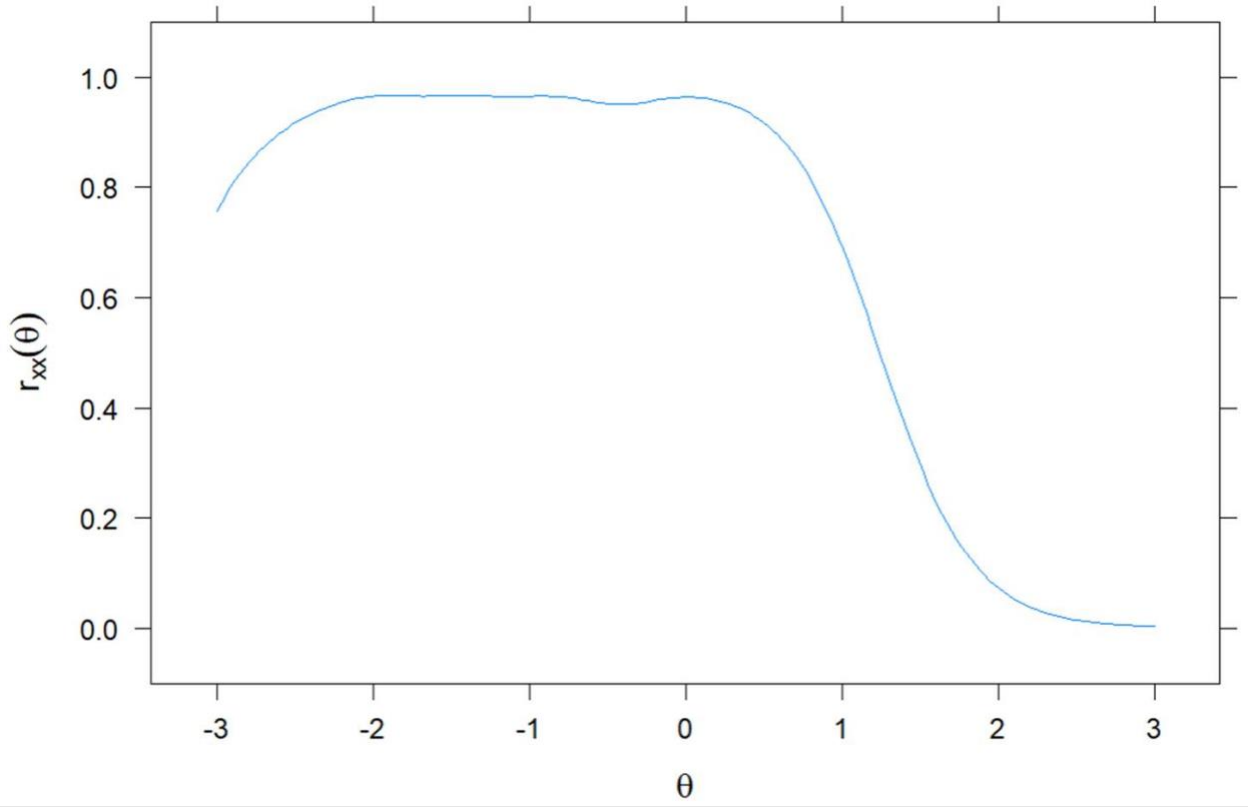


Figure 3. Conditional Reliability Curve

The curve in figure 3 indicates that score estimates are most reliable on a wide range of attitudinal scale; with an overall IRT marginal reliability estimate of 0.84.

The item discrimination parameter indicates the strength of the relationship between an item and the measured construct, i.e., experience of instruction. It determines the rate at which the probability of positively endorsing an item changes given the individual attitude/perception levels (Thorpe & Favia, 2012). The higher the discrimination parameter, the steeper the slope will be on the item characteristic curve, indicating a stronger ability to detect differences in the attitude/perception of respondents compared with less steep slopes.

The MLRT model was compared to a base IRT model (with no covariates) and to a one-level full model (with the same number of covariates as the MLRT model). The one-level full model performed better than the base model and the MLRT model on all five comparison criteria (p-values < 0.0001). Based on these comparisons (Table 3), we proceed to present results based on the 1-level full model.

Table 3: IRT Model Comparisons

Model	Criteria*					χ^2	df	p-value
	AIC	SABIC	HQ	BIC	logLik			
Base Model	112820.9	112944.8	112894.8	113040.2	-56380.46			
1-level	112617	112790.4	112720.4	112923.9	-56266.48	228	12	< 0.0001
MLRT	112883	113044.1	112979	113168	-56402.49			
1-level	112617	112790.4	112720.4	112923.9	-56266.48	272	3	< 0.0001

* AIC=Akaike Information, BIC=Bayesian Information, HQ=Hannan Quinn, logLik=Log Likelihood

The item discrimination parameter estimates (slopes) for the 2-parameter IRT models are given in Table 4, for both the new UMI 2021 survey questions and the random sample from the pre-2021 version of the survey (the UMI questions in use prior to 2021). Typically, the larger the discrimination parameter, the steeper the slope, which implies that the item is more effective at discriminating among different attitudes along the continuum. Thus, for a given level of endorsement, UMI question 6 in the pre-2021 SEI survey with a discrimination parameter of 8.67 would have more than 5 times the contribution to the survey information compared to UMI question 1 with a discrimination parameter of 3.62.

Yet a discrimination parameter of 8.67 is quite high, which is an indication that the survey question is not working properly. Reeve and Fayers (2005) suggest the useful range of discrimination values is from 0.5 to 2.5. Discrimination values above 2.5 don't add much to the slope of Item Characteristic Curves (ICC). However, a disproportionally large item slope indicates a disproportionally large contribution to the overall survey information.

Table 4: Item Discrimination Parameter Estimates

Data Source	Discrimination Parameter Estimates					
	UMI 1	UMI 2	UMI 3	UMI 4	UMI 5	UMI 6
UMI from the pre-2021 SEI Survey	3.62	5.38	4.15	2.02	3.28	8.67

UMI from the 2021 SEI Survey	3.26	4.80	3.83	3.15	3.00	5.85

In Table 4, UMI question 4 in the pre-2021 survey that asks if “the evaluation of student learning was fair” (2.02), has the least relative discrimination. However, the new UMI 4 question asking about “useful feedback” has a discrimination parameter that is comparable to other items (3.15), indicating that this item discriminates as much as the other items, among different attitude/perception levels.

Overall, the parameter estimates in the new UMI questions (2021 SEI survey) have been improved compared to those reported for the pre-2021 survey, and they are now more consistent across the items.

Figures 4 and 5 display the Item Information Curves (IIC) for each of the new 2021 SEI survey UMI questions, and for the pre-2021 survey UMI questions, respectively. The IICs measure the statistical information an individual item contributes to the overall survey. The x-axis is the individual’s level of endorsement; a person with an endorsement level of 2 has a more positive attitude regarding the course than someone with a level of -0.2. The y-axis indicates the magnitude of the information provided by each of the survey items. Higher information signifies higher precision (or reliability) in differentiating among respondents (Reeve & Fayers, 2005). In addition, items should be well spaced across the continuum (x-axis).

There are notable differences evident when comparing the item information curves in Figure 4 and 5. Figure 4 indicates improvement in the relative contributions of all new UMI questions to the overall survey information compared with the pre-2021 survey sample, notably for UMI questions 2 and 3 and 4. Furthermore, the newly-worded 2021 UMI items shown in Figure 4 appear to differentiate across a broader range on the x-axis than the pre-2021 survey UMI items shown in Figure 5. The y-axis scales differ between Figures 4 and 5 as a result of the disproportionately large UMI 6 discrimination parameter (8.67) in Figure 5. Although UMI 6 has a relatively large discrimination parameter estimate in the new 2021 survey (5.85), it appears to discriminate across a similar range on the x-axis, but it displays sharp peaks on the information curve, which implies that the item is not functioning as well as it could. However, the new UMI 6 peaks (Figure 4) were less jagged and show improvement compared to that of the pre-2021 UMI 6 (Figure 5).

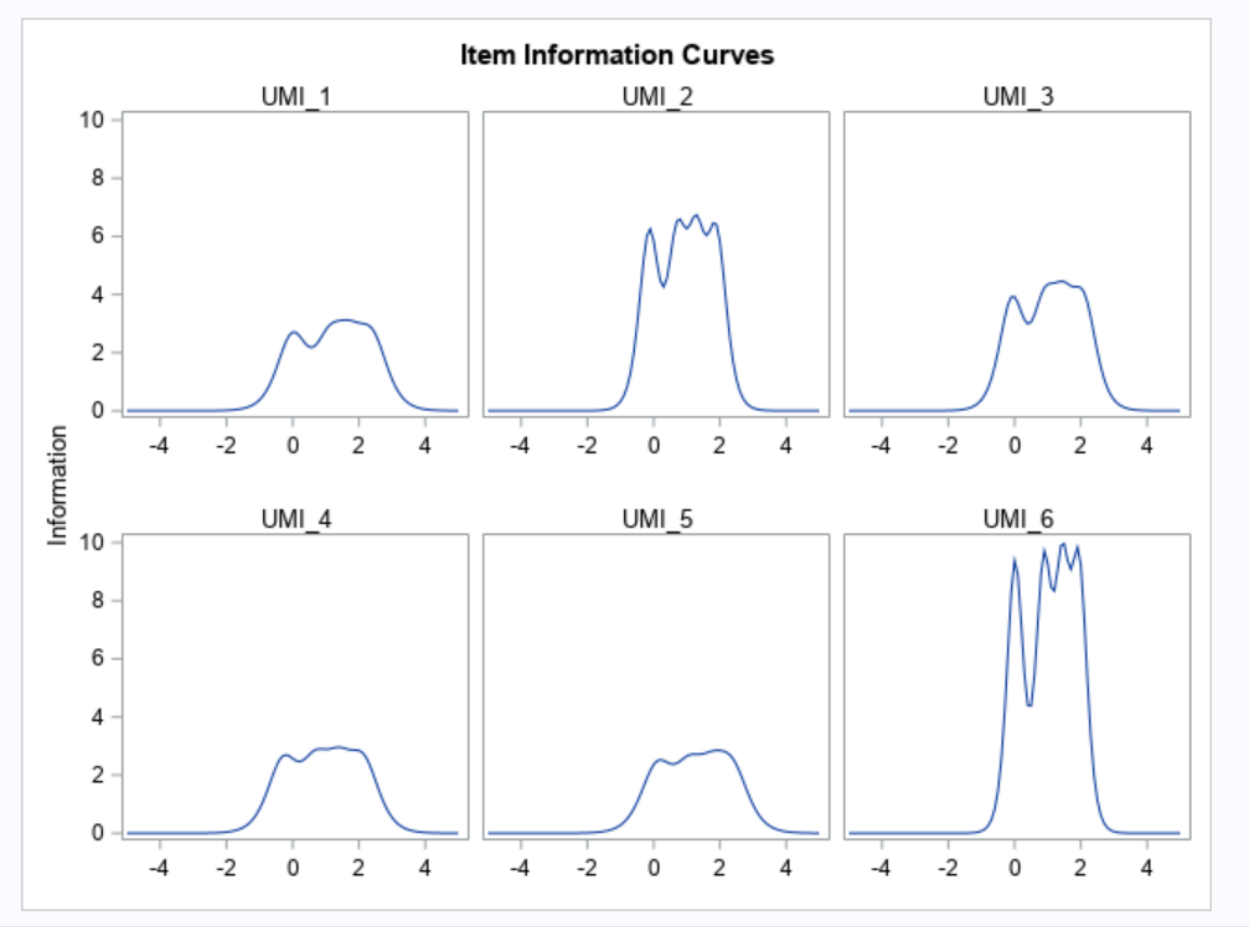


Figure 4: Item Information Curves for the new 2021 SEI Survey UMI questions

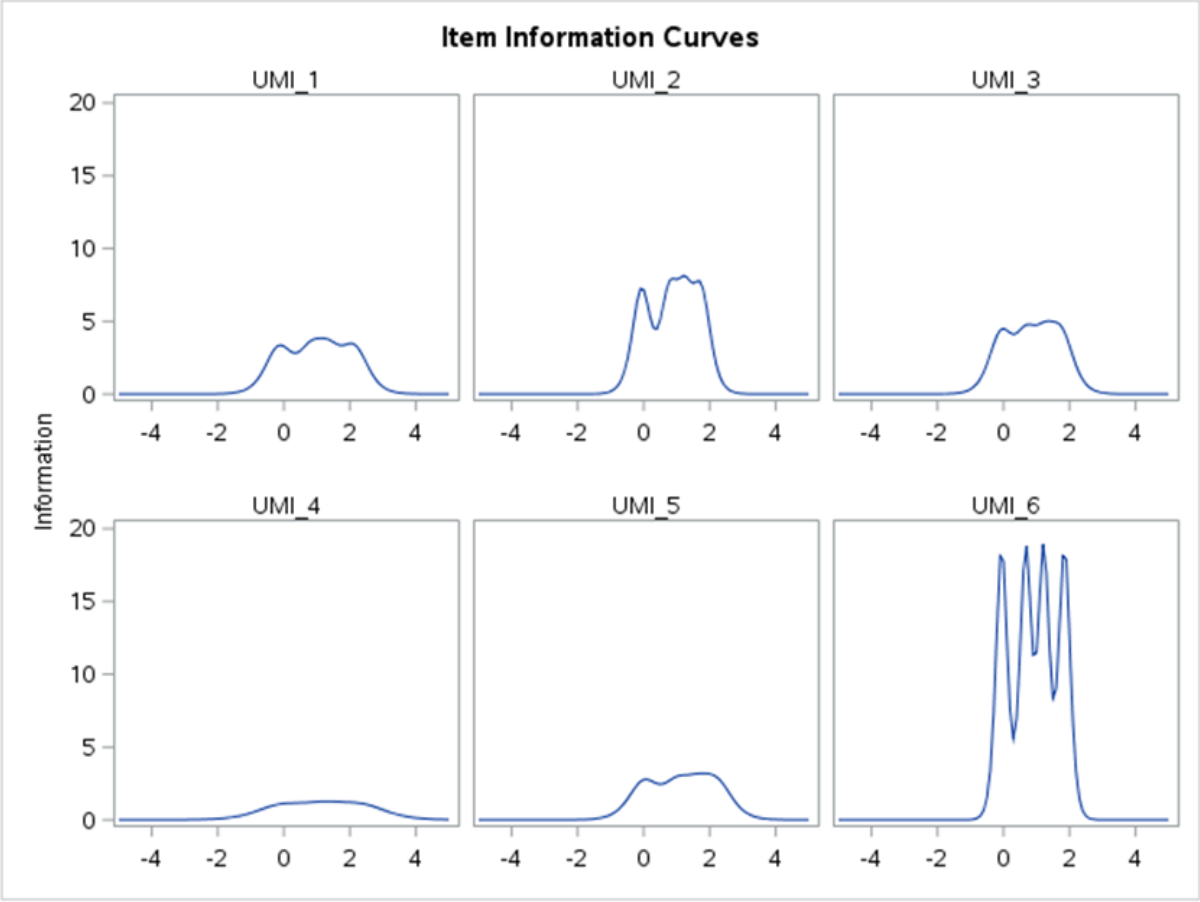


Figure 5: Item Information Curves for the pre-2021 SEI Survey UMI questions

Looking at Figure 5, the IICs for the pre-2021 UMI questions show that UMI 6 disproportionally contributes to the overall survey information; however, for the new set of UMI questions, the contribution of each item seems to be more consistent. Overall, the proposed changes to the UMI questions appear to have improved their relative discrimination among students with varying levels of endorsements for most items.

3.4 GENERALIZED LINEAR MIXED MODELS

We used a Generalized Linear Mixed Model (GLMM) approach to model variation in SEI scores within 5 fields of study (Sciences, Humanities, Health Sciences, Engineering and Social Sciences). In this approach, respondents to SEI surveys are considered to be clustered within fields of study (grouping variable the GLMM with a random intercept). Proc GLIMMIX in the SAS statistical software was used to fit the cumulative logit of the probability of higher SEI ratings in the response profile (corresponding to the 5-point Likert scale) as a function of course attributes (year level and meeting time), instructor demographics (rank and gender) and student gender; and with the field of study as a grouping variable.

The estimated covariance parameters, which measures the variation in field of study effects, for the six UMI questions are shown in Table 5. For each UMI question, the estimated variance of the field of study random intercepts is given along with standard error and p-value for testing if the variance is significantly different from zero.

Table 5: Estimated variance of the field of study random intercepts in the GLMM

Question	Covariate Estimate	Standard Error	Z value	p-value
UMI 1	0.0092	0.0081	1.13	0.1282
UMI 2	0.0302	0.0230	1.32	0.094
UMI 3	0.0314	0.0239	1.31	0.0943
UMI 4	0.0355	0.0266	1.33	0.0911
UMI 5	0.0315	0.0239	1.32	0.0936
UMI 6	0.0301	0.0230	1.31	0.095

The estimated values for all UMI questions in Table 5 are not significantly larger than 0 (p-values > 0.05), which indicates that there is no significant variation in the field of study effect on SEI ratings (no significant random effect). A Generalized Linear Model (GLM) across all fields of study (no field of study random intercept) was also fitted to the data. There are minor differences between the GLM and GLMM model. However, the GLMM model is preferred as it explained added variance (though not statistically significant) that could impact the effect of other variables in the model. Tests of the model fixed effects are shown in Table 6.

Table 6: P-values for the model fixed effects

Question	Instructor Rank	Instructor Gender	Student Gender	Year Level	Meeting Time
UMI 1	**< 0.001**	0.050	**0.025**	**0.002**	0.055
UMI 2	**< 0.001**	0.142	**0.025**	**< 0.001**	0.105
UMI 3	**< 0.001**	**0.004**	**0.023**	**< 0.001**	0.643
UMI 4	**< 0.001**	0.080	0.071	**< 0.001**	0.154
UMI 5	**< 0.001**	**0.012**	0.148	**< 0.001**	0.109
UMI 6	**< 0.001**	0.266	**0.007**	**< 0.001**	0.225

Model parameter estimates and associated statistics for fixed effects are shown in the Appendix to this report. For all UMI questions, there were no significant differences in SEI ratings between course sections that met before or after 11:00 AM.

SEI ratings for 1st, 2nd and 3rd year courses were consistently significantly lower compared to 4th and 5th year courses. It is important to note that these differences are not due to Differential Item Functioning (see table 2 for DIF results). Recall that DIF is conceptualized as occurring when survey respondents who have similar attitudes/perceptions on a measured trait respond differently due to construct-irrelevant factors, i.e., DIF analysis takes into consideration the sum of scores for all UMI questions as a measure of respondent attitude/perception.

Female instructors received relatively higher ratings compared to their male counterparts in UMI questions 3 (“The Instructor presented the course material in a way that I could understand”) and 5 (“The instructor showed genuine interest in supporting my learning throughout this course”). However, the odds ratio for the two questions were relatively small (1.3 and 1.2, respectively). Chen, Patricia Cohen & Sophie Chen (2010) showed that odd ratios < 1.5 translate to small effect size. There were no instructor gender differences in the other 4 UMI questions.

Female students rated their experience of instruction significantly higher compared to male students in UMI questions 1, 2, 3 and 6. Again, though statistically significant, odds ratios were close to 1.0 (1.1 for UMI questions 1, 2, and 3 and 1.2 for UMI 6).

There were also differences in ratings depending on instructor rank for all UMI questions. However, differences between instructor ranks and their magnitudes vary across questions, but odds ratios were relatively small (< 1.4), with slightly higher ratings for assistant professors and lecturers. Also, it is important to note that instructor rank was based on SEI survey data which reports “Standard Job Title” and does not consider tenure or other relevant appointment information.

Finally, there were consistent and significant differences in SEI ratings between fields of study, with Humanities rated higher compared to the overall average, but with odd ratios not exceeding 1.2 for all UMI questions.

4.0 CONCLUSION

The Item Response Theory (IRT) results indicated that the new UMI questions implemented in 2021 seem to function better than previous UMI questions. In the old version, UMI question 6 provided most of the statistical information for the overall survey, but did not differentiate broadly among respondents’ attitudes/perceptions. Furthermore, the presence of sharp peaks in the item information curve indicates the item was not functioning well. The Item Information results were similar to those obtained in a 2021 pilot study (McKeown, Zumrawi & Pena, 2021) and provide further evidence that the new UMI questions are more consistent in their contribution to the overall survey, and are more widespread across the attitudinal continuum (x-axis).

While most of the new 2021 survey UMI questions showed no DIF among different grouping by student, instructor or class attributes, UMI 1 exhibited moderate to large DIF, and UMI 6 exhibited moderate DIF between class sizes. Moderate DIF between genders was also detected for UMI 6, with female students positively endorsing that question more than male students (recall that only binary data were used for gender based on challenges with using Employment Equity Survey data in these analyses). However, this result was not consistent across test methods and thus was not conclusive. Negligible/moderate DIF in instructor gender was also detected for UMI 3, with female instructors receiving slightly more positive endorsement on this item, however, the direction (favouring female instructors) was consistent with previous studies at UBC (CTLT, 2010).

GLMM results showed that SEI ratings for 1st, 2nd and 3rd year courses were consistently significantly lower compared to 4th and 5th year courses. Also, female instructors received slightly higher ratings (on UMI 3 and 5) and female students rated their instructors slightly higher (on UMI 1, 2, 3 & 6) compared to their male counterparts. However, in both cases the effect sizes were small. Finally, there were also significant differences in ratings depending on instructor rank for all UMI questions. Differences between instructor ranks and their magnitudes vary across questions, but odds ratios were relatively small (< 1.4), mostly favouring assistant professors and lecturers.

Due to the lack of sufficient Employment Equity Survey data, we were not able to test how the new UMI questions function across other variables of interest, e.g., gender identity, ethnicity, disability, and more. Thus, and based on these results, we recommend that further IRT and DIF analysis be carried out on the new UMI questions. Furthermore, we will continue to monitor the Employment Equity Survey response rate and examine the randomness of missing data.

References

Centre for Teaching, Learning & Technology. (2010). An investigation into the effects of instructor gender, field of study, and student respondent gender on UMI Scores in the 2008-09 SEoT administration. https://seoi.ubc.ca/files/2020/10/SEoT-Gender-X-Field-of-Study-Analysis-Revised-10-27-09.pdf.

Columbia University Mailman School of Public Health. (2019*).* Item Response Theory. https://www.publichealth.columbia.edu/research/population-health-methods/item-response-theory.

Hambleton, R.K., Swaminathan, H., & Rogers H. J. (1991). *Fundamentals of Item Response Theory*. Sage Publications.

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics—Simulation and Computation*, *39*(4), 860-864. https://doi.org/10.1080/03610911003650383

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Dirs.), *Test validity*. Lawrence Erlbaum Associates.

Li, Z. (2015). A power formula for the Mantel-Haenszel test for Differential Item Functioning. *Applied Psychological Measurement*, *39*(5), 373–388.

Linacre, J. M. (n.d.). Mantel & Mantel-Haenszel DIF statistics. http://www.winsteps.com/winman/mantel_and_mantel-haenszel_dif.htm

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, *7*(4), 328.

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

McKeown S., Zumrawi A., & Pena C. (2021). Re-envisioning the Student Experience of Instruction survey questions from the student perspective. Report to the University of British Columbia Senate Teaching & Learning Committee. https://seoi.ubc.ca/files/2021/10/SEI-Report-to-Senate-Committees-August-2021.pdf.

Paek, I., & Wilson, M. (2011). Formulating the Rasch Differential Item Functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel– Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement*, *71*(6), 1023-1046.

Reeve, B.,& Fayers, P. M. (2005). Applying Item Response Theory modelling for evaluating questionnaire item and scale properties. In P. M. Fayers & R. D. Hays (Eds.), *Assessing Quality of Life in Clinical Trials: Methods and Practice,* 2nd ed. *(*pp. 55-73). Oxford University Press.

Ryan, K., Gannon-Slater, N., & Culbertson, M. J. (2012). Improving survey methods with cognitive interviews in small- and medium-scale evaluations. *American Journal of Evaluation*, *33*(3)*,* 414-430.

Şahin, A., & Anil, D. (2017). The effects of test length and sample size on item parameters in Item Response Theory. *Educational Sciences: Theory & Practice*, *17(*1), 321-335.

Thorpe, G. L., & Favia, A. (2012). Data analysis using Item Response Theory methodology: An introduction to selected programs and applications. *Psychology Faculty Scholarship*, *20*. https://digitalcommons.library.umaine.edu/psy_facpub/20

Trenor, J. M., Miller, M .K., & Gipson, K. G. (2011). *Utilization of a think-aloud protocol to cognitively validate a survey instrument identifying social capital resources of engineering undergraduates.* Paper presented at 2011 ASEE Annual Conference & Exposition, Vancouver, BC.

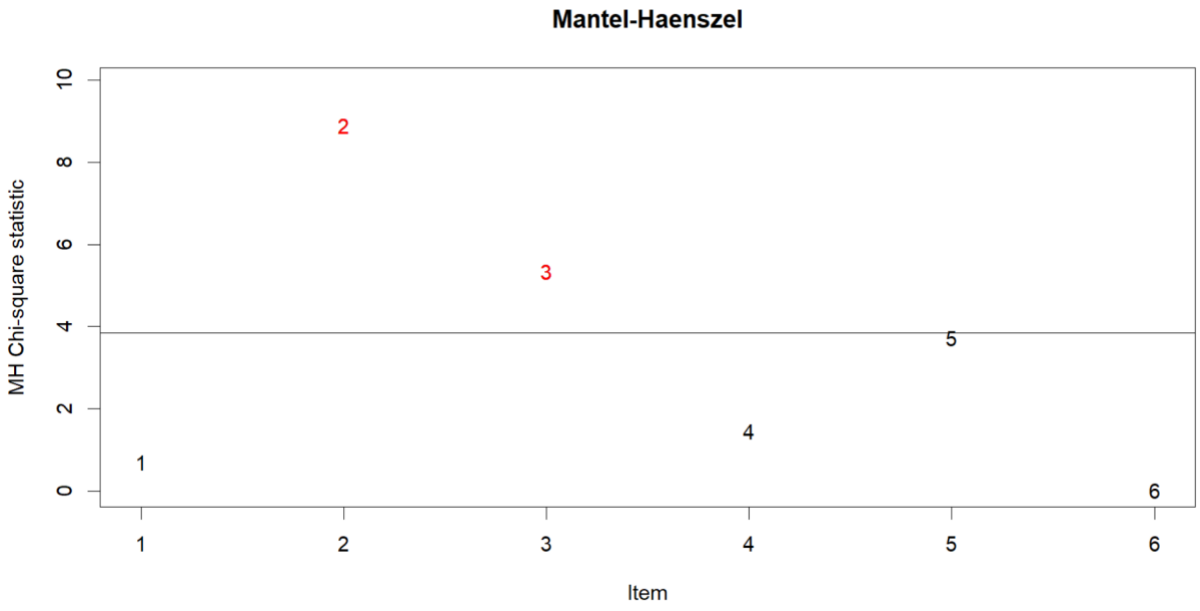
Zumbo, B. (1999). *A handbook on the theory and methods of Differential Item Functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

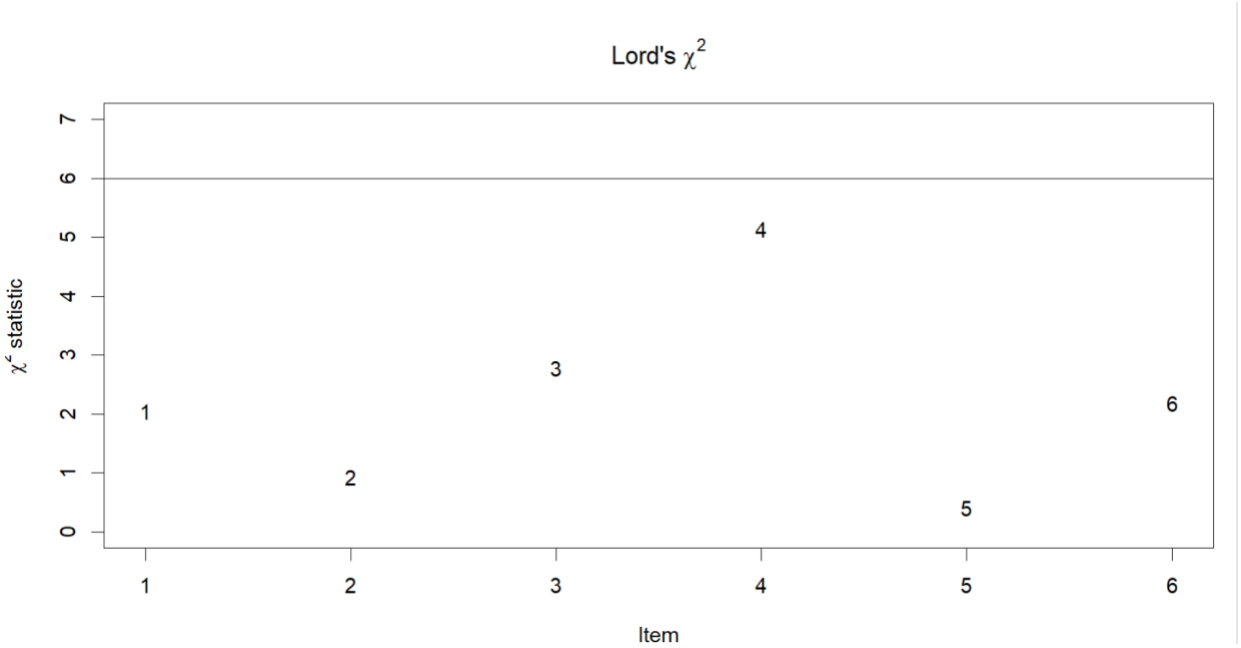
Zwick, R., Thayer, D. T., & Lewis, C. (1999) An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*(1), 1-28.

Appendix 4A

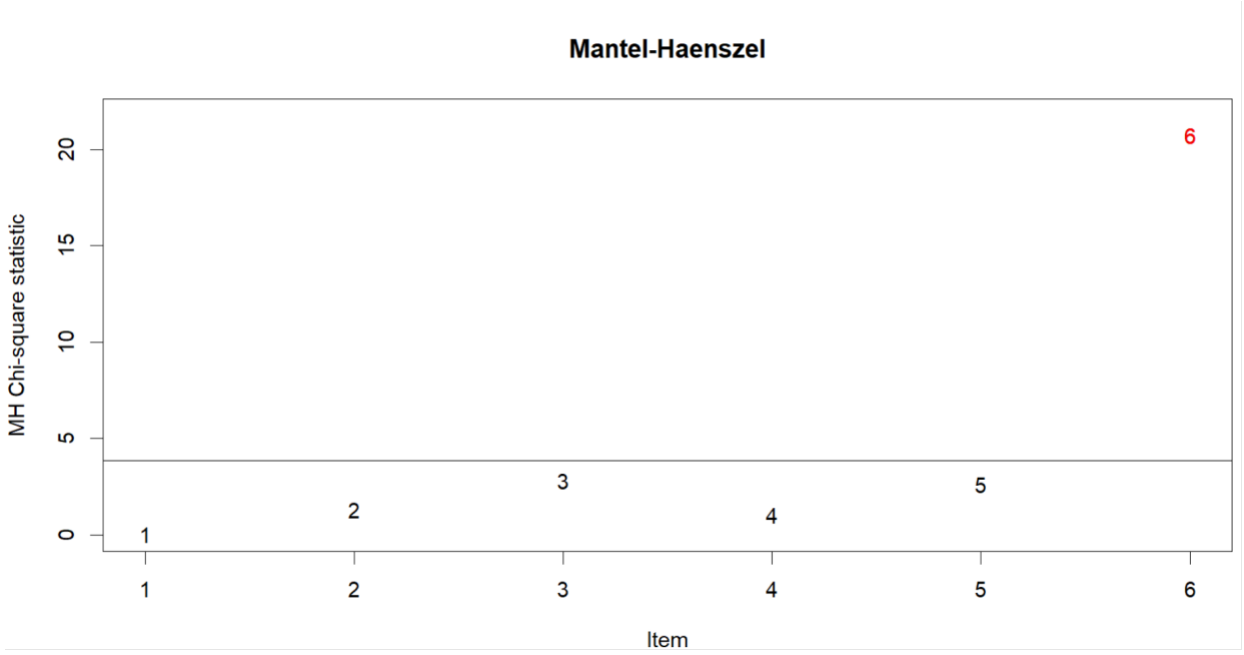
Graphical Representations of the Mantel-Haenszel and Lord's DIF Statistics

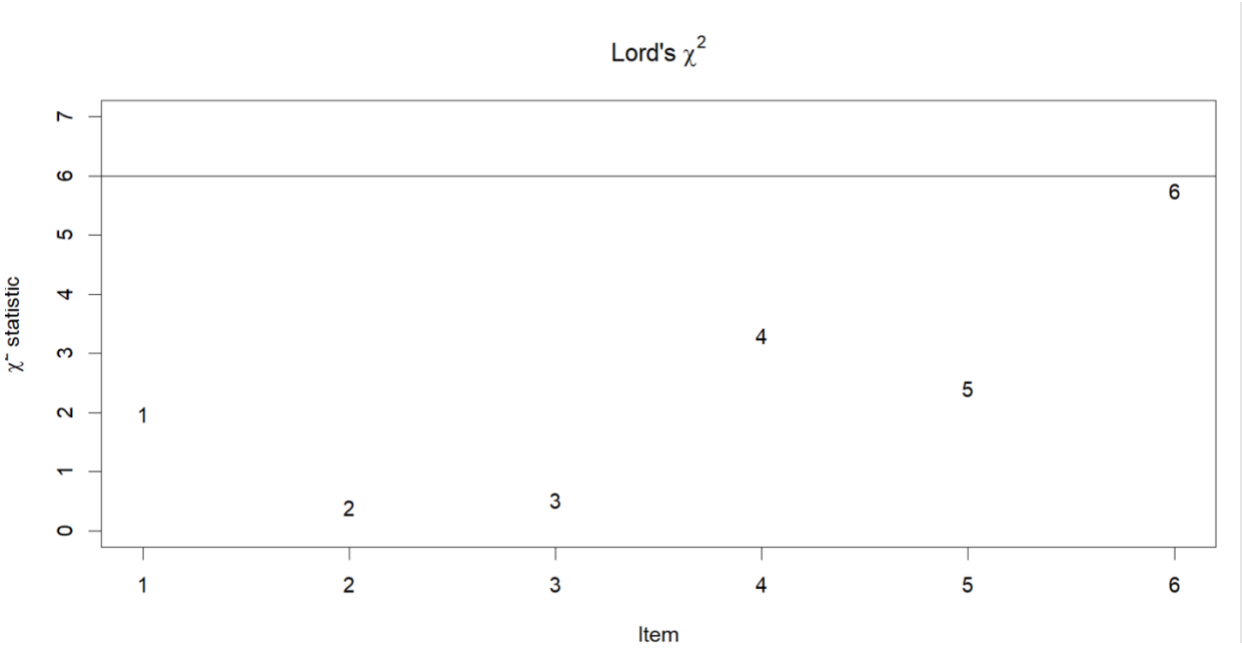
Campus (UBCO vs UBCV)



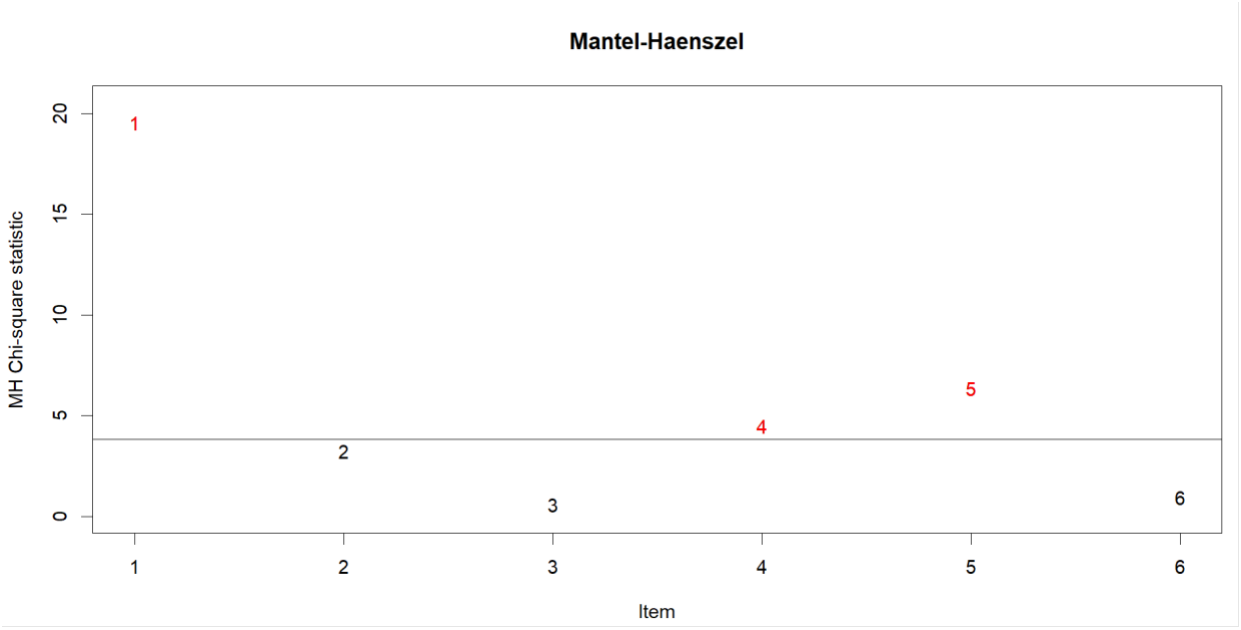


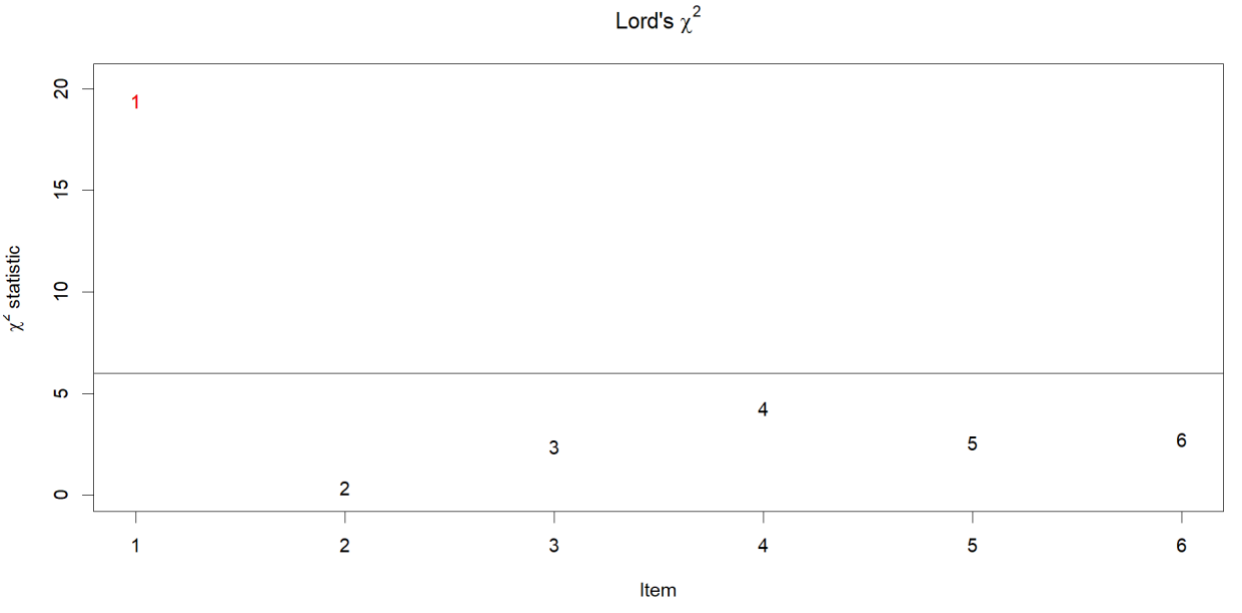
Student Gender



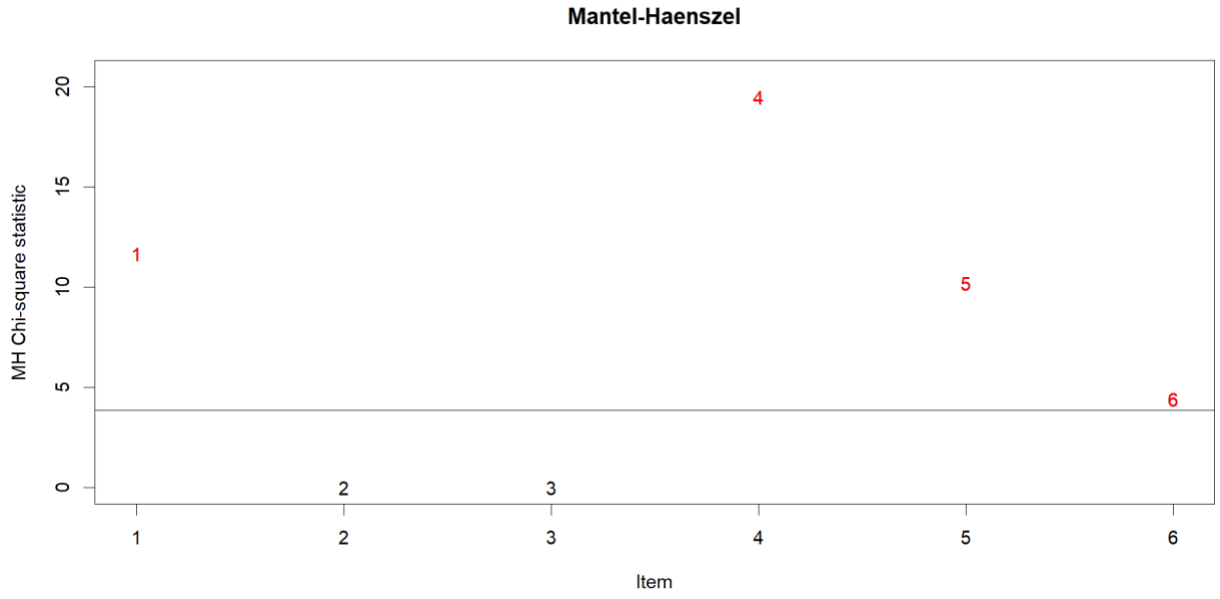

Lord's χ^2
χ^2 statistic
0
1
2
3
4
5
6
7
1
2
3
4
5
6
Item

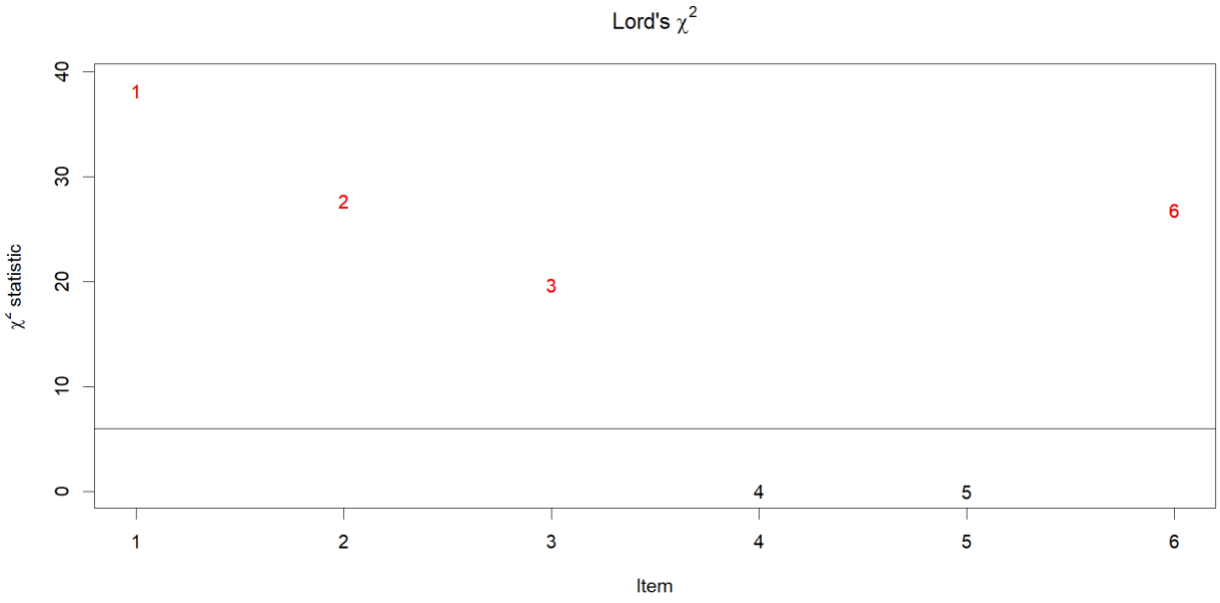
Class Size (< 100 vs 100+)



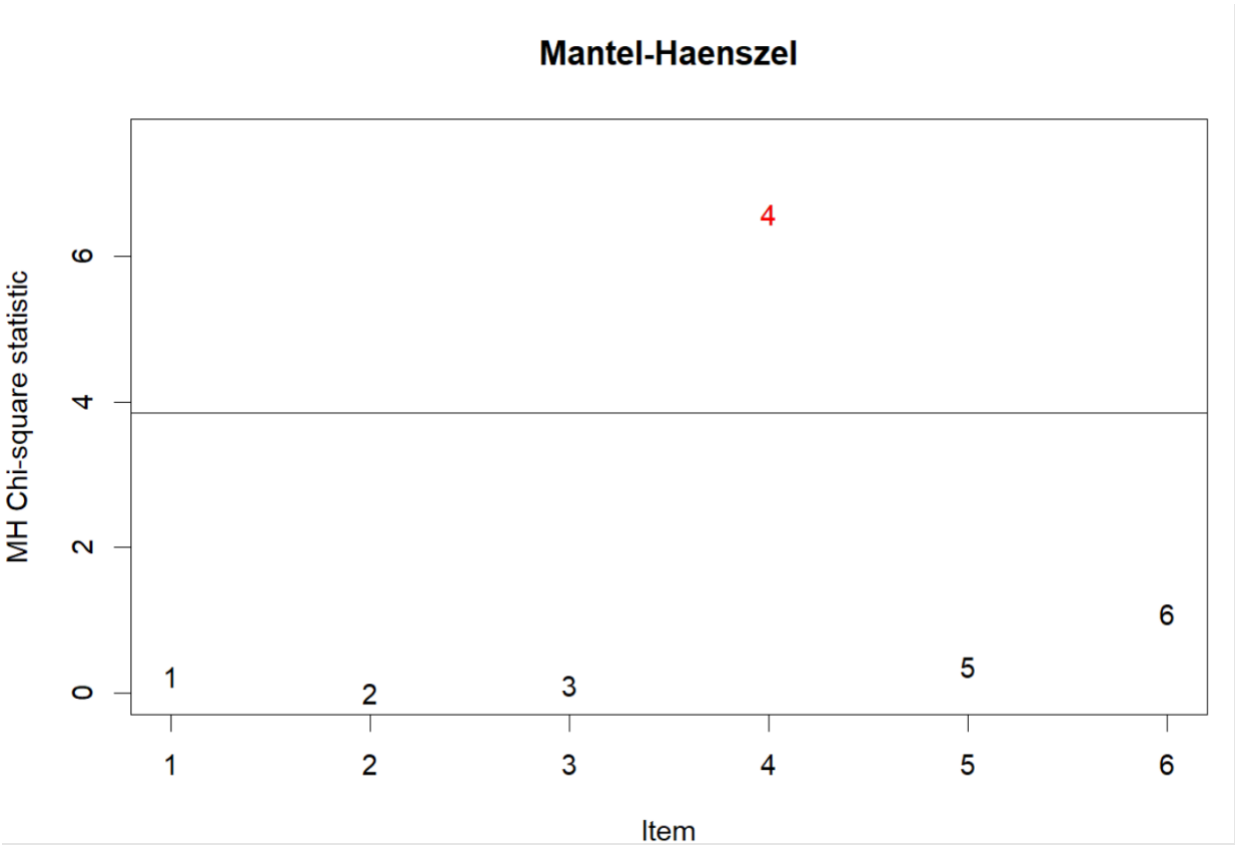


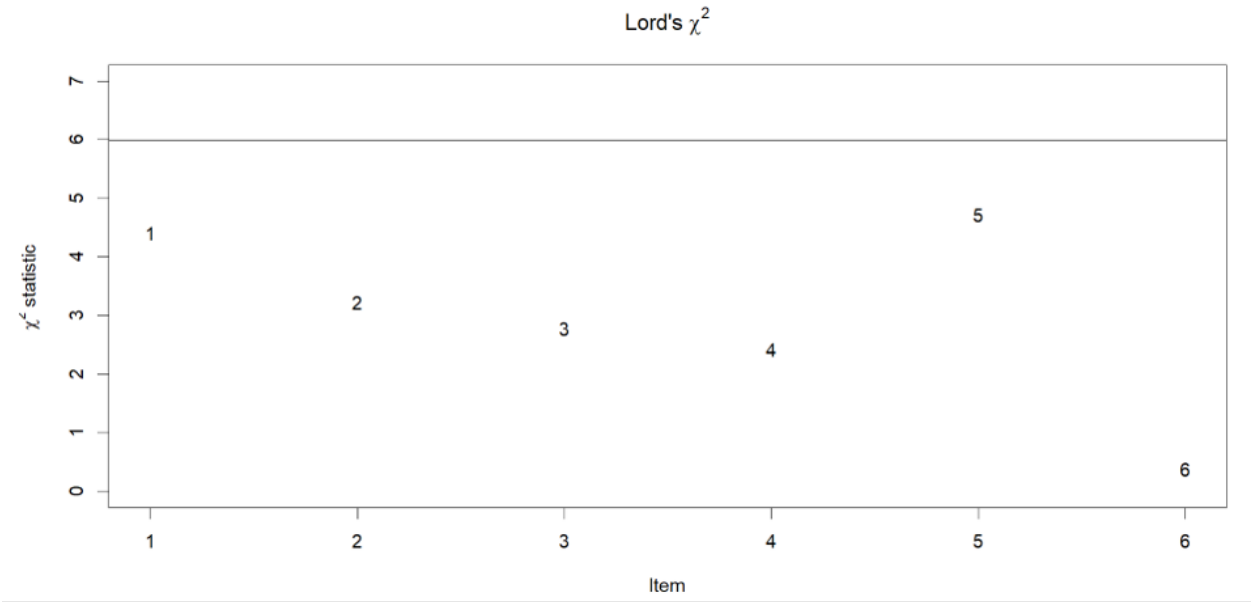
Class Size (1-49 vs 200+)



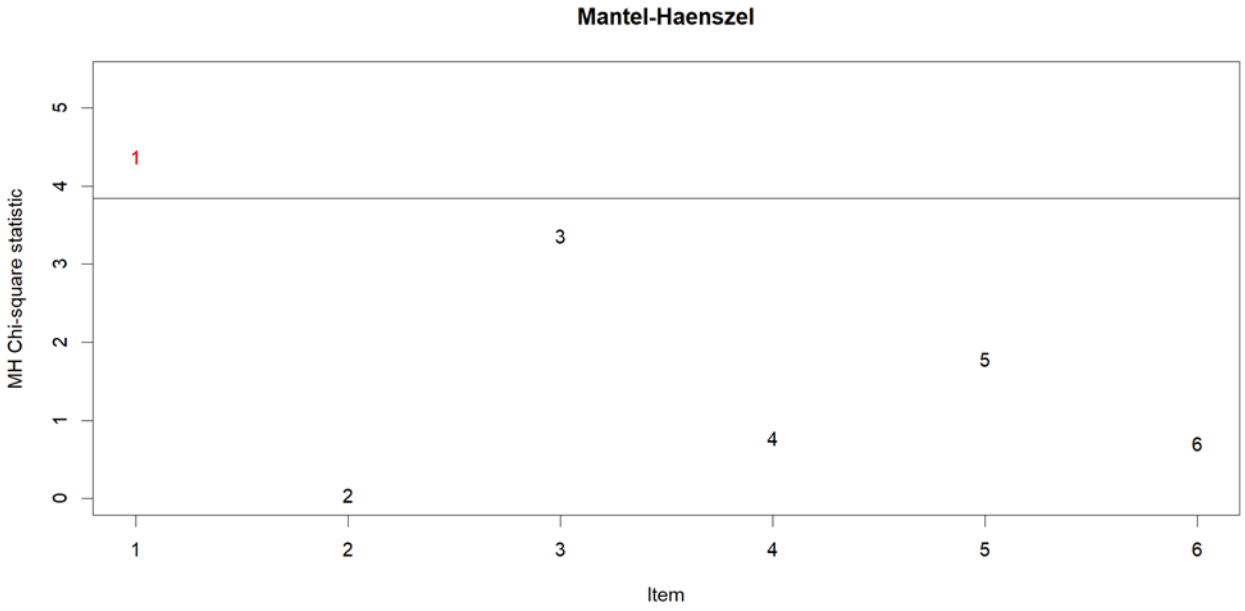


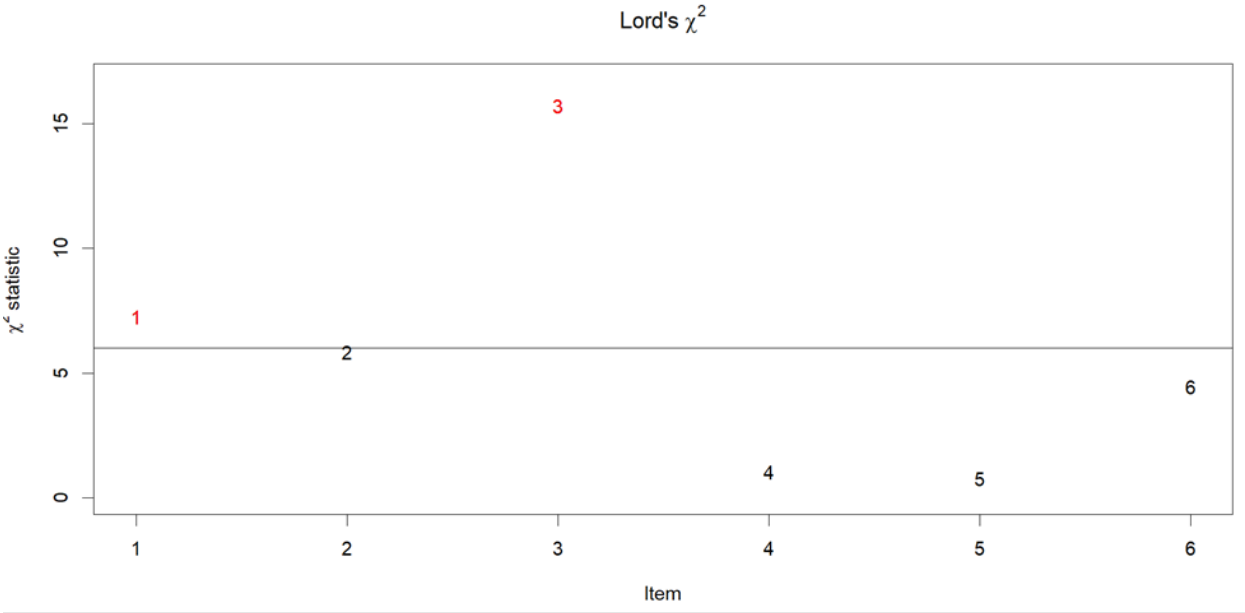
Class Meeting Time (before 11:00 AM vs after)



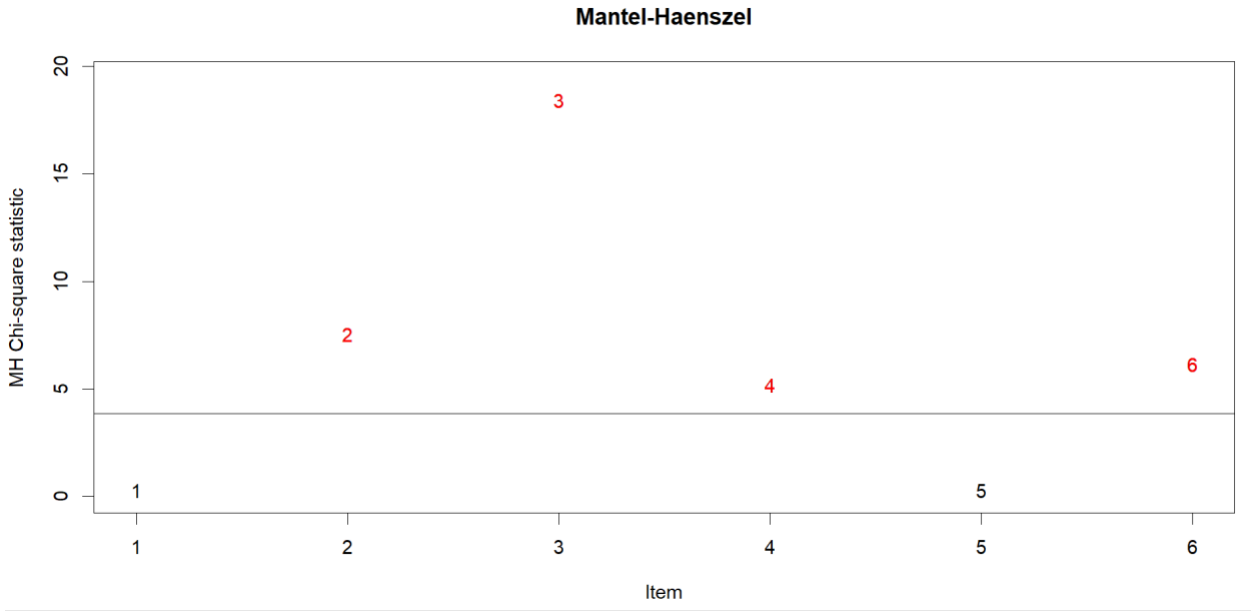


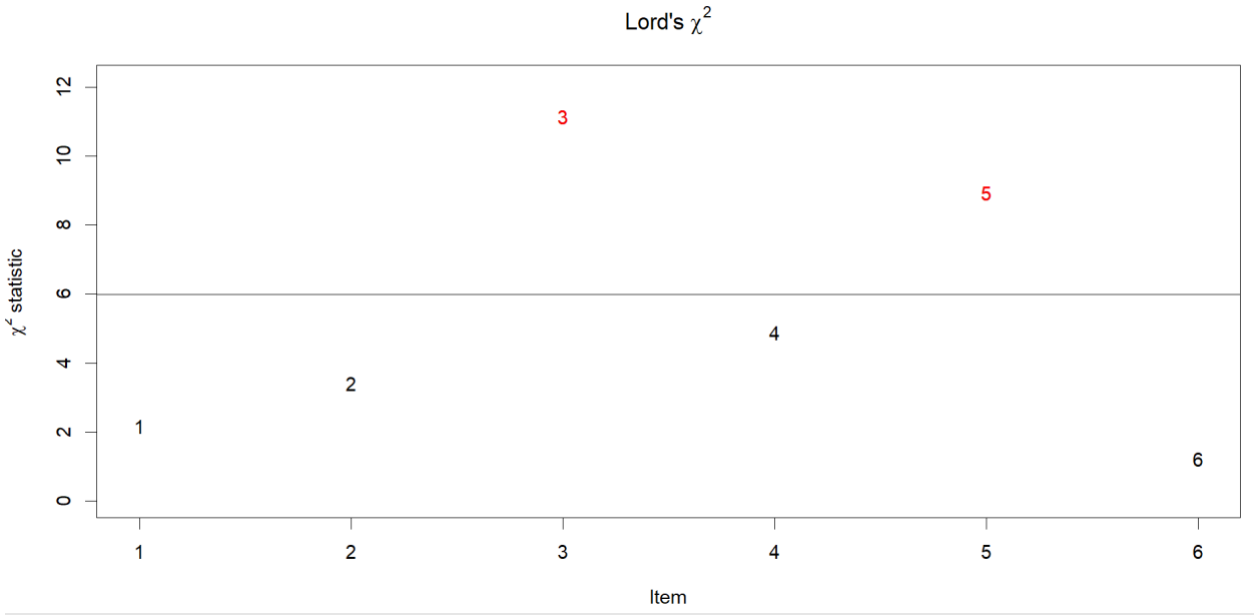
Year Level (1st, 2nd & 3rd vs 4th & 5th)





Instructor Gender





IRT Model Parameter Estimates and Associated Statistics

Item	Parameter	Estimate
UMI_1	Threshold 1	7.04
	Threshold 2	4.85
	Threshold 3	2.90
	Threshold 4	-0.84
	Slope	3.26
UMI_2	Threshold 1	8.43
	Threshold 2	5.48
	Threshold 3	2.69
	Threshold 4	-1.47
	Slope	4.80
UMI_3	Threshold 1	7.29
	Threshold 2	4.92
	Threshold 3	2.74
	Threshold 4	-1.12
	Slope	3.83
UMI_4	Threshold 1	5.99
	Threshold 2	3.66
	Threshold 3	1.36
	Threshold 4	-1.75
	Slope	3.15
UMI_5	Threshold 1	6.40
	Threshold 2	4.67
	Threshold 3	2.46
	Threshold 4	-0.48
	Slope	3.00
UMI_6	Threshold 1	10.73
	Threshold 2	7.78
	Threshold 3	4.47
	Threshold 4	-0.73

Item	Parameter	Estimate
	Slope	5.85

Generalized Linear Mixed Model Fixed Effects Estimates and Associated Statistics

UMI 1

Solutions for Fixed Effects											
Effect	UMI_1	Inst_Gend	Stud_Gen	Meeting_t	Rank	Level	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	5						0.05708	0.08514	1	0.67	0.624
Intercept	4						1.8405	0.08732	1	21.08	0.0302
Intercept	3						2.8789	0.09237	1	31.17	0.0204
Intercept	2						4.1038	0.1091	1	37.6	0.0169
Rank					Assoc. Prof		-0.1791	0.0596	16	-3.01	0.0084
Rank					Asst. Prof		0.2012	0.053	16	3.8	0.0016
Rank					Lecturer		0.1794	0.06143	16	2.92	0.01
Rank					Professor		-0.2548	0.05917	16	-4.31	0.0005
Rank					Sessional		0	.	.	.	.
Inst_Gender		F					0.1122	0.04029	4	2.79	0.0495
Inst_Gender		M					0	.	.	.	.
Stud_Gender			F				0.1333	0.03833	4	3.48	0.0254
Stud_Gender			M				0	.	.	.	.
Level						1	-0.291	0.06793	16	-4.28	0.0006
Level						2	-0.1728	0.06771	16	-2.55	0.0213
Level						3	-0.2693	0.0686	16	-3.93	0.0012
Level						4	-0.06403	0.08565	16	-0.75	0.4656
Level						5	0	.	.	.	.
Meeting_time				Early			-0.1081	0.0403	4	-2.68	0.0551
Meeting_time				Late			0	.	.	.	.

UMI 2

Solutions for Fixed Effects											
Effect	UMI_2	Inst_Gend	Stud_Gen	Meeting_t	Rank	Level	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	5						-0.02282	0.1061	1	-0.22	0.8651
Intercept	4						1.3846	0.107	1	12.94	0.0491
Intercept	3						2.4251	0.1093	1	22.18	0.0287
Intercept	2						3.5921	0.1168	1	30.75	0.0207
Rank					Assoc. Prof		-0.09432	0.05813	16	-1.62	0.1242
Rank					Asst. Prof		0.2344	0.05151	16	4.55	0.0003
Rank					Lecturer		0.2741	0.05989	16	4.58	0.0003
Rank					Professor		-0.1642	0.05767	16	-2.85	0.0117
Rank					Sessional		0	.	.	.	.
Inst_Gender		F					0.07186	0.03935	4	1.83	0.1419
Inst_Gender		M					0	.	.	.	.
Stud_Gender			F				0.1304	0.03751	4	3.48	0.0254
Stud_Gender			M				0	.	.	.	.
Level						1	-0.5646	0.0669	16	-8.44	<.0001
Level						2	-0.3635	0.0665	16	-5.47	<.0001
Level						3	-0.4312	0.06741	16	-6.4	<.0001
Level						4	-0.09493	0.08437	16	-1.13	0.2771
Level						5	0	.	.	.	.
Meeting_time				Early			-0.0821	0.03936	4	-2.09	0.1053
Meeting_time				Late			0	.	.	.	.

UMI 3

Solutions for Fixed Effects											
Effect	UMI_3	Inst_Gend	Stud_Gen	Meeting_t	Rank	Level	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	5						-0.1452	0.1075	1	-1.35	0.4057
Intercept	4						1.4667	0.1086	1	13.51	0.047
Intercept	3						2.488	0.1115	1	22.32	0.0285
Intercept	2						3.6359	0.1208	1	30.11	0.0211
Rank					Assoc. Prof		-0.2326	0.05882	16	-3.95	0.0011
Rank					Asst. Prof		0.2267	0.05234	16	4.33	0.0005
Rank					Lecturer		0.2338	0.06093	16	3.84	0.0015
Rank					Professor		-0.1638	0.05842	16	-2.8	0.0127
Rank					Sessional		0	.	.	.	.
Inst_Gender		F					0.2331	0.04003	4	5.82	0.0043
Inst_Gender		M					0	.	.	.	.
Stud_Gender			F				0.1373	0.03801	4	3.61	0.0225
Stud_Gender			M				0	.	.	.	.
Level						1	-0.3256	0.06722	16	-4.84	0.0002
Level						2	-0.1725	0.06687	16	-2.58	0.0202
Level						3	-0.3186	0.06772	16	-4.7	0.0002
Level						4	0.08401	0.08513	16	0.99	0.3384
Level						5	0	.	.	.	.
Meeting_time				Early			-0.01998	0.0399	4	-0.5	0.6428
Meeting_time				Late			0	.	.	.	

Solutions for Fixed Effects											
Effect	UMI_3	Inst_Gend	Stud_Gen	Meeting_t	Rank	Level	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	5						-0.1452	0.1075	1	-1.35	0.4057
Intercept	4						1.4667	0.1086	1	13.51	0.047
Intercept	3						2.488	0.1115	1	22.32	0.0285
Intercept	2						3.6359	0.1208	1	30.11	0.0211
Rank					Assoc. Prof		-0.2326	0.05882	16	-3.95	0.0011
Rank					Asst. Prof		0.2267	0.05234	16	4.33	0.0005
Rank					Lecturer		0.2338	0.06093	16	3.84	0.0015
Rank					Professor		-0.1638	0.05842	16	-2.8	0.0127
Rank					Sessional		0	.	.	.	.
Inst_Gender		F					0.2331	0.04003	4	5.82	0.0043
Inst_Gender		M					0	.	.	.	.
Stud_Gender			F				0.1373	0.03801	4	3.61	0.0225
Stud_Gender			M				0	.	.	.	.
Level						1	-0.3256	0.06722	16	-4.84	0.0002
Level						2	-0.1725	0.06687	16	-2.58	0.0202
Level						3	-0.3186	0.06772	16	-4.7	0.0002
Level						4	0.08401	0.08513	16	0.99	0.3384
Level						5	0	.	.	.	.
Meeting_time				Early			-0.01998	0.0399	4	-0.5	0.6428
Meeting_time				Late			0	.	.	.	

UMI 4

Solutions for Fixed Effects											
Effect	UMI_4	Inst_Gend	Stud_Gen	Meeting_t	Rank	Level	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	5						-0.2062	0.1105	1	-1.87	0.3132
Intercept	4						1.2403	0.1111	1	11.16	0.0569
Intercept	3						2.4616	0.1136	1	21.66	0.0294
Intercept	2						3.7564	0.1224	1	30.69	0.0207
Rank					Assoc. Prof		-0.03483	0.05769	16	-0.6	0.5545
Rank					Asst. Prof		0.2854	0.05098	16	5.6	<.0001
Rank					Lecturer		0.3256	0.05928	16	5.49	<.0001
Rank					Professor		-0.0981	0.05727	16	-1.71	0.1061
Rank					Sessional		0	.	.	.	.
Inst_Gender		F					0.09095	0.039	4	2.33	0.0801
Inst_Gender		M					0	.	.	.	.
Stud_Gender			F				0.09108	0.03722	4	2.45	0.0707
Stud_Gender			M				0	.	.	.	.
Level						1	-0.6765	0.06618	16	-10.22	<.0001
Level						2	-0.3806	0.06574	16	-5.79	<.0001
Level						3	-0.4439	0.06665	16	-6.66	<.0001
Level						4	-0.06917	0.08341	16	-0.83	0.4191
Level						5	0	.	.	.	.
Meeting_time				Early			-0.06867	0.03906	4	-1.76	0.1536
Meeting_time				Late			0	.	.	.	.

UMI 5

Solutions for Fixed Effects											
Effect	UMI_5	Inst_Gend	Stud_Gen	Meeting_t	Rank	Level	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	5						0.5129	0.1103	1	4.65	0.1349
Intercept	4						2.0383	0.1121	1	18.18	0.035
Intercept	3						3.3121	0.1172	1	28.26	0.0225
Intercept	2						4.3322	0.1288	1	33.65	0.0189
Rank					Assoc. Prof		-0.1554	0.06012	16	-2.59	0.0199
Rank					Asst. Prof		0.2464	0.05393	16	4.57	0.0003
Rank					Lecturer		0.2482	0.06258	16	3.97	0.0011
Rank					Professor		-0.145	0.05972	16	-2.43	0.0273
Rank					Sessional		0	.	.	.	.
Inst_Gender		F					0.1802	0.04111	4	4.38	0.0118
Inst_Gender		M					0	.	.	.	.
Stud_Gender			F				0.06989	0.03904	4	1.79	0.1479
Stud_Gender			M				0	.	.	.	.
Level						1	-0.8056	0.07128	16	-11.3	<.0001
Level						2	-0.5272	0.07098	16	-7.43	<.0001
Level						3	-0.4657	0.07209	16	-6.46	<.0001
Level						4	-0.1501	0.09028	16	-1.66	0.1158
Level						5	0	.	.	.	.
Meeting_time				Early			-0.08403	0.04092	4	-2.05	0.1093
Meeting_time				Late			0	.	.	.	.

UMI 6

Solutions for Fixed Effects											
Effect	UMI_6	Inst_Gend	Stud_Gen	Meeting_t	Rank	Level	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	5						0.1479	0.1071	1	1.38	0.3989
Intercept	4						1.6941	0.1084	1	15.63	0.0407
Intercept	3						2.7523	0.1117	1	24.64	0.0258
Intercept	2						3.7461	0.1198	1	31.26	0.0204
Rank					Assoc. Prof		-0.1526	0.05917	16	-2.58	0.0202
Rank					Asst. Prof		0.2213	0.05269	16	4.2	0.0007
Rank					Lecturer		0.2147	0.06115	16	3.51	0.0029
Rank					Professor		-0.1806	0.05873	16	-3.08	0.0072
Rank					Sessional		0	.	.	.	.
Inst_Gender		F					0.05188	0.04017	4	1.29	0.2661
Inst_Gender		M					0	.	.	.	.
Stud_Gender			F				0.1926	0.03822	4	5.04	0.0073
Stud_Gender			M				0	.	.	.	.
Level						1	-0.5245	0.06848	16	-7.66	<.0001
Level						2	-0.3148	0.06812	16	-4.62	0.0003
Level						3	-0.4047	0.06901	16	-5.86	<.0001
Level						4	-0.1263	0.08625	16	-1.46	0.1625
Level						5	0	.	.	.	.
Meeting_time				Early			-0.05758	0.04013	4	-1.44	0.2246
Meeting_time				Late			0	.	.	.	.

Academic Units/Programs within Fields of Study

Engineering
Engineering programs (Faculty of Applied Science except Nursing)

Health Sciences
UBCV faculties of Medicine; Pharmaceutical Sciences; Dentistry and School of Kinesiology
UBCO Faculty of Health & Social Develop except Social Work

Humanities
Programs in: Art History, Visual Art and Theory; Asian Studies; Central, Eastern, and Northern European Studies; Classical, Near Eastern and Religious Studies; English; French, Hispanic, and Italian Studies; Philosophy; History; African Studies; Arts Studies; Creative Writing; First Nations and Endangered Languages; Library, Archival and Information Studies; Linguistics; Medieval Studies; Theatre and Film; Art History; Creative and Critical Studies; German; Japanese; World literature

Sciences
Faculties of Science (UBCO & UBCV), Land and Food Systems and Forestry

Social Sciences
Faculty of Education except Kinesiology
UBCO Faculty of Management
Programs in: Anthropology; Economics; Geography; Political Science; Psychology; Sociology; Gender, Race, Sexuality and Social Justice; Asian Canadian and Asian Migration Studies; Journalism; Public Policy and Global Affairs; Social Work; Gender and Women's Studies; Indigenous Studies; Commerce; Cultural Studies

Appendix 5 – Integrative approach to evaluation of teaching paper

Moving Towards an Integrative Approach to the Evaluation of Teaching at UBC

A Report prepared for the Senate Working Group

Authors: Tanya Forneris, Brendan D'Souza, Christina Hendricks, Sajni Lacey, Jackie Stewart

Date: 27th October, 2021

Background and Executive Summary

UBC highly values teaching and providing high-quality education. As such, one of the goals outlined in UBC's Strategic plan is to "Inspire and enable students through excellence in transformative teaching, mentoring, advising and the student experience." Thus, evaluation of teaching should be held to the same high-quality standards as other forms of assessment through the use of reliable and valid methods. There have been a number of advancements in how post-secondary institutions approach the evaluation of teaching over the past 10 years. However, it has been a significant period of time since policies related to the evaluation of teaching have been developed or reviewed at UBC, and currently these policies are different across both campuses. The policy at UBC Vancouver was last revised and approved by Senate in May of 2007. An initial policy at UBC Okanagan was adopted into its academic calendar in 2005-06 when the campus opened, but it has not been revised since that time. In recent years, the need to review policies and practices related to the evaluation of teaching has been recognized by various stakeholders within UBC.

In the Spring of 2019, a Student Evaluation of Teaching working group was formed at UBC with representation from both campuses (please see the terms of reference for further details). This working group was tasked with reassessing UBC's approach to the Student Evaluation of Teaching in light of current trends in the field and examining student evaluation data for potential bias. For over a year, the working group consulted extensively with multiple constituencies on both campuses, and presented a final report that was endorsed by both Senates in May of 2020. The report included sixteen recommendations, some of which extended beyond student evaluations of teaching. This paper focuses on two of the recommendations:

> **Recommendation 10**: Units should be supported to adopt a scholarly and integrative approach to evaluation of teaching.
>
> **Recommendation 15**: The Vancouver Senate should review the policy on Student Evaluations of Teaching and consider a broader policy on the evaluation of teaching writ large. The Okanagan Senate should develop a similar policy for the Okanagan campus.

A cross-campus working group, sponsored by Senate committees on both campuses, is currently being struck to begin work on revisions to the Senate policies. The purpose of this discussion paper is to provide this Senate working group with an understanding of the state of the field on using an integrative approach to the evaluation of teaching with a view towards the development of a broader UBC policy on teaching evaluation. The paper is composed of four sections. The first section focuses on providing an overview of an integrative approach to the evaluation of teaching. Specifically, it discusses how an integrative approach moves beyond just the collection of multiple sources of data by intentionally integrating numerous types and sources of data for a comprehensive interpretation. The second section provides an overview of how other institutions have moved toward an integrative approach to evaluation of teaching. This overview is based on discussions across multiple interviews with a number of people from a variety of institutions outside of UBC. Included in this section are examples of frameworks

developed and/or adapted by other institutions as well as descriptions of how institutions have worked to implement these frameworks. The key take away was that implementation has involved significant on-the-ground work with academic units over time to shift the culture and/or implement new practices with specific tools, templates and protocols that were meaningful and effective for each unit yet supported the high-level integrative framework of the institution. The third section of the paper provides insight into the current state of teaching evaluation practices at UBC, based on focus group discussions. The focus groups revealed that many units across UBC have practices in place that gather multiple sources of data for evaluating teaching. However, these practices vary significantly across units and a major concern is the emphasis or overreliance on the quantitative data from student evaluations of teaching. Many expressed that this overreliance is partly due to the workload involved in evaluating teaching and this work not being viewed as valuable or as "counting" within merit and/or tenure and promotion processes. The fourth and final section of the report outlines a number of outcome-oriented and process-oriented recommendations. These recommendations are meant to focus discussions related to priorities and actions to support academic units in adopting a scholarly and integrative approach to evaluating teaching as well as the development of a new cross-campus policy on the evaluation of teaching writ large.

Overview of Integrative Evaluation of Teaching

Teaching evaluations should be based on a multisource feedback model that stimulates reflection, is linked to faculty development programs, is transparent about purpose and execution, and is connected in part, to building a climate that fosters excellence in teaching and learning amongst all instructors. There are two main types of evaluation which are often applied to the evaluation of teaching in post-secondary institutions. Formative evaluation refers to processes that use timely feedback to allow for adjustments and progressive betterment of teaching skills and knowledge while summative evaluation is used to assess overarching teaching effectiveness, usually at the end of a formal period of evaluation (Eberly, Center, n.d.).

Teaching evaluations comprised of multiple sources of information such as student evaluations individual reflections and evidence, and peer and/or administrative perspectives is best practice. Specific examples of data include but are not limited to: Student ratings, classroom observations (by peers or administrators), self-evaluation, videos, student interviews, alumni ratings and feedback, employer ratings and reviews, teaching awards, learning outcome measures, teaching portfolios and rubrics with behaviourally-anchored rating scales. Ideally, there are both summative and formative evaluation processes that include both quantitative and qualitative data. Evaluation criteria should be carefully selected to match the purpose of the teaching evaluation (e.g., for tenure and promotion, professional development, mentorship, etc.) through the mapping of a plan within the faculty or department (Berk, 2005; Berk, 2018; Boerboom, et al., 2011; Hornstein, 2017; Lohman, 2021; Shao et al., 2007; Campbell et al., 2010).

An integrative approach moves beyond just the collection of multiple sources of data. It involves bringing together and integrating all the sources of evidence collected as part of the evaluation of one's teaching, including formative and summative as well as qualitative and quantitative for interpretation. One may look to the field of mixed-methods research where quantitative and qualitative forms of evidence are collected and analyzed and then integrated or converged for an overall interpretation and understanding of the phenomenon of interest (Creswell, 2005). There are several advantages of integrating data from different sources, such as being able to use one source or type of data to explain or expand upon the findings of another source or type. Within the field of mixed-methods research several designs exist that could inform future work on an integrative approach to the evaluation of teaching. For example, some designs integrate at the methods level where data from one method of data collection informs another, or two methods of data collection are planned to be merged together for interpretation. In the evaluation of teaching, examples include having an instructor reflect on their end-of-course student surveys using the same platform (e.g., once student surveys are collected the instructor is prompted to log in and provide reflective responses to those provided by the students), or the sharing of a teaching dossier to guide the peer review process. Other mixed-method designs have methods and data collection quite separate and then only integrate at the interpretation and reporting stages, either through a data conversion process or a narrative or visual integration (Fetters et al., 2013). In the evaluation of teaching this could mean having instructors and/or heads develop a narrative or portfolio that speaks to the various sources of evidence and integrates them through an institutionally-developed framework. Another possible approach would be for UBC to develop a system guided by a framework that facilitates the integration of the various sources (e.g., an interactive dashboard that permits one to bring together the quantitative and qualitative data from student surveys, formative and summative peer reviews as well as personal reflections).

In sum, working towards adopting an integrative approach to the evaluation of teaching begins with the adoption of a holistic system that includes multiple sources of data. Once the sources of data have been decided, work is needed to develop a framework that facilitates the integration of these multiple sources in a meaningful and comprehensive manner. The following section provides insight into how other institutions have adopted and implemented a more integrative approach to the evaluation of teaching that could be helpful in guiding change UBC regarding the evaluation of teaching writ large.

Overview of Integrative Evaluation of Teaching Practices Elsewhere

During the summer of 2021, several meetings were held with other institutions who have either adopted or have made considerable progress in the adoption of an integrative approach to the evaluation of teaching. These institutions included University of Colorado Boulder, University of Kansas and University of Massachusetts Amherst (all three are part of the large TEval project focused on this work in the US), as well as the University of Oregon and Simon Fraser University

who have also independently undertaken work in this area. There were a number of common themes that emerged from these meetings.

First, all of the institutions had adopted an approach using the same three sources of evidence.

> Student voice in the form of end-of-term student evaluation surveys.
>
> Peer voice from some form of peer review of teaching (PRT).
>
> Instructor voice, typically in the form of personal reflection through a teaching philosophy statement, a dossier and/or specific reflections on a course-by-course basis in response to the end-of-term student evaluations.

Second, all of the institutions emphasized the value of having a high-level multidimensional framework that clearly outlines expectations in terms of teaching effectiveness and the incorporation of multiple sources of evidence (e.g., Benchmarks Framework from University of Kansas and the Teaching Quality Framework (TQF) from the University of Colorado Boulder – See Appendix A for more resources). These institutions noted that a first critical step is defining what teaching excellence is within the institution, and some spoke at length about how this definition was grounded in the institution's values and/or principles. The challenge faced by many of the institutions was how to integrate the three sources of evidence into something useable by the various individuals who needed to use the evaluation for decision-making (e.g., instructors, unit heads and/or promotion and tenure committees). It was also clear that each institution had worked to either develop or adapt a framework to suit their own context (campuses), particularly on how to integrate the various sources of evidence. The work to develop or adapt a framework across the various institutions was largely informed by the five principles outlined by (Weaver, et al., 2020) in the TEval project.

> Principle # 1: Evaluation includes multiple dimensions of teaching (e.g., activities that capture teaching in its totality, including aspects inside and outside the classroom).
>
> Principle # 2: Evaluation includes multiple lenses (e.g., multiple sources and types of data such as various forms of faculty self-report, peer input and student voice).
>
> Principle # 3: Evaluation involves triangulation of data - no measure should be used in isolation.
>
> Principle # 4: Both formative and summative uses of the data are needed to maximize the impact on teaching effectiveness.
>
> Principle # 5: There must be a balance between uniformity across departments and customization to maximize usefulness at the institutional level.

Third, equally noted was the importance of setting up supports and resources via the institution's teaching and learning centre and/or the Provost Office. For example, small teams composed of staff, teaching fellows and/or post-doctoral fellows in teaching and learning. These small teams then work closely with individual academic units to develop and implement practical and efficient tools, protocols, and strategies that could be adapted to the needs of the unit but still held true to

the framework the institution had developed (See Appendix A for examples of tools from the various institutions listed above). Once the framework was developed and adopted, work with each individual academic unit would start (e.g., 2-3 units at a time). As mentioned above, the work with academic units focused on creating and piloting tools, templates and protocols for instructor reflections, portfolio development as well as peer review processes that would work for their specific disciplines/contexts. In addition, support was often provided to heads of the academic unit to help ensure that the processes they implemented adequately reflected the high-level framework or policy.

Fourth, although these institutions have all taken different approaches due to their specific contexts while working on adopting a more integrative approach to the evaluation of teaching, they all discussed the importance of parallel work on high-level policy and on-the-ground change support. For some institutions, a policy that reflected an integrative approach with multiple sources of evidence had been in place for a significant period of time, yet the practices in the evaluation of teaching did not reflect this policy. Thus, work was initiated by those involved in the institutions' centres of teaching and learning to support academic units in evolving their practices to better align with the policy. Other institutions had yet to or were in the process of developing and implementing new policy or university agreements, alongside work to change teaching evaluation practices at the academic unit level.

Fifth, it was also noted by these various institutions that significant human and financial resources were needed to shift the culture around the evaluation of teaching to an integrative approach. Thus, careful consideration is needed of how work on policy as well as on how to change practices and processes on the ground with academic units can happen concurrently. Many noted that they had advocated within their institutions to support bringing on board faculty champions who received teaching reduction and recognition for this work and/or funded post-doctoral fellowships in teaching and learning. These individuals often formed small working groups that facilitated the "on-the-ground work" with the individual academic units. As outlined above, institutions shared that a successful approach in their experience is working alongside 2-3 academic units at a time to help shift the culture around the evaluation of teaching and implement newly created or adapted tools, templates and protocols. Thus, this can take significant time.

Finally, these institutions also noted that they struggled with the fact that policies are needed to reflect an integrative approach, but since these are inevitably linked to promotion and/or tenure, this can also inhibit the adoption or embracing of a culture shift that is truly about the advancement of high-quality teaching within the post-secondary environment. On the ground, the goal is to have individuals and units engage with the process intrinsically to improve one's experience and confidence with teaching. In reality, there are limits to this without a policy and there is a fine balance to be addressed of having policy that helps drive a culture shift without being perceived as a heavy-handed, top-down, or stress-inducing process.

It is believed that the themes identified above will be informative and helpful as UBC embarks on work to action the two recommendations endorsed by Senate on developing and implementing

an integrative approach to evaluating teaching. However, equally valuable in this process is an understanding of the current practices within UBC, which are summarized in the next section.

Summary of Teaching Evaluation Practices at UBC: The Current State

UBC policies and guidelines

Summative evaluation of teaching at UBC is governed by the Collective Agreement (CA) between the University and the Faculty Association, with the *Senior Appointments Committee (SAC) Guide to Reappointment, Promotion and Tenure* providing more specific guidance within the broader Collective Agreement framework. Teaching evaluation is an essential aspect in the process of promotion and tenure in the tenure-track streams (CA Part 4, Sections 3.04-3.09), and demonstration of excellence in teaching is required for reappointment for lecturers (CA Part 4, Section 2.02). In addition, the teaching performance of sessional lecturers is to be evaluated on a "regular basis" (Part 7, Section 8.01).

The Collective Agreement Part 4, Section 4.02 lays out a list of criteria on which judgments of teaching effectiveness shall be based:

> Evaluation of teaching shall be based on the effectiveness rather than the popularity of the faculty member, as indicated by command over subject matter, familiarity with recent developments in the field, preparedness, presentation, accessibility to students and influence on the intellectual and scholarly development of students.

Those reviewing candidates for tenure and promotion are asked to do so in light of these requirements. In the same section, the CA also lists possible types of evidence that could be used for evaluation of teaching, though without requiring any source specifically:

The methods of teaching evaluation may vary; they may include student opinion, assessment by colleagues of performance in university lectures, outside references concerning teaching at other institutions, course material and examinations, the caliber of supervised essays and theses, and other relevant considerations. When the opinions of students or of colleagues are sought, this shall be done through formal procedures. Consideration shall be given to the ability and willingness of the candidate to teach a range of subject matter and at various levels of instruction.

The *SAC Guide* provides more detailed suggestions on sources of evidence for summative evaluations of teaching:

> The methods of teaching evaluation may vary in face-to-face, online and blended formats, but will normally include Student Evaluations of Teaching (SEoT – UBCV) or scores from the Teaching Evaluation Questionnaire (TEQ – UBCO) and a Summative Peer Review of Teaching. The summative review will normally be based on an examination of the following: quantitative Student Evaluations of Teaching (SEoT) – the University module questions, and in particular Q6 (UBCV) or Q20 (UBCO), with comparative Departmental/Faculty norms; qualitative comments from SEoTs about classroom teaching practices; the candidate's course materials, assignments and grading practices; the caliber of supervised essays and theses; peer reviews of teaching; and other relevant considerations. (Section 3.2.4)

Appendix 2 of the *SAC Guide* notes that a summative review of teaching should be included when a candidate's file is considered by the Senior Appointments Committee, usually written by the Head or Director, or the Chair of a summative peer review of teaching committee in the unit. Data sources that should be summarized in this report, according to the SAC Guide, include: student experience of instruction results, peer review of teaching reports and highlights from them, contributions to graduate or professional training, contributions to educational leadership (required for educational leadership faculty), and a summary of other qualitative evidence of the candidate's teaching effectiveness (such as professional development undertaken, awards or other recognition for teaching). This summative assessment of teaching could be a place to integrate these various sources of evidence, as well as summarize them, though the SAC Guide does not provide guidance on how this might be accomplished. It simply lists which kinds of evidence should be included and summarized in the report.

Notably, there is particular emphasis in the *SAC Guide* on student evaluations of teaching scores, and a limited subset of them at that. Appendix 2 of the *SAC Guide* states that the summative review of teaching report should include a table of scores from student evaluations of teaching focusing on questions about "overall effectiveness" (Q6 at UBCV, Q20 at UBCO). Scores from additional questions could also be included if they "provide particularly useful evidence about the candidate's teaching record" (*SAC Guide,* Appendix 2). A sample of student comments from the end-of-course surveys could also be included (optionally) if they are selected by the person writing the summative report, rather than by the candidate. This emphasis on student evaluations of teaching scores in evaluating teaching, particularly on one number, is a source of concern for many across campus, as noted below.

Peer review of teaching practices (PRT), both formative and summative, are governed by policies and procedures at the Faculty or unit level. Examples from some Faculties who have agreed to share are posted on the Summative Peer Review of Teaching section of the Centre for Teaching, Learning and Technology website at UBCV. A few other examples of Faculty-level guidelines were shared with us in support of writing this paper. From reviewing these documents we found that summative PRT practices vary across the institution, including differences in number of reviewers and whether any must be from outside the unit, number of classes visited, number of meetings with the candidate (before and/or after the class visit, or not at all), whether the peer review of teaching report is shared with the candidate or not, and more. This

variation may be due to differing approaches to teaching, and criteria for evaluating such approaches, between disciplines and contexts.

Still, amongst the units whose PRT practices were reviewed, many adhere to a set of Principles of Summative Peer Review put together by a UBCV working group on peer review of teaching, including: having more than one reviewer; using a set of clearly-defined criteria consistent across a Faculty, program, or unit; and paying attention not only to class visits but to other aspects of teaching such as course materials, course design, use of learning technology as appropriate.

Focus group discussions

During the summer of 2021, several focus groups were held with individuals from UBCO and UBCV, including Associate Deans of some Faculties and faculty members who have served as peer reviewers, to gather information on what they felt is working well or could use improvement in teaching evaluation practices. However, not all Faculties or units on both campuses were represented, and thus this section should not be taken to be a comprehensive review of teaching evaluation practices at the institution. Instead, it is meant to provide an overview of some of these practices as well as perceived challenges, as a way to contextualize the recommendations made later in this paper.

There was general consensus in the focus groups that multiple data sources should be used for teaching evaluation, and many Faculties and units do so by including student end-of-course surveys, peer reviews of teaching, reflective summaries of teaching practices by faculty members, sample teaching materials, and other evidence in teaching dossiers as part of summative teaching evaluation. One challenge that emerged in discussion, though, is that while abridged teaching dossiers for educational leadership stream faculty may be sent forward to the Senior Appointments Committee, this is not the case for faculty in the research and teaching stream (see the *SAC Guide* Appendix 2). It is not clear why there should be this difference since teaching quality is an important part of evaluations for promotion and tenure for both faculty streams. Though the Collective Agreement requires that faculty reach different levels of teaching quality in order to be promoted to a higher rank (e.g., promotion to Associate Professor requires “successful” teaching, while promotion to Associate Professor of Teaching requires “excellence” in teaching), this does not mean there should be a difference in the type of evidence provided or considered at the level of the Senior Appointments Committee.

Another concern expressed by some focus group participants is that there tends to be too much reliance on quantitative results from the student experience of instruction (SEI) surveys in summative teaching evaluation for reappointment, tenure and promotion processes, particularly on the single number from the question about overall quality of teaching (as suggested in the *SAC Guide*, quoted above). This may be in part because the quantitative data is relatively

simple, easy to scan and understand quickly, and easy to use for comparisons across courses or time periods.

Some focus group participants also pointed out that this overreliance on quantitative SEI results is likely because summative peer review of teaching reports tend to be mostly or wholly positive. This may be because they are so high stakes that including criticism is viewed as potentially jeopardizing a case for tenure and/or promotion. However, if there are few to no critical comments or constructive suggestions, these reports may not provide a great deal of information as components of *evaluating* teaching, and it is easy to fall back on SEI results because they seem to provide clearer ways to differentiate amongst levels of teaching quality.

Over the past few years, a group of faculty and staff from multiple faculties and units at UBC Vancouver created a summative peer review of teaching rubric that was meant to, among other things, try to address the issue of summative PRT reports being nearly uniformly positive. The rubric includes seven levels, many of them tied to descriptors in the faculty Collective Agreement, with sample descriptors of the levels and examples of the kinds of practices an educator at that level might exhibit. The hope was to show that not everyone needs to be at the very top level, and that very good teaching could be at somewhat "lower" levels and still be both high-quality enough to fulfill the criteria in the Collective Agreement and yet include possible room for improvement. The rubric is open to any unit in the institution to revise and use as they wish.

Another theme that emerged in relation to PRT was that it, and practices of evaluating teaching more broadly, seem to be mostly focused on tenure and promotion processes, rather than on improvement of teaching at various stages in one's career. Several focus group participants noted that there is not as much emphasis placed on evaluation of teaching post-tenure or promotion. One suggestion was to consider instituting more formative peer reviews of teaching where feasible, from early on in one's career (while teaching habits are being formed) to every few years for all faculty, even after tenure. Another suggestion was to do more to celebrate and promote excellent teaching within units as something all faculty should be striving for, such as through regular faculty-led sessions devoted to sharing ideas and good practices with their colleagues.

Focus group participants also discussed, however, that PRT takes a great deal of time, so instituting more formative PRT in addition to summative is challenging, particularly in smaller faculties or units with fewer peer reviewers available. This work needs to be resourced, including training for reviewers. Another challenge is with recognizing/rewarding peer review activities: given the amount of time and effort it takes to do well, doing peer reviews should be recognized as a significant part of one's service work. One participant in the focus groups noted that in their unit if someone is the PRT representative for their unit and doing quite a few PRTs then they are provided a course release.

In summary, a number of units already include multiple sources of data when evaluating teaching, and the *SAC Guide* instructs heads of units to do so in summative reports on teaching. Student experience of instruction (SEI) questionnaires, peer observations, and teaching dossiers are standard practices to varying degrees. However, the extent to which the various forms of evidence are brought together in an integrative fashion is not entirely clear, and an overreliance on quantitative SEI scores is a significant concern. In addition, there are a variety of practices of peer review of teaching across the institution, but no concerns about this variation were raised amongst the focus group participants, and we do not draw any conclusions about it here. A number of challenges with practices of teaching evaluation, including the workload involved, were noted amongst focus group participants and warrant further investigation and discussion.

Recommendations for an Integrative Approach for Evaluation of Teaching at UBC

This section outlines both outcome-focused and process-focused recommendations. It is hoped that the outcome-focused recommendations can help guide the "what is needed" discussions around changes to the evaluation of teaching writ large at UBC while the process-focused recommendations help guide discussions on "how" these outcome-focused recommendations may be implemented and/or achieved effectively.

Outcome-Focused Recommendations

- As a first step in developing an integrative approach to the evaluation of teaching, UBC needs to establish a working definition of teaching effectiveness to define what teaching effectiveness is within our own context or institution. Establishing such a definition was recognized as a necessary first step by all institutions that we met with. The process involved in establishing such a definition was best exemplified by the University of Oregon and the University of Massachusetts (Amherst). The University of Oregon established a definition of "teaching quality" within the context of the values of the university. These values were agreed upon by various stakeholders including the Faculty Union. In the case of the University of Massachusetts (Amherst), the working group that was tasked with developing a "multi-faceted approach" to teaching evaluation established a definition of "teaching quality" based on the views of different departments on teaching quality as well as on "emerging" definitions of quality from the literature. This in turn led to establishment of aspects/dimensions of teaching that can be evaluated and adopted university-wide with individual departments having autonomy over defining different levels of achievement (developing, proficient and expert) for each aspect/dimension of teaching.

- UBC also needs to develop a high-level framework that clearly outlines what constitutes an integrative approach to the evaluation of teaching at UBC. This framework should be grounded in the values, principles, and definition (discussed in the above

recommendation). Based on reviewing frameworks developed and adopted by other institutions it should clearly identify the different aspects/dimensions of teaching being evaluated, the sources (multiple) of evidence used to evaluate each dimension, the extent of achievement of the dimension of teaching and how these are to be integrated. Finally, having this framework reflected in the new policy would be valuable as it would foster consistency in the adoption of an integrative approach across units while recognizing that the specific tools, templates and/or protocols adopted by individual units can and should be adaptable to meet the needs of different disciplines and contexts.

Process-Focused Recommendations

- To adopt an integrative approach, UBC should establish a centralized system with personnel trained to support individual academic units or faculty members with transitioning to an integrative approach to the evaluation of teaching. This work will require a multi-year commitment and change management process and cannot be downloaded to individual units or faculty members without such centralized supports. As outlined above, other institutions engaged in these change processes have had success with smaller working groups composed of staff from their centres of teaching working with faculty teaching fellows with teaching release and/or post-doctoral fellowships in teaching and learning who work progressively with the academic units (2-3 units at a time) to identify, develop and/or adapt a repertoire of tools that can be used to collect multiple types of data across the institution to support the change process.

- To effectively sustain an integrative approach to the evaluation of teaching, there is a need to recognize the adoption of these practices as an important and valued part of faculty workload. As outlined above, units have been successful in implementing both formative and summative peer review of teaching when that work is recognized as valued service contributions, or considered in teaching workloads, teaching award criteria and/or merit processes.

- Those working on policy should connect regularly with those that will be working on the ground to supporting the academic units and instructors with this change. One option would be to have representation from the CTL and CTLT from both campuses as members of this Senate-endorsed working group. Inclusion of such roles would allow for the higher-level policy and framework development to work in tandem with on-the-ground implementation and adoption of new practices and tools designed to collect and integrate multiple sources of data.

- Careful attention is needed on how policy implementation and on-the-ground work can nurture a shift away from an anxiety, stress and/or remiss culture to one that fosters a real aspiration and support for excellence in teaching and learning at UBC. Fostering culture change throughout the process may be best accomplished by engaging and

empowering instructors to contribute to the development of the new processes and frameworks. On-the-ground support from units such as the CTL, CTLT and/or teaching fellows could strengthen this cultural shift. The institutions consulted to date shared that it was on-the-ground support that often-helped instructors feel supported, capable, and invested in change practices around the evaluation of teaching.

- Finally, it is recognized that this discussion paper serves as an initial foundation for this work. Further engagement with the university community on both campuses is needed to provide more comprehensive information about current teaching evaluation practices within units, including current challenges and successful practices. Regular engagement and consultation with faculty, students, staff, and academic leaders throughout the process of developing, adopting, and implementing an integrative approach to the evaluation of teaching will be critical.

References

Benton, S. L., & Young, S. (2018). *Best practices in the evaluation of teaching: IDEA paper #69*. IDEA Center, Inc. https://eric.ed.gov/?id=ED588352

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International journal of teaching and learning in higher education, 17*(1), 48-62. https://www.isetl.org/ijtlhe/pdf/IJTLHE8.pdf

Berk, R. A. (2018). Start spreading the news: Use multiple sources of evidence to evaluate teaching. *The Journal of Faculty Development, 32*(1), 73-81. https://www.schreyerinstitute.psu.edu/pdf/UseMultipleSourcesSRs_Berk_JFacDev1-11-2018.pdf

Boerboom, T. B., Jaarsma, D., Dolmans, D. H., Scherpbier, A. J., Mastenbroek, N. J., & Van Beukelen, P. (2011). Peer group reflection helps clinical teachers to critically reflect on their teaching. *Medical teacher, 33*(11), e615-e623. https://doi.org/10.3109/0142159X.2011.610840

Campbell, N., Wozniak, H., Philip, R. L., & Damarell, R. A. (2019). Peer-supported faculty development and workplace teaching: An integrative review. *Medical Education, 53*(10), 978-988. https://doi.org/10.1111/medu.13896

Creswell, J. W., & Creswell, J. D. (2005). Mixed methods research: Developments, debates, and dilemmas. In R.A. Swanson and E.F. Halton III (Eds.), *Research in organizations: Foundations and methods of inquiry*, (pp. 315-326). Berrett-Koehler Publishers.

Eberly Center, Carnegie Mellon University. (n.d.). What is the difference between formative and summative assessment? https://www.cmu.edu/teaching/assessment/basics/formative-summative.html

Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs—principles and practices. *Health services research, 48*(6pt2), 2134-2156. https://onlinelibrary.wiley.com/doi/10.1111/1475-6773.12117

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education, 4*(1), 1304016. https://doi.org/10.1080/2331186X.2017.1304016

Lohman, L. (2021). Evaluation of university teaching as sound performance appraisal. *Studies in Educational Evaluation, 70,* 1-11. https://doi.org/10.1016/j.stueduc.2021.101008

Shao, L. P., Anderson, L. P., & Newsome, M. (2007). Evaluating teaching effectiveness: Where we are and where we should be. *Assessment & Evaluation in Higher Education, 32*(3), 355-371. https://doi.org/10.1080/02602930600801886

Weaver, G. C., Austin, A. E., Greenhoot, A. F., & Finkelstein, N. D. (2020). Establishing a better approach for evaluating teaching: The TEval Project. *Change: The Magazine of Higher Learning, 52*(3), 25-31. https://go.exlibris.link/W5K1YfF4

Appendix 5A - Additional Resources

The TEval Project (Transforming Higher Education – Multidimensional Evaluation of Teaching)

The TEval project is a multi-institutional initiative that works to advance how teaching is evaluated within post-secondary institutions. Below are three links to provide further context and examples of work as many of the institutions met with in the writing of this paper are part of this larger project.

- Overview of the TEval project: https://teval.net/index.html
- Weaver, G. C., Austin, A. E., Greenhoot, A. F., & Finkelstein, N. D. (2020). Establishing a better approach for evaluating teaching: The TEval Project. *Change: The Magazine of Higher Learning, 52*(3), 25-31. UBC Permalink: https://go.exlibris.link/W5K1YfF4
- Examples of Frameworks, Rubrics & Tools: https://teval.net/resources.html

Below are further examples of institutions working under the larger TEval project and the frameworks, rubrics, tools, and/or processes developed and implemented.

- **University of Kansas -** Framework, Rubric & Tools developed by the KU Center for Teaching Excellence https://cte.ku.edu/benchmarks-teaching-effectiveness-project
- **University of Colorado Boulder -** Framework, Rubrics & Tools: https://www.colorado.edu/teaching-quality-framework/resources
- **University of Massachusetts Amherst -** Summary of the work at UMass regarding the process of adopting and implementing changes to transforming how teaching is evaluated: http://www.umass.edu/oapa/program-assessment/instructional-innovation-assessment/evaluation-teaching-new-approach

University of Oregon

The University of Oregon has also embarked on this work but the work has been a joint project between the Provost's office and University Senate.

- Background: https://provost.uoregon.edu/revising-uos-teaching-evaluations
- Definition and Principles of Teaching Excellence: U of O Principles of Teaching Excellence
- Framework & Resources: https://teaching.uoregon.edu/resources/teaching-evaluation

Simon Fraser University

SFU has also initiated work to develop and implement a multi-dimensional teaching assessment and the information and resources are available here: https://www.sfu.ca/cee/services/assessment.html

Appendix 6 - Report on investigation of options for automated text analysis

Automated analysis of SEI text comments: Report on options reviewed

Submitted by the SEI Implementation Committee as part of the final report to Senates on the SEI Implementation Project

September 2022

Executive Summary

In May 2020 a UBC Student Evaluations of Teaching working group submitted a report to both Senates with multiple recommendations related to what were then called Student Evaluations of Teaching. Recommendation 12 in that report was to engage in a pilot project to investigate the possibility of automated analyses of open text comments from these surveys:

> Many faculty members report the free-text student comments as sources of rich data to support reflection and enhancement of their course and teaching. It is recommended that a pilot investigation be undertaken, with one or more Faculties, to investigate the potential of automated approaches to extract useful information from large volumes of text submissions. The pilot should engage with appropriate research expertise in Faculties in these areas, and aim initially for formative purposes. (p. 6)

A small project team, made up of members of the Student Experience of Instruction (SEI) Implementation Committee, has reviewed four options for automated processing of open text comments, which are detailed in this report. They are:

1. A natural language processing application developed by faculty and students in UBCV Computer Science (CS).
2. An Arts Instructional Support and Information Technology (ISIT) pilot in 2018-2019 using machine learning to extract suggestions from text comments (UBCV).
3. Blue Text Analytics (BTA): an add-on product within Blue, UBC's current SEI software, which is part of UBC's current license with Explorance (the software company that created Blue).
4. Blue Machine Learning (BlueML), a standalone product from Explorance that currently has no direct integration option with Blue.

This report provides an overview of the functionality of these systems and recommendations for possible next steps. The report also notes that there is significant interest at the institution in finding methods to locate and then remove discriminatory, abusive, or otherwise harmful comments before faculty members access the set of comments. The committee has not found a straightforward method for doing this yet; some of the work in this area seems to be in relatively early stages.

Background and Context

Prior to the implementation of the new University Module Items in Fall 2021, there were various open-ended questions asked on the surveys across the two campuses. The new UMI in SEI surveys on both campuses include Likert-style questions (i.e. "closed questions") as well as three common, open questions that invite students to write free text comments:

- Do you have any suggestions for what the instructor could have done differently to further support your learning?
- Please identify what you consider to be the strengths of this course.
- Please provide suggestions on how this course might be improved.

Individual SEI reports available to instructors include statistics for the quantitative questions (interpolated median, dispersion index, and percent favorable), as well as a list of all text responses. Such comments can be sources of in-depth information about students' experiences in courses that, as noted above, can inform formative reflection and possibly inspire changes in teaching. However, in some cases these comments can be quite extensive, making it challenging to discern patterns simply by reading through them. It is also important to recognize that the comments sometimes include harmful and abusive language, including racist, sexist, ableist and other discriminatory comments.

In the summer of 2021, the Implementation Committee formed a small project team to begin investigating different options for implementing Recommendation 12, as discussed above, to investigate automated systems for summarizing themes from text comments for instructors to use for formative purposes. We reviewed four systems, discussed in this report.

There are multiple tools for undertaking various aspects of natural language processing (NLP), such as tokenization (breaking a text up into sentences, words, symbols, etc. called "tokens"), part-of-speech tagging (tagging words as, e.g., noun, pronoun, verb, adverb, etc.), topic analysis (putting phrases or sentences into topics that group similar ideas together), sentiment analysis (tagging phrases or sentences with a polarity, such as positive, negative, neutral), and more. Many of these either are stand-alone tools, or collected into packages to be used with languages such as Python or R.

The Implementation Committee has not reviewed such options, but has focused on platforms that bring these functions together into a system that could be used by individual faculty members to review analyses of their own student comments data, such as through a dashboard or a report.

Systems investigated

1. Natural language processing system developed in Computer Science, UBC Vancouver

Raymond Ng, Giuseppe Carenini and colleagues in Computer Science and the Natural Language Processing Group (NLP) at UBC Vancouver have developed an NLP application that extracts themes from text data and performs binary sentiment analysis (positive or negative).

Review of functionality

One can either begin with a pre-defined list of themes or the application can generate them from raw SEI comments data to create a lexicon. The lexicon can be refined manually to ensure that the system is picking up on meaningful themes for the data and its purpose. Categories of similar themes can also be created. Then, data is run through the system using the refined lexicon and categories to generate information that users can interact with on a dashboard.

The user dashboard provides multiple options for parsing and viewing the data, including viewing by theme, multiple themes in a category, positive and negative sentiments in comments

by theme, filtering by year or course, filtering by SEI question, comparing across years, and more. The data can be viewed in tables or visualized in charts.

In the fall of 2021, the Implementation Committee worked with the team in Computer Science to pilot test the system on text comments from SEI surveys, with volunteers who agreed to have their results used for this purpose. A focus group of faculty and staff met to review the system using the pilot SEI data and discuss the feasibility for individual faculty members to possibly use it for formative purposes.

Feedback from the focus group was overwhelmingly positive, with significant interest in continuing to investigate this system. Participants appreciated how the system encourages focus on both positive comments as well as those that are attached to negative sentiments, since it is quite easy to focus mostly on the negative ones otherwise. They also appreciated how using a system like this can provide a better summary of trends and outliers in a large body of comments, instead of faculty members having to manually review all comments to gauge the general themes and sentiments.

The focus group was interested in discussing whether the system could be used to find and remove harmful and abusive comments. The answer is that it may be possible in future to include functionality in the system that could locate at least some of the harmful comments, though tools to automatically recognize such comments are in nascent stages, and review by people of comments that a system might tag as potentially harmful should always be done. Removing them before faculty members access them would not be automatic as this system is standalone and not integrated with any other systems at UBC.

Possible next steps

The pilot done so far was very small, and a next step could be to do a larger pilot, such as with an entire department or program. Further items that might be investigated in such a pilot could include: developing and testing a way for individual instructors to access the dashboard (in the earlier pilot the CS team uploaded the data to the system and displayed it for others as “view only”); developing and testing the ability to edit sentiments as well as the lexicon (not yet possible in the system); and developing an approach that might help to flag harmful comments (also not yet possible in the system).

2. Arts ISIT – pilot work undertaken in the Faculty of Arts, UBC Vancouver

The UBCV Faculty of Arts Instructional Support and Instructional Technology (Arts ISIT) team conducted a pilot in 2018/19 using machine learning algorithms to extract suggestions from student comments to support course improvement.

They created the algorithm by manually coding a set of comments as either containing an explicit suggestion or not, then analyzing the linguistic features of the comments with explicit suggestions to create a set of grammar rules. They then trained the machine learning model with a training data set and refined it by comparing with human coders.

Review of functionality

The SEI Implementation Committee did not test this system, but received a briefing presentation and written information from Arts ISIT about the pilot.

The machine learning system developed by Arts ISIT can automatically locate and highlight explicit suggestions from student comments. Explicit suggestions refer to comments that provide clear recommendations for changes that are immediately actionable, e.g., “The topics could be explained in more detail, especially important concepts.”

Using the algorithm, Arts ISIT was able to quickly extract students’ explicit suggestions on courses and instructors from large sets of comments. They were able to achieve a high degree of accuracy with the machine learning system as compared with human reviewers.

The team created a dashboard that listed the full set of comments in a box at the top, with the set of explicit suggestions in a box at the bottom. This could provide useful information for instructors to consider specific areas that students felt could use improvement by allowing for easier focus on explicit suggestions out of a larger set of comments.

Possible next steps

This project is an interesting proof of concept that yielded a dashboard that could be helpful for individual faculty. Note that in the new UMI, one of the open-text questions now explicitly asks students for suggestions, so what the algorithm in its current form does (pull out explicit suggestions) may be less needed (though still useful, since there may be explicit suggestions in other comments).

One option could be to expand the work Arts ISIT has done to create a new algorithm with a different purpose. For example, the Arts ISIT team working on this project noted that another step could be to develop an algorithm to map sentiments and aspects.

3. Blue Text Analytics (BTA)

Blue Text Analytics (BTA) is a tool developed and supported by UBC’s SEI survey system vendor, Explorance. PAIR currently has access to BTA and could run reports for individual instructors.

BTA consists of two components – the BTA engine and Explorance’s dictionaries. The BTA engine uses natural language processing methods to categorize comments into themes that are predetermined by the dictionaries. The BTA dictionaries have been created by Explorance and cannot be altered by individual users or institutions. There are currently four dictionaries available for use in analysing students’ feedback:

- Two Teaching and Learning Dictionaries – American English and British English
- Two Sentiments Dictionaries – American English and British English

The Teaching and Learning Dictionaries include three categories:

- **Teaching and Learning Attributes:** This category includes positive, neutral, negative, and ambiguous attributes. For example, “interesting” or “enthusiastic” are usually positive, and will be labeled as positive attributes, while “boring” or “stressful” will be labeled as negative attributes.
- **Elements mentioned**: This category provides an analysis of elements mentioned in feedback comments, such as assessments/grading, feedback, content/materials, lectures.
- **Alerts**: This category focuses on comments that are related to health and safety issues such as mentions of violence or bullying, or discrimination such as racism or sexism.

More information about BTA can be found in the BTA User Guide from Explorance.

Review of functionality

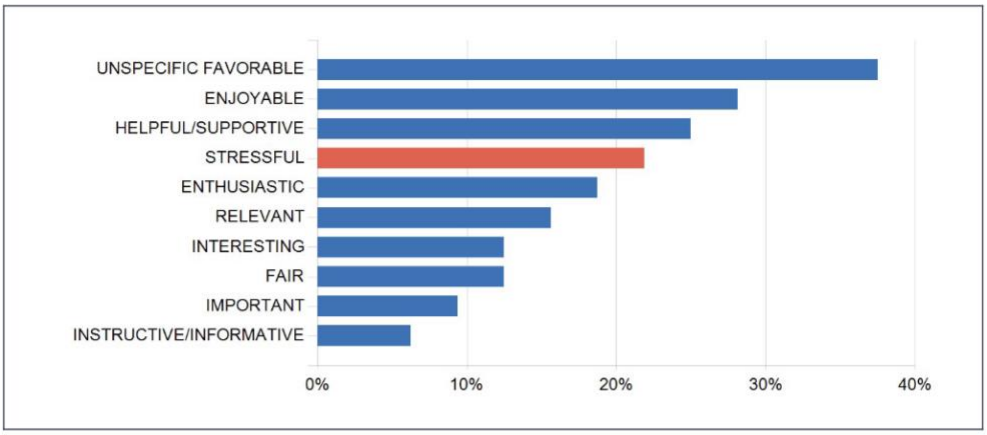
Some members of the SEI team along with the Chair of the Implementation Committee reviewed how the system works, and also viewed reports with SEI data from faculty who consented to have their data used for this purpose.

Because BTA is integrated directly with Blue, there is no need to upload data into the system separately; it can ingest SEI data directly from the system UBC already uses for SEI surveys. This is a significant advantage over other systems reviewed in this report, as it can take a great deal of time to ensure the data is in the right format for the systems before it is uploaded.

BTA analyses could be run by the SEI team in PAIR as part of the SEI reports provided to instructors. These analyses would appear as additional areas in the reports to what is currently provided (statistical data plus a list of text comments).

Below is a screen shot of a report using the “Teaching and Learning attributes” category for a question asking about the strengths of a course. This image shows three different ways of reporting on the same data. Note that “positive” and “negative” attributes are indicated by different colours in the bar chart and table (blue and red, respectively; also, bars in bar charts can be hatched to improve accessibility).

What were the strengths of the course?



Attributes - t&l [No. of comments]	Overall [32]
UNSPECIFIC FAVORABLE	37.50 %
ENJOYABLE	28.13 %
HELPFUL/SUPPORTIVE	25.00 %
STRESSFUL	21.88 %
ENTHUSIASTIC	18.75 %
RELEVANT	15.63 %
INTERESTING	12.50 %
FAIR	12.50 %
IMPORTANT	9.38 %
INSTRUCTIVE/INFORMATIVE	6.25 %

The addition of some of these analyses to individual instructor reports could provide some basic information about trends and patterns that may not be as obvious to instructors by simply scanning the list of text comments, such as being able to notice at a glance that a significant percentage of students made comments related to helpfulness or enjoyableness, or that there were about equal numbers of comments related to stressfulness and helpfulness.

One limitation to the system is that the tables, charts, or word clouds in the BTA reports don't show which comments were labeled with which themes. This is possible by exporting the data into a CSV file, which can only be done by SEI staff rather than individual faculty themselves. In addition, as noted above, the available dictionaries cannot be altered by users or institutions.

The Alerts category within the Teaching and Learning Attributes dictionary can search for themes related to discrimination and harassment, but to do a proper test of this functionality would require a larger dataset, as the small sets of comments we used for testing were not enough to indicate what kinds of comments the Alerts function would flag.[5] Sample keywords that BTA uses to put comments into Alerts are available in the BTA dictionaries documentation from Explorance.

If the Alerts category were to be used, it is vital to also have a clear set of guidelines, roles and responsibilities for reviewing the alerts, determining which need action, and directing the information to the responsible parties or offices to respond.

Possible next steps

One next step could be to do a pilot test of BTA functionality and reports with faculty members from multiple disciplines to gather their feedback on the value of the system for reviewing text comments for formative purposes. From there a decision could be taken as to whether it would be worth implementing the BTA reports into the SEI data reports already made available to instructors. As noted above, since BTA is integrated with Blue it is fairly straightforward to include this information in instructor reports.

4. Blue Machine Learning (BlueML)

Blue Machine Learning (BlueML) is a standalone text comment analysis solution developed by Explorance. It is based on proprietary machine learning algorithms and automatically detects themes and sentiments in qualitative feedback. The tool features a dashboard that allows administrative users to upload a spreadsheet of qualitative data to be analyzed by the BlueML system and then visualize the results in a number of dimensions.

Blue ML has several machine learning models to choose from; in our testing we focused on the Student Learning Categorization model, which groups comments using a large set of topics and categories such as course materials, assessments, lectures, use of technology, and more. These topics and categories are created and updated by the vendor. This model also includes sentiments: positive, negative, neutral, or not explicit.

---

[5] Hum, Wuetherick, and Yang (2021) provide a useful discussion and review of the Alerts function in BTA, as well as other functions. They note that using the Alerts dictionary required a good deal of manual review to address false positives, and there were some important concerning comments the dictionary missed. They found that the Alerts function was particularly useful for identifying comments that could indicate problems with words or actions of the teaching team in classes, or that suggest issues of concern for student safety or wellbeing. Hum, G., Wuetherick B., Jang, Y. (2021). Supporting practical use and understanding of student evaluations of teaching through text analytics design, policies, and practices. In E. Zaitseva, B. Tucker, & E. Santhanam(Eds.). *Analysing Student Feedback in Higher Education: Using Text-Mining to Interpret the Student Voice* (1st ed.). Routledge. https://doi.org/10.4324/9781003138785

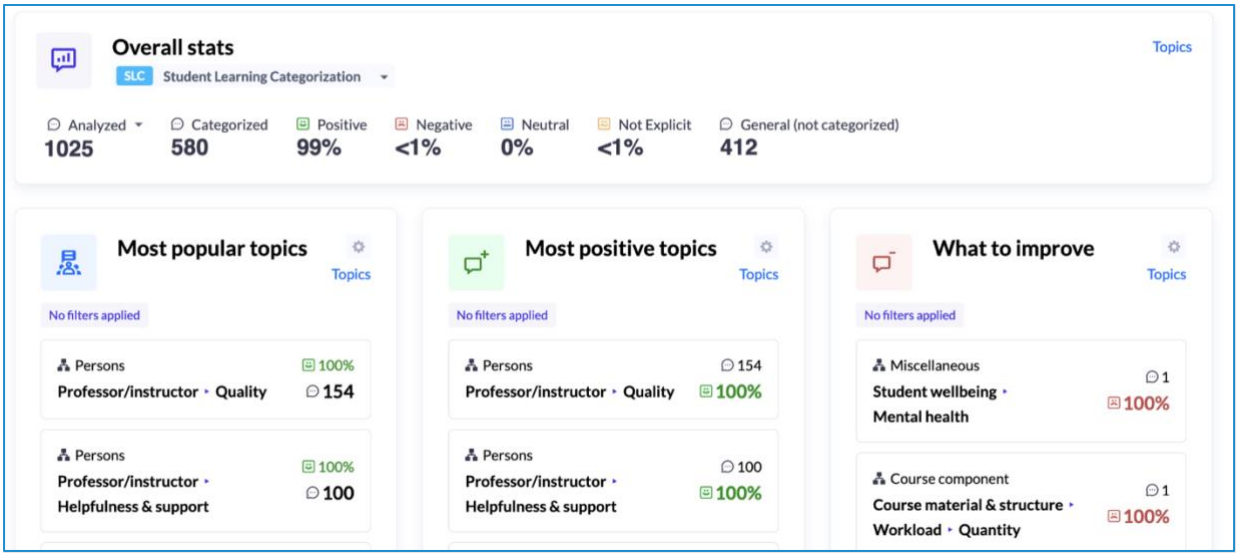
For more information on Blue ML, review [the Blue ML documentation from Explorance.]

Review of functionality

Several members of the Implementation Committee as well as SEI staff in PAIR were provided with sandbox accounts for Blue ML from the vendor in order to test out the platform. Blue ML is not integrated with Blue at this time; each user must upload data directly into the platform in a CSV file. The file must be formatted in a particular way to get useful data from the system, which can take a good deal of time and effort, particularly if this were to be done with large sets of data.

Once data is uploaded and analyzed, the results are presented in a dashboard that administrative users can view—note that Blue ML is not set up for individual instructors to have access to the dashboard. The dashboard includes information such as how many comments are in the data set, the number of comments that were categorized into topics, the percentage that were tagged with positive, negative or neutral sentiments. It also has widgets that focus on the most popular topics (those with the highest number of comments), the most positive (topics with the highest number of positive comments) and what to improve (topics with the highest number of negative comments). Note that single comments can have multiple topics and sentiments assigned.

Below is a sample screen shot of the dashboard.



Pilot testers noted that once the data was uploaded and analyzed, the dashboard provided a clear and helpful at-a-glance breakdown of the distribution of sentiments and which category areas were mentioned most often, which were mentioned with the most positive sentiments, and which areas could use attention for possible improvement. One can click on any of the stats at the top of the dashboard or topics in the widgets below to drill down to find specific comments in those areas, how they were categorized, and sentiments attached to them.

The testers noted some limitations with Blue ML. There were a number of errors in the categorization of some of the comments into topics or sentiments, and the only way to address these currently is to provide manual feedback on each comment with suggestions for changes. These go to a team at Explorance who use them to update the model. We could not find a way to fix the errors within the system itself beyond waiting for an update from the vendor and re-running the data analysis.

In addition, as noted above, the dashboard is not designed for broad access by individual faculty members at this time. Instead, PAIR staff would need to run and export the analyses. The tool provides the option to export the results in an Excel format that includes the question, the comment, the sentiment, and all of the categories to which the comment was attached. The data in this raw format is less digestible and useful for faculty than what is provided by the dashboard.

Finally, Blue ML recently developed an Alerts model that is currently in Beta, that is designed to find comments that mention keywords or topics related to racism, sexism, bullying, harassment, insults, threats, and more. This model is in early stages of development, and was not tested by the committee. A list of topics and keywords the model is meant to locate in comments can be reviewed in the documentation for the Alerts model from Blue ML.

Possible next steps

Since individual instructors would not have access to the dashboard, it's not clear if a broader pilot test of the dashboard functionality with faculty would be useful. Faculty could view the exported data in an Excel file, but that raw data may not be very useful for individual faculty members without a way to easily review the patterns and other information the dashboard provides. A license for Blue ML does include access to an API, and one option could be to investigate whether PAIR might be able to ingest data through the API into a customized reporting dashboard, but that would need to be further investigated.

Summary and possible next steps

The SEI Implementation Committee finishes its work in early Fall, 2022, wrapping up after the final report is presented to both Senates. We suggest below some possible next steps the institution could take.

Pilot testing

One or more of the options above could be further investigated through further pilot testing. For example, a working group could be struck specifically for this purpose; it would be useful to have at least some people on the working group with expertise in the area of natural language processing.

If further pilot testing were to be explored, we recommend focusing on one or both of the following, based on our investigations so far.

- **Computer Science NLP system:** A broader pilot of this system could be useful, perhaps with a full department. This pilot could potentially test some of the new functionality the

focus group suggested, and how the dashboard might be made available to individual faculty members.

- **Blue Text Analytics (BTA):** It could be useful to gather a group of faculty from multiple disciplines to review the types of reports that can be generated with their own data. As discussed above, a fulsome test of the “Alerts” category in the BTA dictionary would be helpful.

Investigating other options

There may be more options available beyond those which the SEI Implementation Committee has investigated so far. This is a quickly-evolving space, and new options are likely to develop in the near future. If a working group is struck to conduct a pilot test of one or more of the systems discussed here, they could also be tasked with investigating other possibilities.

One area that needs further investigation is tagging harmful and abusive comments and potentially being able to remove them before they are shared with faculty members. As noted above, BTA and Blue ML may flag such comments, but further detailed testing is required to better understand the value of these systems for this purpose. There are other options for screening for and possibly removing harmful comments,[6] but at this stage there does not seem to be a straightforward, easy-to-implement way to do so. Further investigation would be useful.

[6] For example, an article published in July 2022 by Cunningham, Laundon, Cathcart, Bashar, and Nayak discusses work at Queensland University of Technology combining machine learning with a dictionary approach to locate and remove harmful comments. This work was built on a foundation of a definition of unacceptable comments that the institution had established through community consultation. A dictionary was then that fit the definition, and that was applied while the survey was live, using functionality in Qualtrics (where their surveys are hosted). This allowed for staff to reach out to individual students to edit comments before the survey closed. Then, a machine learning algorithm was used to review comments after the surveys closed. In both cases, staff reviewed the flagged comments and determined if they fit the definition of unacceptable comments; if so, they were removed before results were shared with faculty members. Cunningham, S., Laundon, M., Cathcart, A. Bashar, A. & Nayak, R. (2022). First, do no harm: automated detection of abusive comments in student evaluation of teaching surveys. *Assessment & Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2022.2081668.