Evaluation of University of Missouri’s Instruction and Course Evaluation

Prepared by
Ze Wang (wangze@missouri.edu), Associate Professor, College of Education, MU
Chia-Lin Tsai
Paula McFarling

September 18, 2017

Executive Summary

In this evaluation report, we examined the internal structure of key constructs of the University of Missouri’s (MU) instruction and course evaluations (ICE), as well as relationships between these constructs and relevant variables. Based on the results, we conclude:

- The four key constructs (i.e., Course Content and Structure, Teaching Delivery, Learning Environment, and Assessment), each measured by individual items on the ICE forms, have good internal structure and reliability.
- There is an overall teaching effectiveness construct that could be represented by the 20 items supposedly measuring the four key constructs.
- The general teaching effectiveness item (*This instructor taught effectively considering both the possibilities and limitations of the subject matter and the course (including class size and facilities.)* could be used to replace the overall teaching effectiveness construct (20 items), but only for larger classes and classes with a high number of respondents.
- There is more variability among students than variability among classes and instructors. That is, the major source of different ratings is differences between students instead of differences between classes and instructors.
- Class average GPA was not strongly related to student rating of teaching.
- Instructor’s sex was not strongly related to student rating of teaching.
- Student gender was not strongly related to their rating, for classes taught by a male instructor, or a female instructor.
- There are some group differences. They are summarized below:
    - Graduate-level courses received ratings that were about 0.14 to 0.17 standard deviations (SDs) higher than undergraduate-level courses.
    - Larger classes tend to receive lower ratings than smaller classes. For undergraduate courses, compared to classes with an enrollment of 30 or fewer students, classes with 31-99 students were rated about 0.05 SDs lower on F2 "Teaching Delivery", F3 "Learning Environment", and F4 "Assessment"; classes with 100-250 students were rated about 0.07 to 0.14 SDs lower on the four key ICE constructs; and classes with sizes>250 were rated 0.112 to 0.208 SDs lower on the four key ICE constructs. For graduate courses, larger classes with an enrollment greater than 30 were rated 0.067 SDs lower on F1 "Course Content and Structure" than smaller classes with size of 30 or fewer students.
    - For undergraduate courses, Traditional classes with no online components received the highest ratings compared to classes with other instruction modes. Compared to Traditional classes with no online component, ratings for E-Learning and Online classes were lower on F2 "Teaching Delivery" (0.163 SDs) and F3 "Learning "Environment" (0.146 SDs); ratings for Web-facilitated classes

were 0.088 to 0.227 SDs lower on all four constructs; and ratings for Blended classes were 0.147 to 0.226 SDs lower on all four constructs.
- For graduate courses, compared to Traditional classes with no online component, ratings for E-Learning classes were (0.108 SDs) higher on F1 "Course Content and Structure" and (0.136 SDs) lower on F2 "Teaching Delivery"; Online classes were rated lower on F2 "Teaching Delivery" (by 0.155 SDs) and F3 "Learning "Environment" (by 0.106 SDs).
- Students rated elective courses higher than required courses. For undergraduate courses, when the course was elective, students gave ratings that were 0.087 to 0.176 SDs higher on all four constructs than when the course was a requirement. For graduate courses, when the course was elective, students gave ratings that were 0.074 to 0.099 SDs higher on all four constructs than when the course was a requirement.
- There are some small gender differences. For undergraduate courses, female students gave 0.02 to 0.088 SDs **higher** ratings than male students for all four key ICE constructs. For graduate courses, female students gave 0.037 SD **lower** ratings than male students on F4 "Assessment."
- Compared to freshmen, juniors rated 0.040 SDs higher on F4 "Assessment", and seniors rated 0.061 to 0.100 SDs higher on all four constructs.

We also make the following recommendations:

- The ICE forms, especially forms with items that measure the four key constructs of Course Content and Structure, Teaching Delivery, Learning Environment, and Assessment, continue to be used for instruction evaluation.
- When administrators use ICE to assist decision making, consider the level of courses (undergraduate vs. graduate courses), class sizes, instruction mode, and required and elective courses. For example, if departmental or divisional averages are to be calculated, whenever possible, pool courses at the same level, of similar class sizes and instruction mode, and that are either required or elective. In order to have more comparable courses, departmental or divisional averages may be calculated across multiple semesters, particularly for graduate level courses.
- At the same time, when high-stakes decisions are made that use ICE results as supplementary information, it is recommended that coarse categories be used (e.g., Not effective, Effective, Highly effective) instead of many fine categories. This is because there is more variability among students than variability among classes and instructors for the ratings.
- Departments and divisions should promote the importance of instruction and teaching evaluation in order to get a higher response.

Analyses for this report are based on five semesters of ICE data collected for the MU campus. While we have some interesting findings, we did not conduct analysis separately for individual colleges, divisions, or departments. As a result, the conclusions and recommendations are at a relatively broad level for the MU campus. Individual colleges, divisions, and departments may have unique features that are not revealed in this report.

Background

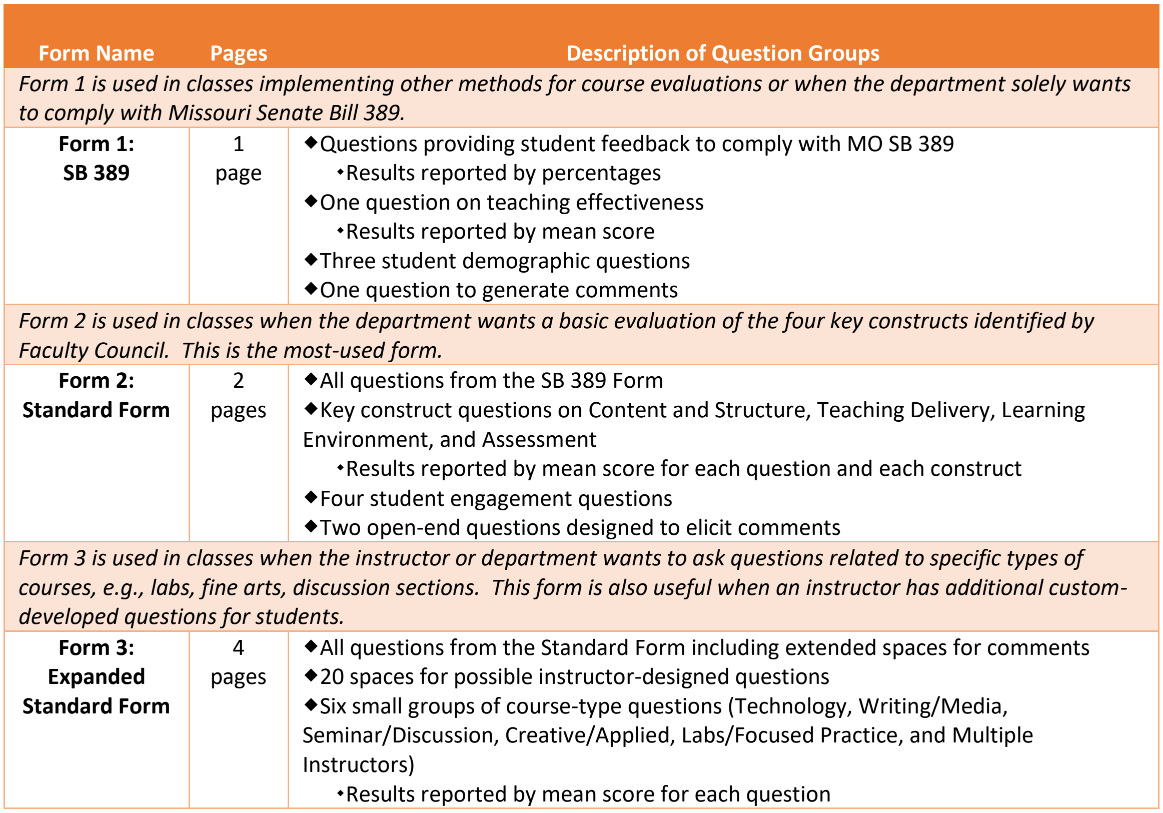
At the University of Missouri (MU), a new course evaluation system was designed to provide information that would promote excellence in teaching. In 2014, MU implemented a new system designed to improve the information aggregated from student ratings with the hopes of 1) aiding faculty and instructors in their instructional design; 2) assisting administrators with decision-making; and 3) helping future students select courses.

Beginning in 2012, the Assessment Resource Center (ARC) was asked to develop new course evaluations using four key constructs established by MU Faculty Council, i.e., Course Content and Structure, Teaching Delivery, Learning Environment, and Assessment. Using surveys, focus groups, and discussion sessions with MU faculty, staff, students, and administrators, ARC developed twenty Likert-scale questions to represent the four constructs. After adding a question on teaching effectiveness which was carried over from the earlier forms, Faculty Council approved the new Evaluation of Instruction and Course forms in 2013 stating, “the revised forms are a better, streamlined, and more flexible MU-specific instrument for the evaluation of teaching.” Use of the new forms and their reports began in Fall 2013 and completely replaced the previous forms by Fall 2014. In addition, a new online platform using these forms was implemented in Fall 2014, providing a choice between paper and online evaluation forms.

Student gender, requirement vs. elective class, and student status (i.e., freshman, sophomore, junior, senior, graduate, other) were self-reported. The new forms included a gender question which was deleted from all forms in August 2016 due to student concerns.

Missouri Senate Bill 389 (MO SB 389) requires public institutions of higher education to collect instructor ratings from students and to post these on the institution’s website. These institution-designed questions collect data considered “consumer” information for both current and incoming students. In 2014, the five new SB 389-compliant questions designed by ARC and approved by MU’s Faculty Council were implemented campus-wide as the Feedback for Other Students section of the new forms. To protect student confidentiality, any course with five or fewer completed evaluations will not have their SB389 evaluation results posted. These questions ask students if they would recommend this class to others according to each construct. The responses to these questions are meant to inform “consumers” and are not intended to be used for any type of internal evaluation, e.g., annual evaluations or promotion and tenure dossiers; however, these ratings should mirror the ratings from the twenty Likert-scale questions.

For consistency across campus, each department or program is encouraged to use one of the three ARC-provided course evaluation forms.



Form Name	Pages	Description of Question Groups
Form 1 is used in classes implementing other methods for course evaluations or when the department solely wants to comply with Missouri Senate Bill 389.		
Form 1: **SB 389**	1 page	◆Questions providing student feedback to comply with MO SB 389 •Results reported by percentages ◆One question on teaching effectiveness •Results reported by mean score ◆Three student demographic questions ◆One question to generate comments
Form 2 is used in classes when the department wants a basic evaluation of the four key constructs identified by Faculty Council. This is the most-used form.		
Form 2: **Standard Form**	2 pages	◆All questions from the SB 389 Form ◆Key construct questions on Content and Structure, Teaching Delivery, Learning Environment, and Assessment •Results reported by mean score for each question and each construct ◆Four student engagement questions ◆Two open-end questions designed to elicit comments
Form 3 is used in classes when the instructor or department wants to ask questions related to specific types of courses, e.g., labs, fine arts, discussion sections. This form is also useful when an instructor has additional custom-developed questions for students.		
Form 3: **Expanded Standard Form**	4 pages	◆All questions from the Standard Form including extended spaces for comments ◆20 spaces for possible instructor-designed questions ◆Six small groups of course-type questions (Technology, Writing/Media, Seminar/Discussion, Creative/Applied, Labs/Focused Practice, and Multiple Instructors) •Results reported by mean score for each question

Figure 1. Three forms of Evaluation of Instruction and Course developed by the Assessment Resource Center at the University of Missouri. Source: Guide to the Evaluation of Instruction and Course, 2013, revised 2017.

Individual students complete the Evaluation of Instruction and Course forms near the end of their course. Results from individual surveys are aggregated, analyzed, and reported for each class-and-instructor pair. Evaluation reports are only available in portable document format (PDF) on the course evaluation website. Each semester, evaluation reports begin to be released 36 hours after the date grades are due. All instructors can view and print their own evaluation reports including the full set of student-written comments. Department-designated support staff with myZou-security-approval can also access all reports online.

The Likert-scale response choices are *Strongly Disagree* (1), *Disagree* (2), *Neutral* (3), *Agree* (4), and *Strongly Agree* (5). Response data for the reports are stated in two ways: for a single question using response choice percentages and a mean score, and a mean score for the group of questions for each construct. One of the past SB 389 questions on general teaching effectiveness is included at the request of Faculty Council and is now reported using a 5-point scale rather than a 4-point scale, consistent with the new questions.

Using the SB 389 questions, students report their recommendations to other students regarding the construct areas and these are reported as percentages.

ARC is responsible for maintaining and distributing the course evaluations, analyzing results, and providing official instructor evaluation reports. Working closely with the Vice Provost for Undergraduate Studies, ARC maintains up-to-date forms and reports and provides additional campus reports when requested.

The Present Evaluation

This evaluation report focuses on the four key constructs of MU's Evaluation of Instruction and Course: Course Content and Structure, Teaching Delivery, Learning Environment, and Assessment; as well as a general teaching effectiveness item, five Missouri Senate Bill 389 (MO SB389) items, instructor's sex and class average GPA. For short, the Evaluation of Instruction and Course system is called ICE (instruction and course evaluation).

The principal guiding question is "Is MU's ICE reliable and valid?"

We follow the *Standards for Educational and Psychological Testing* (American Education Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; hereafter the *Standards*) to answer this question. For validity, the *Standards* emphasizes collecting relevant evidence to support the intended interpretation and use of test scores. According to the *Standards*, there are six specific forms of validity evidence and some forms are more relevant at the test development stage. Pertaining to this report, two forms of validity evidence are emphasized: evidence regarding internal structure, and evidence regarding relationships with conceptually related constructs. Another form of validity evidence, evidence regarding relationships with criteria, which could be very useful, is not considered in this report due to lack of clearly defined criteria in the context for the use of MU's ICE.

Throughout its development, MU's ICE is intended to evaluate faculty's teaching effectiveness. Some departments and colleges at MU also use ICE scores for promotion and merit-based performance evaluation. Although such use might be legitimate for specific departments and colleges, in this report, we focus on the primary intended use for faculty's teaching effectiveness. Therefore, the validity part of the principal guiding question becomes "**Can MU's ICE be used to assess faculty's teaching effectiveness**"? Also, we would like to point out that teaching effectiveness in this context is not equivalent to student learning. Although the ultimate goal for any teaching is student learning, research has shown that student evaluation of teaching ratings are not related to student learning (e.g., Uttl et al., 2016).

Instead of treating reliability as separate from validity, the *Standards* position the reliability of scores as having implications for validity because the level of reliability of scores has consequences on the intended interpretation of those scores. Therefore, the reliability part of the principal guiding question is subsumed under the validity part of the question, specifically, validity evidence regarding internal structure.

Another concern is reliability of group means. For each class-and-instructor pair, there are usually multiple students who rate the teaching. While the items are designed for use at the student level, individual students' ratings are evidently aggregated in order to evaluate the performance of a particular instructor for a particular class. The aggregation is usually done by calculating the mean of all students' ratings for that instructor and class. Because not every student at MU rates every class and instructor, it is common to assume, and it can be tested, that the variation due to the sampling of students (in this context, the variation due to course selection and choice to complete the course evaluation form or not) can be a major source of error, especially if class sizes, or the numbers of students who choose to complete the evaluation, are small. In fact, the error associated with the sampling of students could be a significant source of error.

Descriptive Statistics of MU's ICE

For this report, only courses with at least six students enrolled that used Standard Form (Form 2) or Expanded Standard Form (Form 3) were included. Across five semesters (Fall 2014, Spring 2015, Fall 2015, Spring 2016, and Fall 2016), there were 386,016 ratings by students for 16,169 unique class-and-instructor pairs. The number of students who rated the same class and instructor at a given semester ranged from 1 to 480, with a standard deviation of 32.61. From Table 1 and Figure 2, while the average ratings for the four key ICE constructs across all class-and-instructor pairs in a given semester were usually high (about 4.2 to 4.5 on a 1-5 point scale), the standard deviations of average ratings across class-and-instructor pairs were about 0.3 to 0.6. While the highest average rating for any key construct in a given semester was always the highest possible score (i.e., 5.00), the lowest average rating could be as low as 1.00 and typically at the upper end of 1 or lower end of 2 for undergraduate courses, and below 3 for graduate courses.

Students' ratings for the same class/instructor could also vary. From Table 1 and Figure 3, the standard deviations of students' ratings for the same class/instructor, for the four key ICE constructs, could range from 0 to 2.8, with average standard deviations typically in the 0.5-0.7 range.

These indicate that the variability of ratings is more due to differences among students than due to differences among classes and instructors. Because of this, we recommend that when departments and colleges use student ratings for teaching effectiveness purposes, they use coarse categories (e.g., Not effective, Effective, Highly effective) instead of many fine categories.

Table 1. Descriptive Statistics of Means and Standard Deviations of Student Ratings of Classes/Instructors

Undergraduate Courses	Fall 2014				Spring 2015				Fall 2015				Spring 2016				Fall 2016			
	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD
Construct Means																				
Content	2.25	5.00	4.34	0.31	2.30	5.00	4.38	0.30	2.25	5.00	4.37	0.31	2.00	5.00	4.40	0.33	2.11	5.00	4.36	0.36
Delivery	2.36	5.00	4.39	0.42	2.32	5.00	4.43	0.41	2.00	5.00	4.40	0.44	1.00	5.00	4.45	0.43	1.25	5.00	4.39	0.48
Environment	2.29	5.00	4.42	0.38	2.50	5.00	4.45	0.37	2.22	5.00	4.44	0.39	1.00	5.00	4.46	0.39	1.73	5.00	4.43	0.43
Assessment	1.00	5.00	4.19	0.47	1.00	5.00	4.24	0.46	1.61	5.00	4.23	0.46	1.67	5.00	4.27	0.47	1.33	5.00	4.22	0.51
Total Scale	2.20	5.00	4.36	0.37	2.47	5.00	4.40	0.36	2.28	5.00	4.38	0.37	1.30	5.00	4.42	0.38	1.82	5.00	4.37	0.42
Construct SD																				
Content	0.00	2.00	0.58	0.21	0.00	2.83	0.56	0.21	0.00	2.83	0.57	0.21	0.00	2.83	0.56	0.22	0.00	2.18	0.58	0.24
Delivery	0.00	2.42	0.59	0.25	0.00	2.83	0.57	0.26	0.00	2.83	0.59	0.27	0.00	2.73	0.57	0.27	0.00	2.12	0.60	0.29
Environment	0.00	2.83	0.58	0.25	0.00	2.83	0.58	0.25	0.00	2.83	0.58	0.25	0.00	2.83	0.57	0.26	0.00	2.12	0.59	0.28
Assessment	0.00	2.12	0.71	0.26	0.00	2.83	0.70	0.27	0.00	2.36	0.70	0.27	0.00	2.13	0.69	0.28	0.00	2.36	0.70	0.29
Total Scale	0.00	2.00	0.55	0.22	0.00	2.83	0.54	0.22	0.00	2.76	0.55	0.23	0.00	2.69	0.54	0.24	0.00	2.12	0.56	0.25

Table 1 (cont.) Descriptive Statistics of Means and Standard Deviations of Student Ratings of Classes/Instructors

Graduate Courses	Fall 2014				Spring 2015				Fall 2015				Spring 2016				Fall 2016			
	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD
Construct Means																				
Content	2.86	5.00	4.46	0.34	2.73	5.00	4.46	0.37	2.95	5.00	4.49	0.36	2.72	5.00	4.49	0.38	2.50	5.00	4.49	0.36
Delivery	2.52	5.00	4.50	0.42	2.00	5.00	4.50	0.44	1.17	5.00	4.52	0.45	2.68	5.00	4.51	0.45	2.50	5.00	4.51	0.43
Environment	2.34	5.00	4.51	0.40	2.60	5.00	4.53	0.41	1.33	5.00	4.55	0.42	2.17	5.00	4.53	0.42	2.50	5.00	4.54	0.39
Assessment	1.00	5.00	4.28	0.51	1.25	5.00	4.32	0.55	1.33	5.00	4.34	0.53	2.00	5.00	4.33	0.51	1.00	5.00	4.32	0.58
Total Scale	2.76	5.00	4.46	0.38	2.53	5.00	4.47	0.41	1.74	5.00	4.50	0.40	2.70	5.00	4.49	0.41	2.50	5.00	4.48	0.40
Construct SD																				
Content	0.00	1.83	0.54	0.29	0.00	2.83	0.54	0.32	0.00	2.00	0.48	0.26	0.00	2.12	0.51	0.29	0.00	2.12	0.50	0.28
Delivery	0.00	2.12	0.53	0.33	0.00	2.83	0.54	0.36	0.00	1.91	0.50	0.31	0.00	2.03	0.52	0.32	0.00	2.12	0.52	0.33
Environment	0.00	2.47	0.53	0.32	0.00	2.83	0.53	0.36	0.00	2.24	0.49	0.32	0.00	1.95	0.51	0.31	0.00	2.24	0.51	0.32
Assessment	0.00	2.59	0.69	0.34	0.00	2.83	0.65	0.37	0.00	2.14	0.64	0.35	0.00	1.92	0.67	0.34	0.00	2.83	0.65	0.38
Total Scale	0.00	1.95	0.50	0.29	0.00	2.83	0.51	0.33	0.00	1.93	0.47	0.28	0.00	1.95	0.49	0.29	0.00	2.12	0.49	0.30

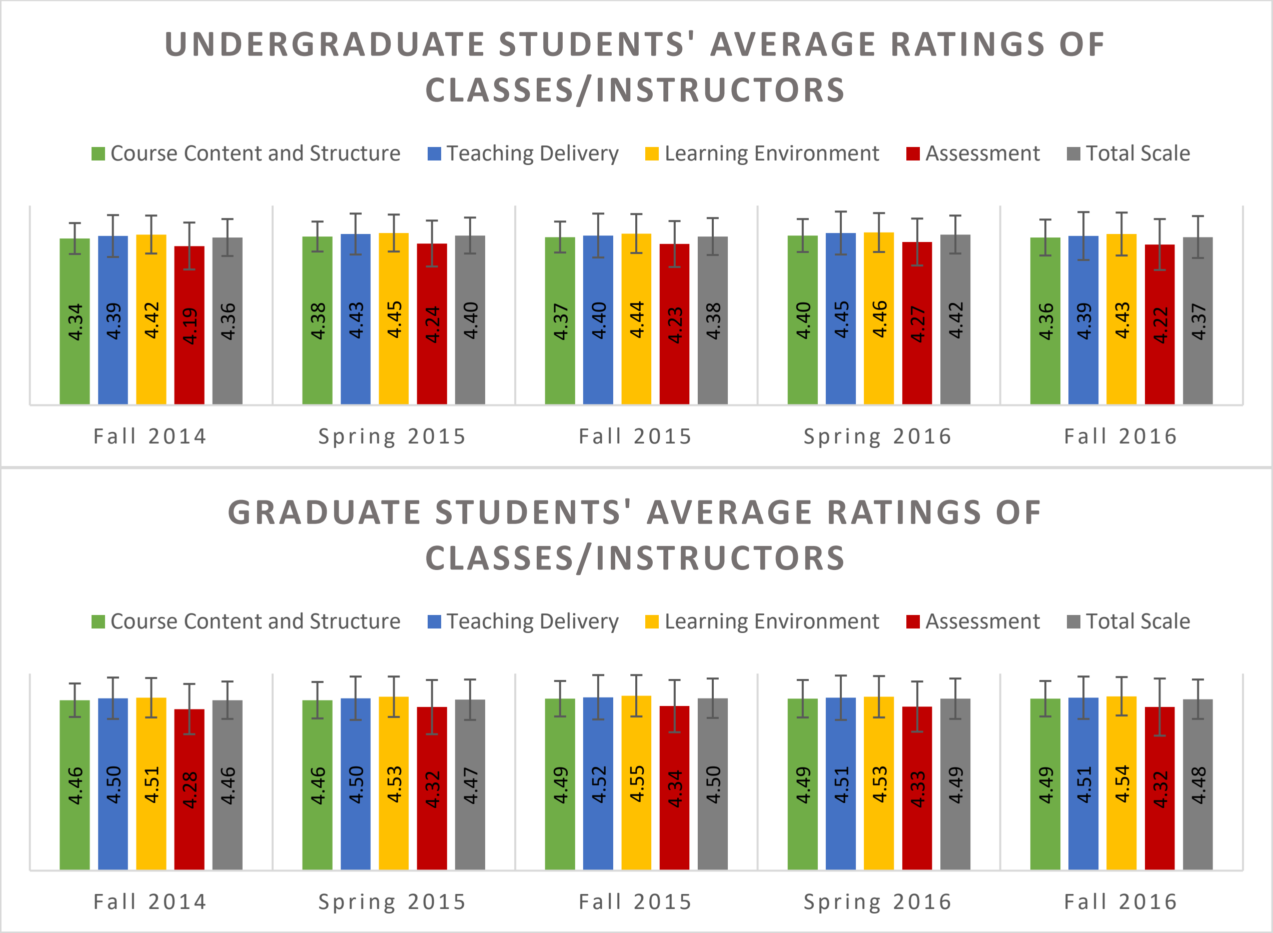


Figure 2. Students' average ratings on the four key ICE constructs and the overall scale. Numbers and colored bars show average ratings and error bars represent standard deviations of student ratings.

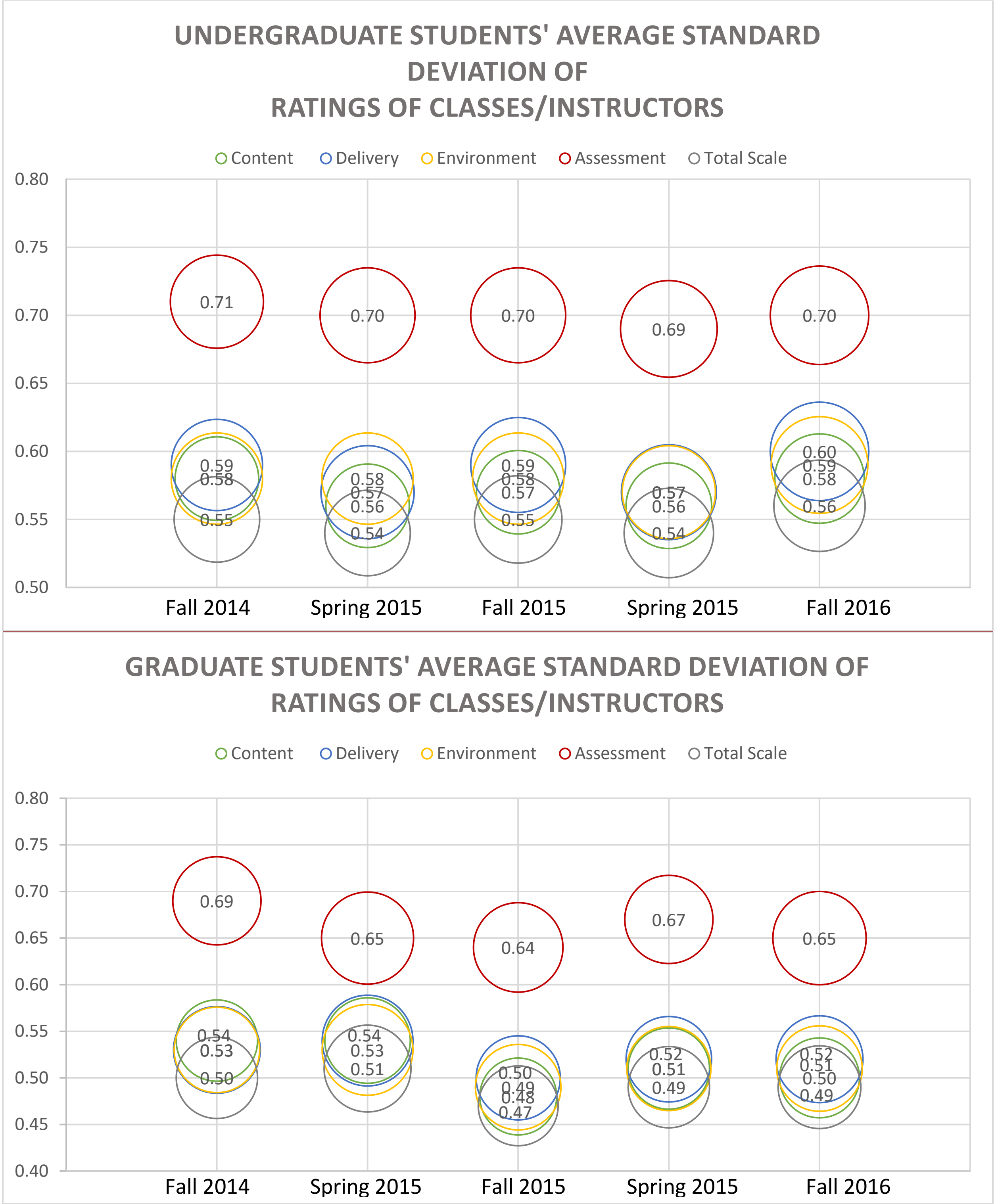


Figure 3. Students' average standard deviations of ratings on the four key ICE constructs and the overall scale. Numbers and center of circles show average ratings and sizes of circles represent standard deviations of the standard deviations of student ratings.

Internal Structure of MU's ICE

The internal structure of MU's ICE was examined through factor analysis. All items were treated as continuous variables. Clustering by class/instructor was taken into account when analysis was conducted.

- Are there four constructs as hypothesized? Is there an overall teaching effectiveness construct based on the 20 Likert scale items that supposedly measure the four constructs?

  Confirmatory factor analysis (CFA) is a common statistical modeling technique used to test factor structure in education, psychology and other fields. In CFA, constructs are usually represented by latent factors, which are unobservable and measured by observable indicators. CFA starts with a set of hypotheses that specify the number of latent factors, the number of observable indicators, and the relationships between latent factors and observable indicators. The hypothesized model can be tested against data collected on the observable indicators to see if it is supported. Using model fit indices, if there is a model-data consistency, the researcher could conclude that the hypothesized model is supported. On the other hand, if the model-data consistency is poor, the hypothesized model is usually concluded as not being supported. Sometimes, the researcher may test several hypothesized models in order to select the one that is best supported by the data; or in the case that multiple hypothesized models are consistent with the data, the researcher may conclude that there are different ways to interpret the construct.

  From the development stage of MU's ICE, the hypothesized factor structure can be represented by Figure 4. In this figure, the four key constructs – Course Content and Structure, Teaching Delivery, Learning Environment, and Assessment – are named f1, f2, f3, and f4, respectively. These are latent factors represented by ovals. Each latent factor is measured by multiple items, corresponding to the questions on the ICE's forms. For example, "Course Content and Structure," or f1, is measured by four items q111, q112, q113, and q114.

  Other considerations of testing a CFA model include whether the observable indicators should be treated as continuous variables or variables with other types of levels of measurement (nominal, or ordinal), whether responses from participants are independent or there is some dependency, estimation methods (maximum likelihood or other estimators), and treatment of missing values. For this project, the 20 statements that supposedly measure the four key constructs of MU's ICE are rated on a five-point Likert scale (Strongly disagree, Disagree, Neutral, Agree, and Strongly agree). While there are arguments among researchers in terms of whether Likert-scale items should be treated as continuous or ordinal variables, for this project, we use them as continuous variables such that 1=Strongly disagree, 2=Disagree, 3=Neutral, 4=Agree, and 5=Strongly agree. This is because students' ratings for each class and instructor are aggregated (i.e., averaged) when results are reported to instructors and departments. Such aggregation requires that the variables be continuous and from the measurement perspective, falls under the classical test theory (CTT) framework.

Students' responses are not independent since multiple students rate the same class and instructor. Therefore, we would assume that students rating the same class and instructor would give similar ratings to each other than students rating different classes or instructors. Such dependency is called clustering (students are nested within class and instructor) and can be accommodated during statistical analysis. Another dependency is that the same student could rate multiple classes and instructors. This is common when the student takes multiple courses and/or when the student rates different instructors (professor, TA) of the same course. For example, some students may have the tendency to always give high ratings. However, due to the anonymous nature of the data (i.e., we do not have unique or identifiable information for students), we cannot accommodate this type of dependency.

For the estimation method, the robust maximum likelihood (MLR) estimator is used. MLR is a maximum likelihood estimator with standard errors and a chi-square test statistic that are robust to non-normality and non-independence of observations for complex data structures. While the parameter estimates from MLR are the same as those from the conventional maximum likelihood estimator, the MLR standard errors are computed using a sandwich estimator. The MLR chi-square test statistic is asymptotically equivalent to the Yuan-Bentler T2* test statistic. In addition, this is a full information maximum likelihood method for missing data in that missing is assumed to be at random and that both complete (no missing) and partial (with some missing) data points are used in the estimation of model fit and model parameters.

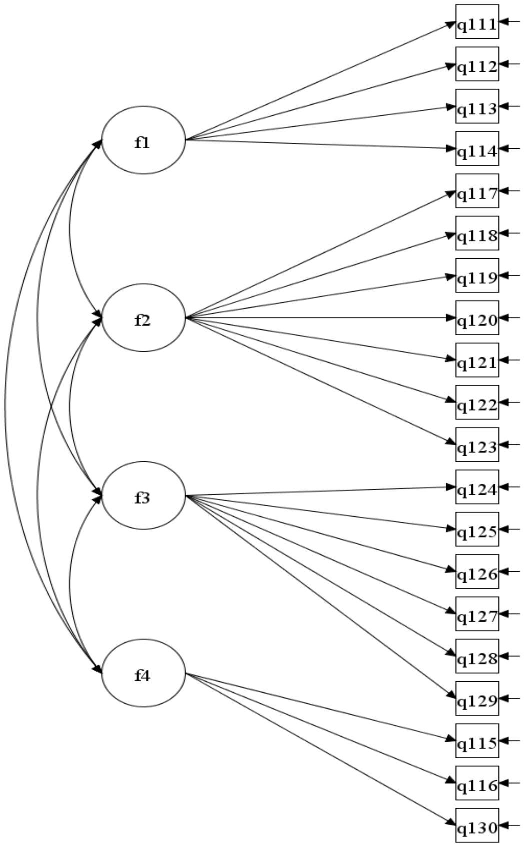


Figure 4. Hypothesized four-factor structure of MU's ICE.

For analysis, we tested a four-factor CFA model that corresponds to the above hypothesized structure for graduate-level courses (course numbers at 7000 levels or above) and undergraduate-level courses (course numbers at 4000 level or below) separately. Fit statistics for this model are shown as bolded in Table 2. For model fit, we use regular cutoffs for CFA models: Comparative Fit Index (CFI) > 0.95, Root Mean Square Error of Approximation (RMSEA) < 0.06, and Standardized Root Mean Square Residual (SRMR) < 0.08. The four-factor is supported by the data. Table 3 includes standardized factor loadings, which reflect the estimated relationship between each observable indicator and its hypothesized construct. Although there is no specific cutoff for factor loadings, researchers typically expect factor loadings from CFA models to be 0.40 or above.

Table 2. Model Fit Statistics

	# parameters	Chi-square	DF	RMSEA	90% CI RMSEA		CFI	TLI	SRMR
Graduate Four Factor CFA (n=34650)	**66**	**9328.86**	**164**	**0.04**	**0.04**	**0.04**	**0.95**	**0.94**	**0.03**
GradSingleFactorCFA	60	14902.57	170	0.05	0.05	0.05	0.92	0.91	0.04
GradTwoFactor CFA	61	11269.22	169	0.04	0.04	0.04	0.94	0.93	0.03
Grad2ndOrderCFA	64	9645.22	166	0.04	0.04	0.04	0.95	0.94	0.03
Graduate Bifactor Model, 4 group factors	80	6768.82	150	0.04	0.04	0.04	0.96	0.95	0.03
Undergrad Four Factor CFA Model (n=343966)	**66**	**77904.02**	**164**	**0.04**	**0.04**	**0.04**	**0.95**	**0.95**	**0.03**
UndergradSingleFactorCFA	60	138227.78	179	0.05	0.05	0.05	0.92	0.91	0.04
UndergradTwoFactorCFA	61	85743.16	169	0.04	0.04	0.04	0.95	0.94	0.03
Undergrad2ndOrderCFA	64	82991.89	166	0.04	0.04	0.04	0.95	0.94	0.03
UndergradBifactor Model, 4 group factors	80	72278.85	150	0.04	0.04	0.04	0.96	0.95	0.03

Table 3. Factor Loadings and Correlations Between Factors

Construct / Items	Graduate	Undergraduate
Course Content and Structure		
The syllabus clearly explained the course objectives, requirements, and grading system.	0.77	0.78
Course content was relevant and useful (e.g., readings, online media, classwork, assignments).	0.84	0.84
Resources (e.g., articles, literature, textbooks, class notes, online resources) were easy to access.	0.75	0.77
This course challenged me.	0.62	0.46
Teaching Delivery		
This instructor was consistently well-prepared.	0.82	0.82
This instructor was audible and clear.	0.79	0.79
This instructor was knowledgeable and enthusiastic about the topic.	0.81	0.83

This instructor effectively used examples/illustrations to promote learning.	0.87	0.87
This instructor fostered questions and/or class participation.	0.83	0.82
This instructor clearly explained important information/ideas/concepts.	0.89	0.89
This instructor effectively used teaching methods appropriate to this class (e.g., critiques, discussion, demonstrations, group work).	0.89	0.88
Learning Environment		
This instructor responded appropriately to questions and comments.	0.85	0.86
This instructor stimulated student thinking and learning.	0.89	0.89
This instructor promoted an atmosphere of mutual respect regarding diversity in student demographics and viewpoints, such as race, gender, or politics.	0.79	0.82
This instructor was approachable and available for extra help.	0.81	0.83
This instructor used class time effectively.	0.83	0.83
This instructor helped students to be independent learners, responsible for their own learning.	0.84	0.85
Assessment		
I was well-informed about my performance during this course.	0.76	0.72
Assignments/projects/exams were graded fairly based on clearly communicated criteria.	0.80	0.78
This instructor provided feedback that helped me improve my skills in this subject area.	0.88	0.85

Table 3 (cont.). Factor Loadings and Correlations Between Factors

	F1	F2	F3	F4
F1: Course Content and Structure		0.83	0.84	0.91
F2: Teaching Delivery	0.87		0.95	0.90
F3: Learning Environment	0.87	0.96		0.93
F4: Assessment	0.88	0.88	0.91	

Note: Correlations for the graduate sample are shaded, and correlations for the undergraduate sample are not shaded.

Despite that the four-factor CFA model fit the data from both graduate- and undergraduate-level courses, the correlations among the factors were high, suggesting that there was too much overlap among the factors – the factor structure may be simpler with fewer factors. We tested four alternative models: a single-factor model, a two-factor CFA model, a second-order factor model and a bifactor model.

For the single-factor model, all 20 ICE items are hypothesized to measure the "Teaching Effectiveness" factor. For the two-factor CFA model, the first factor is hypothesized to be a "Course" factor and measured by the six ICE statements that start with "This Course," and the second factor is hypothesized to be an "Instructor" factor and measured by the 14 ICE statements that start with "This Instructor." For the second-order factor model, the four factors from the earlier four-factor CFA model are hypothesized to measure a higher- (i.e., second-) order factor, which can be called "Teaching Effectiveness." For the bifactor model, it is hypothesized that all 20 ICE items directly measure something in common that is called "Teaching Effectiveness," and that there are four group factors that additionally account for relationships among items. These group factors correspond to the latent factors in the earlier four-factor CFA model, although their meanings are different now that they only account for residual covariation after the general factor supposedly accounts for most of the common variance among the 20 items.

Model fit statistics (see Table 2) suggest that for both graduate and undergraduate students, multiple factor structures were consistent with the data. Of the alternative models, particularly, the bifactor model, which has the most number of parameters and therefore is more complex than the other models, fit the data very well. In addition, the factor loadings on the general factor in the bifactor are high, suggesting that the 20 ICE items measure something in common. Nevertheless, the single factor CFA model had poor model fit for both graduate and undergraduate data, suggesting that the 20 ICE items are not unidimensional.

Based on these, we conclude that the original hypothesized four-factor model, which is consistent with the key constructs proposed for MU's ICE, can be supported by both graduate and undergraduate data. However, the four key constructs are highly correlated (see Table 3). In addition, there is an overall construct based on the 20 ICE items. This overall construct can be best represented by the general factor in the bifactor model.

- Is the measurement model invariant across groups (grouping by semester, graduate/undergraduate classes, class size, instruction mode, student gender, requirement vs. elective, and student status – freshman, sophomore, junior, and senior)?

  Measurement invariance has been increasingly a consideration during scale development and validation. The general idea of measurement invariance is that the "measure" (or scale, instrument, etc.), which is analogous to a ruler in the physical world, should function in a similar way for different groups so that these groups can be compared using this measure. For this project, measurement invariance was tested using the four-factor CFA model across various grouping variables. Consistent with measurement invariance literature, three types of invariance models were tested: configural invariance, metric invariance, and scalar invariance, by sequentially imposing cross-group constraints on model parameters. Recommendations of changes in model fit indices have been proposed for testing measurement invariance (Chen, 2007; Cheung & Rensvold, 2002). According to these recommendations, if CFI does not decrease by at least 0.01 and RMSEA does not increase by at least 0.015, the more restricted model should be chosen. For the various grouping variables, model fit indices always suggest the scalar invariance model was the best considering both model fit and model parsimony (see Table 4), suggesting that relationships between ICE items and the latent factors are comparable across groups based on the various grouping variables and thus latent factor means can be compared.

Table 4. Model Fit Statistics for Testing Measurement Invariance

	# para.	Chi-square	DF	RMSEA	90% CI RMSEA		CFI	TLI	SRMR
nfigural Invariance Model	330	10560.02	820	0.04	0.04	0.04	0.95	0.94	0.03
Metric Invariance Model	266	10138.320	884	0.04	0.04	0.04	0.95	0.95	0.03
Scalar Invariance Model	**202**	**10271.06**	**948**	**0.04**	**0.04**	**0.04**	**0.95**	**0.95**	**0.04**

Note: Multiple semesters. Four-factor CFA-graduate classes; n=34650 (Fall2014 n=7819; Spring2015 n=6725; Fall2015 n=7305; Spring2016 n=5858; Fall2016 n=6943)

	# para.	Chi-square	DF	RMSEA	90% CI RMSEA		CFI	TLI	SRMR
Configural Invariance Model	330	79845.27	820	0.04	0.04	0.04	0.95	0.95	0.03
Metric Invariance Model	266	74940.82	884	0.04	0.04	0.04	0.96	0.95	0.03
Scalar Invariance Model	**202**	**72752.74**	**948**	**0.03**	**0.03**	**0.03**	**0.96**	**0.96**	**0.03**

Note: Multiple semesters. Four-factor CFA-undergraduate classes; n=343966 (Fall2014 n=77492; Spring2015 n=65998; Fall2015 n=73016; Spring2016 n=62220; Fall2016 n=65240)

	# para.	Chi-square	DF	RMSEA	90% CI RMSEA		CFI	TLI	SRMR
Configural Invariance Model	132	89850.63	328	0.04	0.04	0.04	0.95	0.95	0.03
Metric Invariance Model	116	88484.06	344	0.04	0.04	0.04	0.95	0.95	0.03
Scalar Invariance Model	**100**	**89784.73**	**360**	**0.04**	**0.04**	**0.04**	**0.95**	**0.95**	**0.04**

Note: Multiple group 4-factor CFA-undergraduate/graduate; n=378616 (Undergrad n=343966; Grad n=34650)

	# para.	Chi-square	DF	RMSEA	90% CI RMSEA		CFI	TLI	SRMR
Configural Invariance Model	264	79344.150	656	0.04	0.04	0.04	0.94	0.95	0.03
Metric Invariance Model	216	75316.520	704	0.04	0.04	0.04	0.96	0.95	0.03
Scalar Invariance Model	**168**	**75461.72**	**752**	**0.03**	**0.03**	**0.03**	**0.96**	**0.96**	**0.04**

Note: Multiple groups by class size 4-factor CFA-undergraduate; n=343966 (Csize<=30 n=134683; Csize 31-99 n=102535; Csize 100-250 n=49755; Csize>250 n=56993)
Four groups of class sizes: <=30, 31-99, 100-250, >250

	# para.	Chi-square	DF	RMSEA	90% CI RMSEA		CFI	TLI	SRMR
Configural Invariance Model	132	9494.80	328	0.04	0.04	0.04	0.95	0.94	0.03
Metric Invariance Model	116	9155.26	344	0.04	0.04	0.04	0.95	0.94	0.03
Scalar Invariance Model	**100**	**9136.09**	**360**	**0.04**	**0.04**	**0.04**	**0.95**	**0.95**	**0.03**

Note: Multiple groups by class size 4-factor CFA-graduate; n=34650 (Csize<=30 n=26693; Csize 31-99 n=7957)
Two groups of class sizes: <=30, 31-99

	# para.	Chi-square	DF	RMSEA	90% CI RMSEA		CFI	TLI	SRMR
Configural Invariance Model	330	103103.72	820	0.04	0.04	0.04	0.95	0.95	0.03
Metric Invariance Model	266	101593.37	884	0.04	0.04	0.04	0.95	0.95	0.03
Scalar Invariance Model	**202**	**99519.45**	**948**	**0.04**	**0.04**	**0.04**	**0.96**	**0.96**	**0.03**

Note: Multiple groups by instruction mode 4-factor CFA-undergraduate; n=343793 (Traditional with no online n=315140; E-Learning, 100% online n=5893; Web-facilitated <30% online n=14790; Blended class 30-80% online n=6500; Online >80% online n=1470)

Five groups of instructional model: Traditional with no online; E-Learning, 100% online; Web-facilitated <30% online; Blended class 30-80% online; Online >80% online

	# para.	Chi-square	DF	RMSEA	90% CI RMSEA		CFI	TLI	SRMR
Configural Invariance Model	132	75239.31	328	0.04	0.04	0.04	0.95	0.95	0.03
Metric Invariance Model	116	76039.08	344	0.04	0.04	0.04	0.95	0.95	0.03
Scalar Invariance Model	**100**	**78254.50**	**360**	**0.04**	**0.04**	**0.04**	**0.95**	**0.95**	**0.03**

Note: Multiple groups by student gender 4-factor CFA-undergraduate; n=276230 (Male n=117215; Female n=159015)

	# para.	Chi-square	DF	RMSEA	90% CI RMSEA		CFI	TLI	SRMR
Configural Invariance Model	132	9116.19	328	0.04	0.04	0.04	0.95	0.94	0.03
Metric Invariance Model	116	9191.95	344	0.04	0.04	0.04	0.95	0.94	0.03
Scalar Invariance Model	**100**	**9533.84**	**360**	**0.04**	0.04	0.04	**0.95**	**0.94**	**0.03**

Note: Multiple groups by student gender 4-factor CFA-graduate; n=28573 (Male n=11687; Female n=16886)

	# para.	Chi-square	DF	RMSEA	90% CI RMSEA		CFI	TLI	SRMR
Configural Invariance Model	132	82799.20	328	0.04	0.04	0.04	0.95	0.95	0.03
Metric Invariance Model	116	82403.75	344	0.04	0.04	0.04	0.95	0.95	0.03
Scalar Invariance Model	**100**	**84927.36**	**360**	**0.04**	**0.04**	**0.04**	**0.95**	**0.95**	**0.03**

Note: Multiple groups by required/elective 4-factor CFA-undergraduate; n= (Freshman n=; Sophomore n=; Junior n=; Elective n=81911)

	# para.	Chi-square	DF	RMSEA	90% CI RMSEA		CFI	TLI	SRMR
Configural Invariance Model	132	9921.60	328	0.04	0.04	0.04	0.95	0.94	0.03
Metric Invariance Model	116	9858.93	344	0.04	0.04	0.04	0.95	0.94	0.03
Scalar Invariance Model	**100**	**10077.19**	**360**	**0.04**	0.04	0.04	**0.95**	**0.94**	**0.03**

Note: Multiple groups by required/elective 4-factor CFA-graduate; n=32858 (Required n=24188; Elective n=8670)

	# parameters	Chi-square	DF	RMSEA	90% CI RMSEA		CFI	TLI	SRMR
Configural Invariance Model	264	89491.70	656	0.04	0.04	0.04	0.95	0.95	0.03
Metric Invariance Model	216	88670.57	704	0.04	0.04	0.04	0.95	0.95	0.04
Scalar Invariance Model	**168**	**92086.84**	**752**	**0.04**	**0.04**	**0.04**	**0.95**	**0.95**	**0.04**

Note: Multiple groups by class status (Freshman, Sophomore, Junior, Senior) 4-factor CFA-undergraduate; n=323337 (Freshman n=76459; Sophomore n=78618; Junior n=83148; Senior n=85112)

From the scalar invariance models, we also found the following group differences on the latent factors. All reported group differences were statistically significant at the 0.05 level.

- For graduate-level courses, latent factor means are comparable across semesters.
- For undergraduate-level courses, latent factor means for later semesters are higher than for Fall 2014.

  - Ratings for graduate-level classes tend to have higher latent factor means (all four factors) than ratings for undergraduate-level classes.
  - For undergraduate courses, compared to small classes (enrollment<=30), classes with size 31-99 were rated lower on F2 "Teaching Delivery", F3 "Learning Environment", and F4 "Assessment"; classes with size 100-250 and classes with sizes>250 were rated lower on all four factors.
  - For graduate courses, compared to smaller classes (enrollment<=30), larger classes (enrollment>30) were rated lower on F1 "Course Content and Structure".
  - For undergraduate courses, compared to Traditional classes with no online component, ratings for E-Learning and Online classes were lower on F2 "Teaching Delivery" and F3 "Learning "Environment"; ratings for Web-facilitated and Blended classes were lower on all four factors.
  - For graduate courses, compared to Traditional classes with no online component, ratings for E-Learning classes were higher on F1 "Course Content and Structure," and lower on F2 "Teaching Delivery"; Online classes were rated lower on F2 "Teaching Delivery" and F3 "Learning "Environment";
  - For undergraduate courses, female students gave **higher** ratings than male students for all four key ICE constructs.
  - For graduate courses, female students gave **lower** ratings than male students. However, gender difference was only statistically significant for F4 "Assessment" factor at alpha=.05.
  - For undergraduate courses, when the course was elective, the student gave **higher** ratings on all four factors than when the course was a requirement (all significant at alpha=.05).
  - For graduate courses, when the course was elective, the student gave higher ratings on all four factors than when the course was a requirement (all significant at alpha=.05).
  - Compared to freshmen, juniors rated higher on F4 "Assessment", and seniors rated higher on all four factors.

- Reliability

For reporting purposes, we rely on the classical test theory due to its simplicity and straightforward way to calculate scale and subscale scores. Specifically, for each class-instructor pair, we calculated the average student rating on each item; next, we calculated the scale scores and subscale scores for each class-instructor pair. The scale score is the mean of the 20 ICE items and the subscale scores for each construct is the mean of the items that supposedly measure the construct. We checked the internal consistency of items for the total scale and subscales using Cronbach's alpha. These Cronbach's alpha coefficients are high (Table 5 and Figure 5), suggesting that there was high internal consistency among the items for each of the subscales and the total scale of MU's ICE.

Table 5. Reliability for Total Scale and Subscales of MU's ICE

	# items	All	**Undergraduate Classes**				
		Undergrads	Fall2014	Spring2015	Fall2015	Spring2016	Fall2016

Course Content and Structure	4	0.836	0.805	0.836	0.824	0.857	0.850
Teaching Delivery	7	0.972	0.968	0.972	0.971	0.973	0.976
Learning Environment	6	0.967	0.965	0.967	0.965	0.968	0.971
Assessment	3	0.898	0.892	0.901	0.890	0.900	0.903
Total Scale	20	0.980	0.978	0.980	0.979	0.981	0.982
	All	All	**Graduate Classes**				
		Grads	Fall2014	Spring2015	Fall2015	Spring2016	Fall2016
Course Content and Structure	0.842	0.857	0.831	0.881	0.827	0.879	0.868
Teaching Delivery	0.971	0.962	0.956	0.965	0.963	0.964	0.960
Learning Environment	0.964	0.950	0.944	0.957	0.951	0.955	0.946
Assessment	0.897	0.894	0.884	0.915	0.881	0.898	0.895
Total Scale	0.980	0.977	0.974	0.980	0.974	0.980	0.976

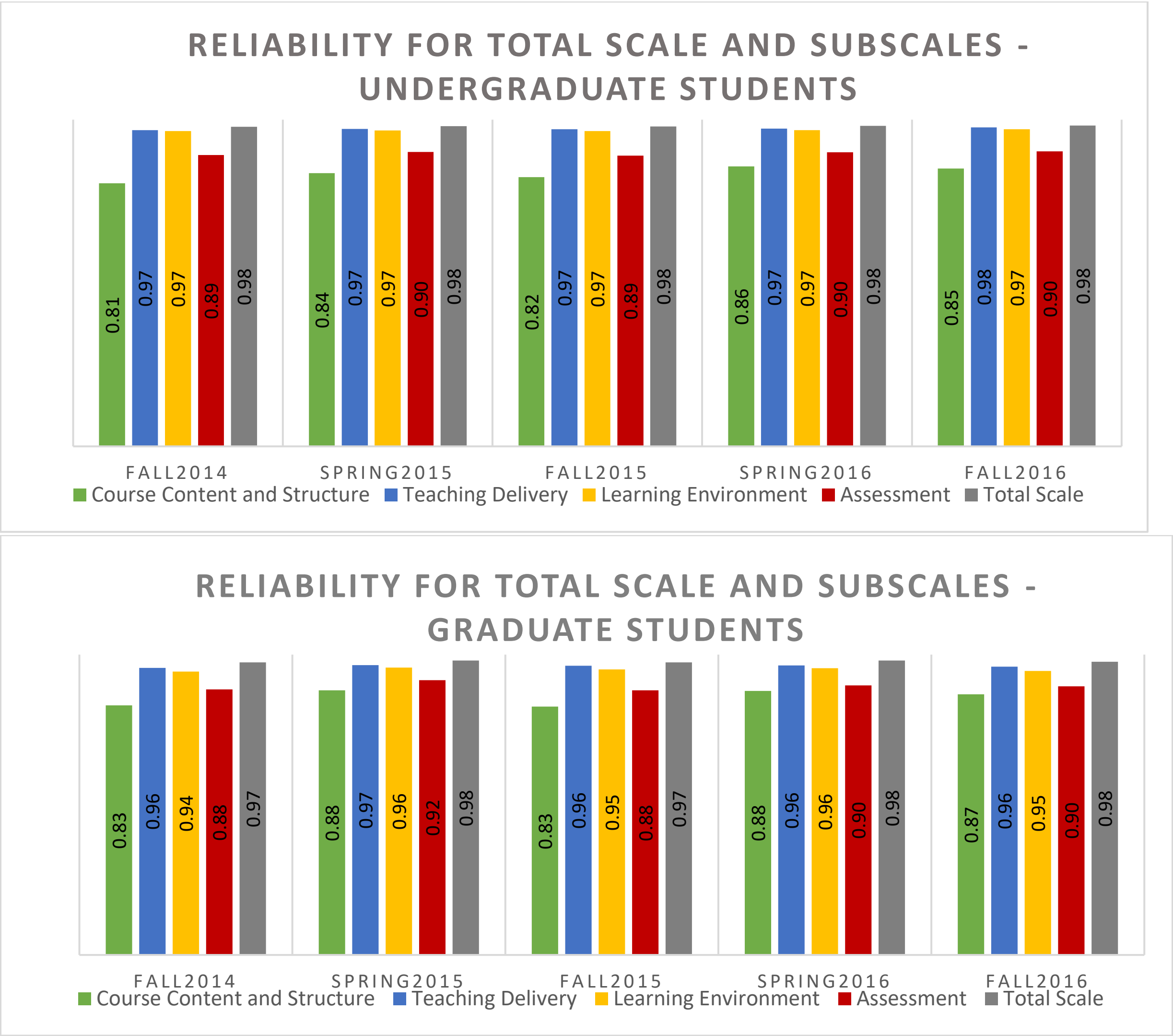


Figure 5. Reliability for the total scale and four subscales of MU's ICE.

- Intraclass Correlation

As mentioned earlier, we were interested in the variation due to the sampling of students. For each of the ICE items, and the subscale and total scale scores, we calculated the intraclass correlation (ICC). ICC is a commonly used statistic for agreement among raters. For this project, multiple students rated the same instructor for the same class. A low ICC reflects large variation (i.e., disagreement or inconsistency) among raters. The ICCs are in the range of 0.10 to 0.30, reflecting large variation due to the sampling of students (see Table 6 and Figure 6).

Table 6. Intraclass Correlations for ICE Items, Subscales, and Total Scale

Undergraduate Sample	Fall 2014	Spring2 015	Fall 2015	Spring2 016	Fall 2016
Sample size	77496	66001	73016	62220	65240
# of clusters	2782	2548	2730	2543	2691
Average cluster size	27.856	25.903	26.746	24.467	24.244
Items					
This instructor taught effectively considering both the possibilities and limitations of the subject matter and the course (including class size and facilities). Q104	0.22	0.22	0.22	0.23	0.26
The syllabus clearly explained the course objectives, requirements, and grading system. Q111	0.15	0.15	0.14	0.16	0.21
Course content was relevant and useful (e.g., readings, online media, classwork, assignments). Q112	0.15	0.14	0.15	0.17	0.18
Resources (e.g., articles, literature, textbooks, class notes, online resources) were easy to access. Q113	0.11	0.10	0.10	0.11	0.14
This course challenged me. Q114	0.15	0.13	0.14	0.14	0.14
I was well-informed about my performance during this course. Q115	0.20	0.21	0.21	0.22	0.25
Assignments/projects/exams were graded fairly based on clearly communicated criteria. Q116	0.19	0.18	0.19	0.20	0.22
This instructor was consistently well-prepared. Q117	0.21	0.22	0.23	0.23	0.27
This instructor was audible and clear. Q118	0.26	0.25	0.27	0.25	0.29
This instructor was knowledgeable and enthusiastic about the topic. Q119	0.19	0.19	0.22	0.20	0.23
This instructor effectively used examples/illustrations to promote learning. Q120	0.18	0.19	0.20	0.20	0.24
This instructor fostered questions and/or class participation. Q121	0.19	0.20	0.21	0.20	0.22
This instructor clearly explained important information/ideas/concepts. Q122	0.22	0.22	0.23	0.24	0.27
This instructor effectively used teaching methods appropriate to this class (e.g., critiques, discussion, demonstrations, group work). Q123	0.20	0.20	0.21	0.21	0.24
This instructor responded appropriately to questions and comments. Q124	0.20	0.20	0.21	0.21	0.24

This instructor stimulated student thinking and learning. Q125	0.18	0.19	0.19	0.21	0.23
This instructor promoted an atmosphere of mutual respect regarding diversity in student demographics and viewpoints, such as race, gender, or politics. Q126	0.13	0.13	0.13	0.14	0.15
This instructor was approachable and available for extra help. Q127	0.17	0.17	0.18	0.18	0.21
This instructor used class time effectively. Q128	0.19	0.20	0.19	0.21	0.23
This instructor helped students to be independent learners, responsible for their own learning. Q129	0.14	0.14	0.14	0.15	0.17
This instructor provided feedback that helped me improve my skills in this subject area. Q130	0.19	0.19	0.19	0.21	0.23
Constructs					
F1: Course Content and Structure	0.15	0.14	0.15	0.16	0.19
F2: Teaching Delivery	0.25	0.25	0.26	0.26	0.30
F3: Learning Environment	0.21	0.21	0.21	0.22	0.25
F4: Assessment	0.22	0.22	0.23	0.24	0.27
ICE Total Scale	0.23	0.23	0.24	0.24	0.28

Graduate Sample	Fall 2014	Spring 2015	Fall 2015	Spring 2016	Fall 2016
Sample size	7819	6725	7305	5858	6943
# of clusters	638	533	618	488	593
Average cluster size	12.255	12.617	11.82	12.004	11.708
Items					
This instructor taught effectively considering both the possibilities and limitations of the subject matter and the course (including class size and facilities). Q104	0.22	0.24	0.25	0.29	0.25
The syllabus clearly explained the course objectives, requirements, and grading system. Q111	0.17	0.20	0.16	0.24	0.18
Course content was relevant and useful (e.g., readings, online media, classwork, assignments). Q112	0.16	0.19	0.18	0.21	0.21
Resources (e.g., articles, literature, textbooks, class notes, online resources) were easy to access. Q113	0.12	0.14	0.15	0.15	0.14
This course challenged me. Q114	0.18	0.17	0.19	0.20	0.18
I was well-informed about my performance during this course. Q115	0.24	0.31	0.29	0.24	0.32
Assignments/projects/exams were graded fairly based on clearly communicated criteria. Q116	0.18	0.22	0.21	0.20	0.27
This instructor was consistently well-prepared. Q117	0.25	0.29	0.28	0.31	0.27
This instructor was audible and clear. Q118	0.21	0.21	0.24	0.24	0.22
This instructor was knowledgeable and enthusiastic about the topic. Q119	0.17	0.18	0.21	0.19	0.20

This instructor effectively used examples/illustrations to promote learning. Q120	0.18	0.21	0.21	0.24	0.22
This instructor fostered questions and/or class participation. Q121	0.20	0.20	0.23	0.26	0.20
This instructor clearly explained important information/ideas/concepts. Q122	0.21	0.24	0.24	0.27	0.27
This instructor effectively used teaching methods appropriate to this class (e.g., critiques, discussion, demonstrations, group work). Q123	0.21	0.20	0.23	0.26	0.22
This instructor responded appropriately to questions and comments. Q124	0.19	0.25	0.24	0.27	0.25
This instructor stimulated student thinking and learning. Q125	0.19	0.20	0.24	0.25	0.22
This instructor promoted an atmosphere of mutual respect regarding diversity in student demographics and viewpoints, such as race, gender, or politics. Q126	0.14	0.16	0.15	0.19	0.16
This instructor was approachable and available for extra help. Q127	0.17	0.24	0.20	0.23	0.20
This instructor used class time effectively. Q128	0.21	0.21	0.24	0.26	0.22
This instructor helped students to be independent learners, responsible for their own learning. Q129	0.14	0.14	0.14	0.17	0.17
This instructor provided feedback that helped me improve my skills in this subject area. Q130	0.20	0.25	0.23	0.23	0.24
Constructs					
F1: Course Content and Structure	0.18	0.21	0.20	0.24	0.22
F2: Teaching Delivery	0.24	0.25	0.28	0.30	0.27
F3: Learning Environment	0.21	0.23	0.24	0.28	0.24
F4: Assessment	0.23	0.30	0.28	0.25	0.32
ICE Total Scale	0.23	0.26	0.27	0.29	0.27

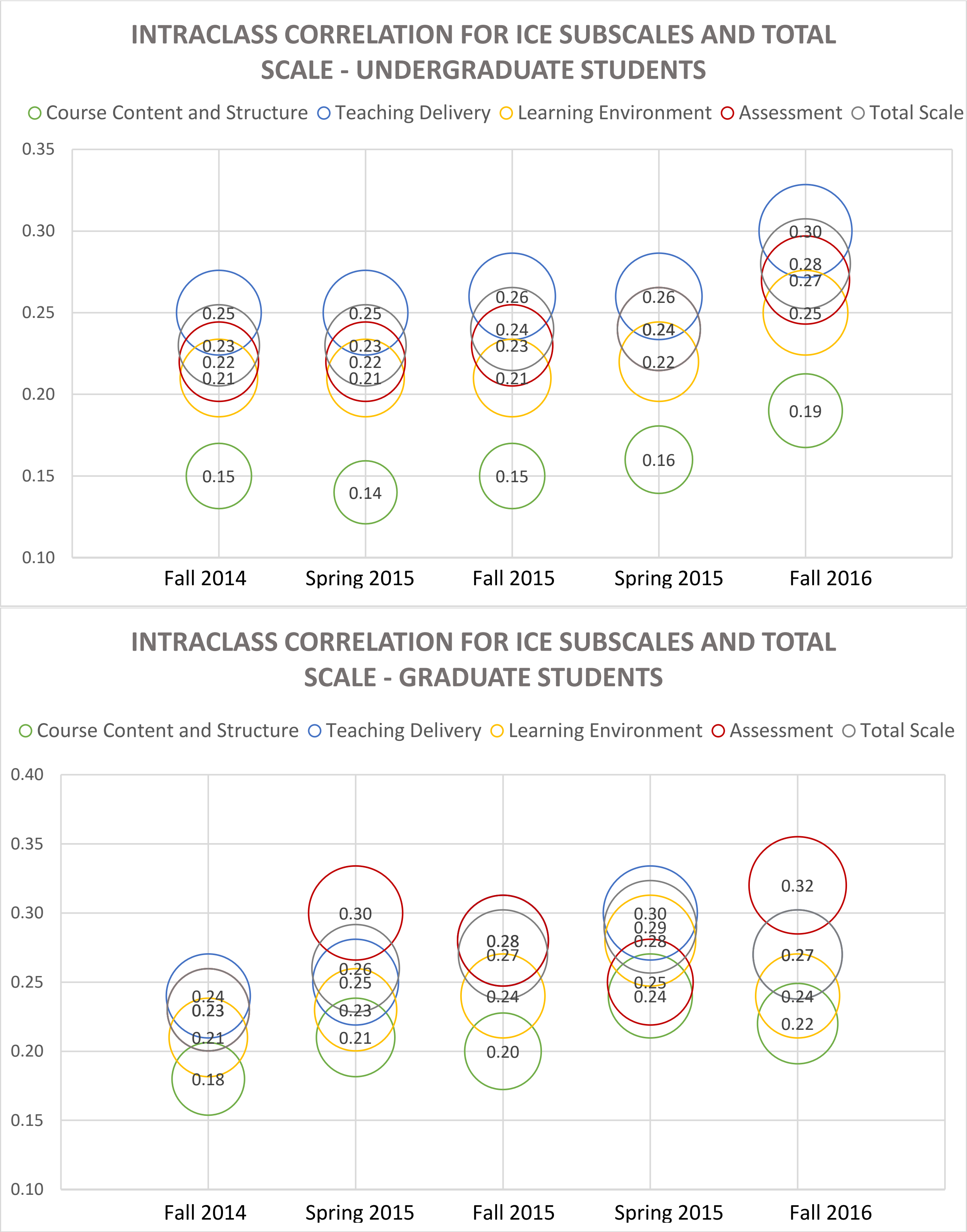


Figure 6. Intraclass correlations (ICCs) for ICE subscales and total scale by semester. Numbers and positions of circles indicate ICC values. Circle sizes represent average cluster sizes.

Relationships with Conceptually Related Constructs

To further collect validity evidence of the 20 items that measure the four key ICE constructs, their relationships with other conceptually related constructs were examined. Two related constructs/variables were used. The first is a general teaching effectiveness item "*This instructor taught effectively considering both the possibilities and limitations of the subject matter and the course (including class size and facilities).*" This item was rated on the same Liker-scale with response options *Strongly Disagree* (1), *Disagree* (2), *Neutral* (3), *Agree* (4), and *Strongly Agree* (5) as the scale for the 20 items measuring the key ICE constructs. This item had been used for MO SB389 in the past. The second related construct is the set of five current MO SB389 items, which asks students to report their recommendations to other students regarding five construct areas (class content, class structure, positive learning environment, instructor's teaching skill/style, and fairness of grading). The response options for the five MO SB389 items are Yes, No, and Don't know. For statistical analysis, only Yes and No responses were used.

The correlations between the four key ICE constructs and the general teaching effectiveness item were high (ranged from 0.81 to 0.89 for the graduate sample; and ranged from 0.76 to 0.89 for the undergraduate sample; see Table 7), suggesting that the general teaching effectiveness item may be used as an overall indicator of teaching effectiveness.

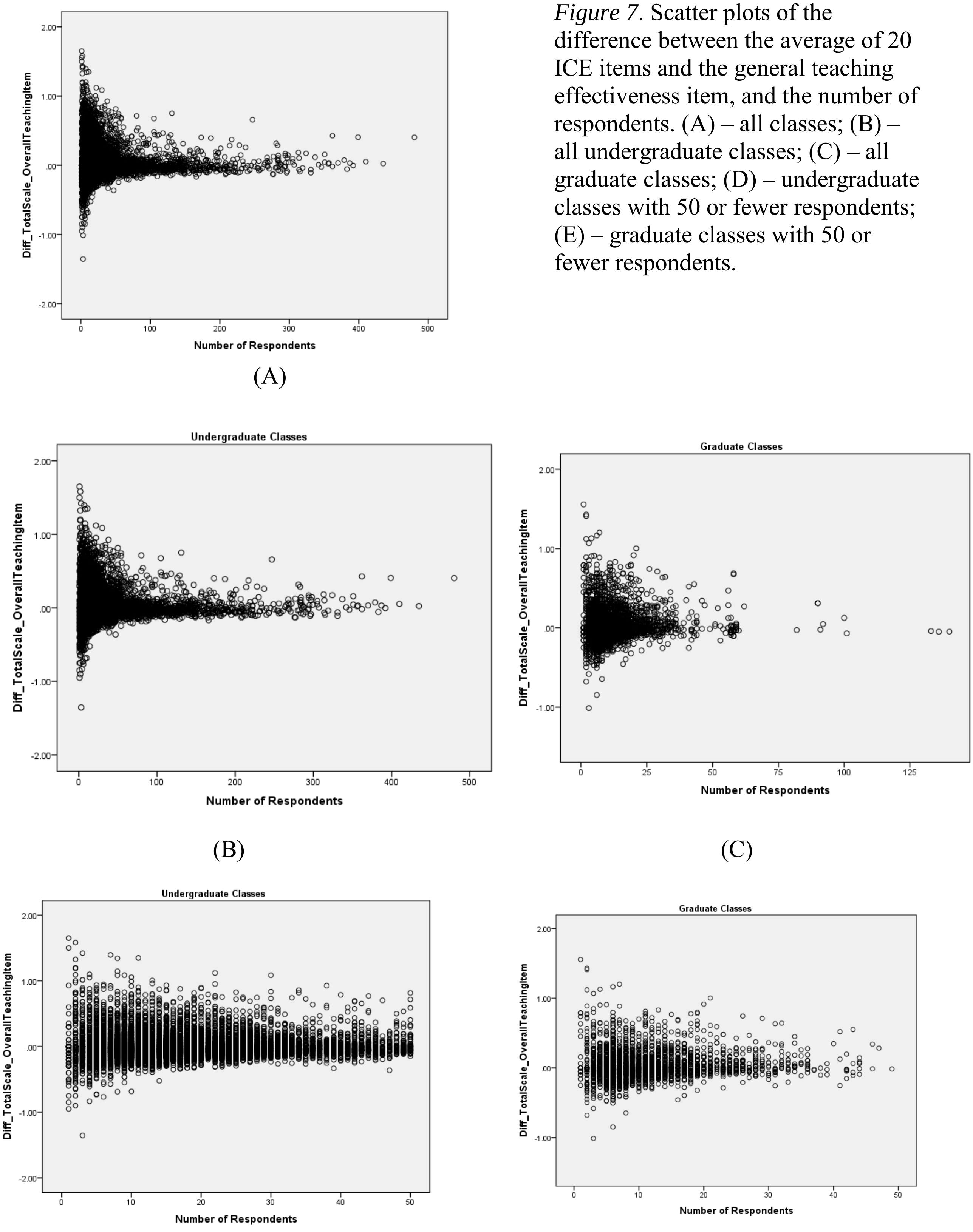
Table 7. Correlations between Key ICE Constructs and General Teaching Effectiveness Item

	F1	F2	F3	F4	Q104
F1: Course Content and Structure		0.83	0.84	0.90	0.76
F2: Teaching Delivery	0.87		0.95	0.90	0.89
F3: Learning Environment	0.87	0.96		0.93	0.88
F4: Assessment	0.88	0.88	0.91		0.87
Q104: This instructor taught effectively considering both the possibilities and limitations of the subject matter and the course (including class size and facilities).	0.81	0.89	0.88	0.84	

Note: Correlations for the graduate sample are shaded, and correlations for the undergraduate sample are not shaded.

Earlier analysis suggested that the 20 ICE items measure something in common. Therefore, another way to look at the relationships between the ICE items and the general teaching effectiveness item is to examine the difference between the average of the 20 ICE items and the general teaching effectiveness item for each class-instructor pair. Based on the data available for 16,148 unique combinations of class and instructor, such differences ranged from -1.35 to 1.65 with a mean of 0.0168. The 1st percentile difference was -0.3268 and the 99th percentile difference was 0.7000. Scatter plots (Figure 7) indicate that the largest discrepancies between the average of the 20 ICE items and the general teaching effectiveness item occurred for classes with fewer students who rated the instructor(s).

Figure 7. Scatter plots of the difference between the average of 20 ICE items and the general teaching effectiveness item, and the number of respondents. (A) – all classes; (B) – all undergraduate classes; (C) – all graduate classes; (D) – undergraduate classes with 50 or fewer respondents; (E) – graduate classes with 50 or fewer respondents.



(A)

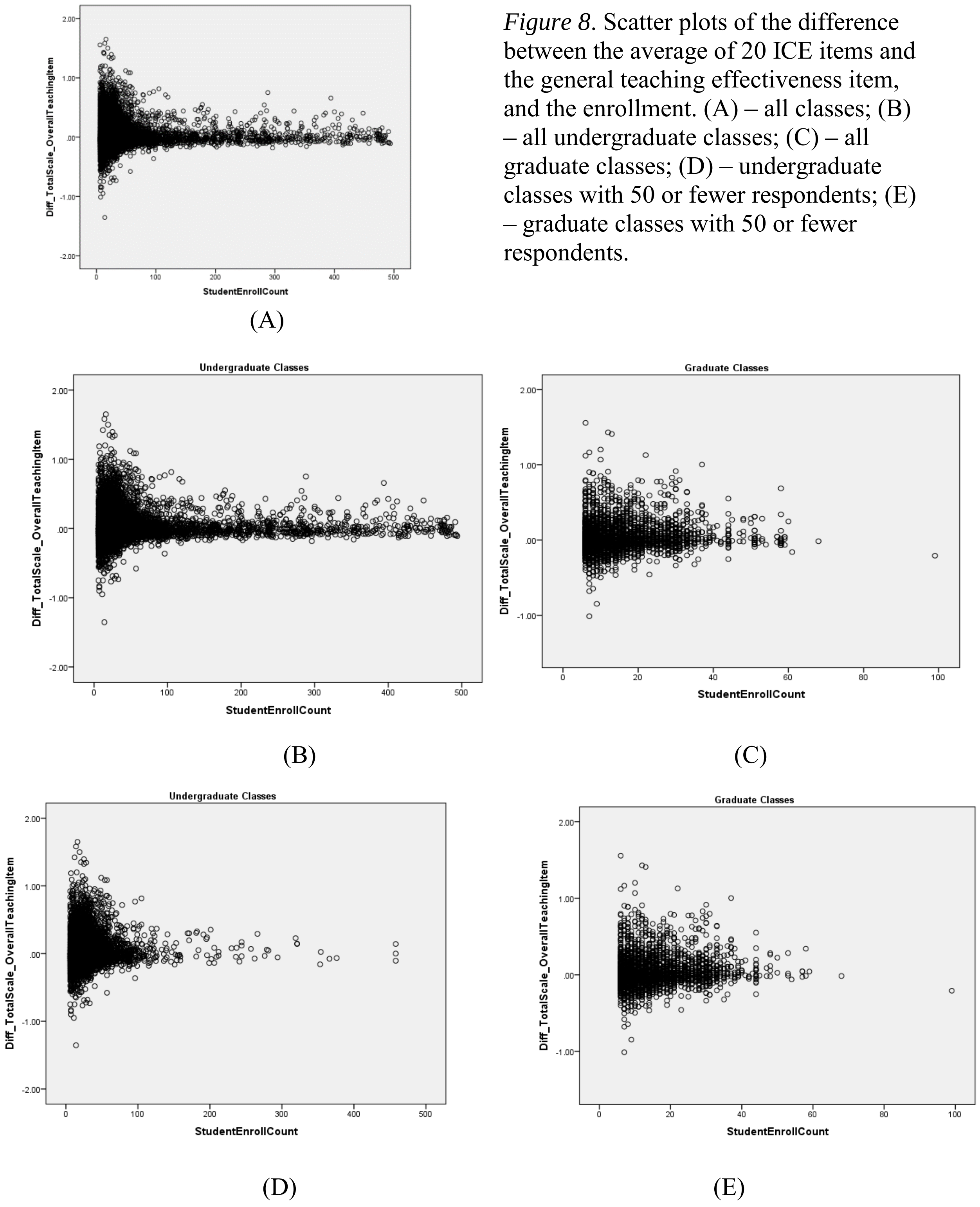
(B)

(C)

(D)

(E)

Another set of scatter plots focuses on relationships between the number of enrollment and the difference between the average of 20 ICE items and the general teaching effectiveness item (Figure 8). The largest discrepancies between the average of the 20 ICE items and the general teaching effectiveness item occurred for smaller classes.



(A)

(B)

(C)

(D)

(E)

Figure 8. Scatter plots of the difference between the average of 20 ICE items and the general teaching effectiveness item, and the enrollment. (A) – all classes; (B) – all undergraduate classes; (C) – all graduate classes; (D) – undergraduate classes with 50 or fewer respondents; (E) – graduate classes with 50 or fewer respondents.

These suggest that while the general teaching effectiveness item alone may be used for larger classes and classes with a high number of respondents, if it is used for smaller classes or classes with few respondents, the rating could be unreliable. We recommend that colleges and departments promote the importance of instruction and teaching evaluation in order to get a higher response rate.

There are five items for the MO SB389 requirement. While the majority of respondents answered "Yes" when asked whether they would recommend the class to other students regarding class content, class structure, positive learning environment, instructor's teaching skill/style, or fairness of grading, there were moderate and statistically significant correlations between the ICE key constructs and the MO SB389 questions. For Table 8, a "Yes" recommendation was coded 1 and a "No" recommendation was coded 2. Therefore, a negative correlation between a MO SB389 item and an ICE construct suggests a higher recommendation rate for classes and instructors rated higher on the construct. From Table 8, higher student ratings were associated with higher likelihood of recommending the class to other students.

Table 8. Correlations between Key ICE Constructs and MO SB389 Items

	F1	F2	F3	F4	Q105	Q106	Q107	Q108	Q109
F1: Course Content and Structure		0.83	0.84	0.91	-0.44	-0.49	-0.42	-0.50	-0.42
F2: Teaching Delivery	0.87		0.95	0.90	-0.41	-0.52	-0.50	-0.64	-0.42
F3: Learning Environment	0.87	0.96							-0.43
F4: Assessment	0.88	0.88	0.91						
Q105 Recommendation regarding Class Content	-0.53	-0.45	-0.44	-0.43					
Q106 Recommendation regarding Class Structure	-0.55	-0.56	-0.54	-0.56	0.49				
Q107 Recommendation regarding Positive Learning Environment	-0.44	-0.51	-0.54	-0.50	0.43	0.48			
Q108 Recommendation regarding Instructor's Teaching Skill/Style	-0.56	-0.66	-0.62	-0.61	0.49	0.68	0.58		
Q109 Recommendation regarding Fairness Of Grading	-0.41	-0.43	-0.45	-0.60	0.36	0.43	0.52	0.48	

Note: Correlations for the graduate sample are shaded, and Correlations for the undergraduate sample are not shaded.

Relationships with Instructor and Class Information

Further, relationships between students' ratings and instructor and class information were examined. Specifically, student average rating for each class/instructor, the standard deviation of student ratings for each class/instructor, the sex of the instructor, and the class average GPA were used.

The numbers of classes for which a male or female instructor were rated by semester and by course level (undergraduate vs. graduate) are in Table 9 and Figure 9.

Table 9. Number of Classes by Instructor's Sex by Semester and by Course Level

Undergraduate	Male	Female	Total
Fall 2014	1364 (53.7%)	1176 (46.3%)	2540
Spring 2015	1293 (54.4%)	1084(45.6%)	2377
Fall 2015	1372 (54.2%)	1159 (45.8%)	2531

Spring 2016	1298 (54.8%)	1069 (45.2%)	2367
Fall 2016	1344 (54.3%)	1129 (45.7%)	2473
Total	6671 (54.3%)	5617 (45.7%)	12288
Graduate	Male	Female	Total
Fall 2014	358 (59.6%)	243 (40.4%)	601
Spring 2015	267 (54.5%)	223 (45.5%)	490
Fall 2015	303 (52.8%)	271 (47.2%)	574
Spring 2016	239 (53.2%)	210 (46.8%)	449
Fall 2016	279 (50.3%)	276 (49.7%)	555
Total	1446 (54.2%)	1223 (45.8%)	2669

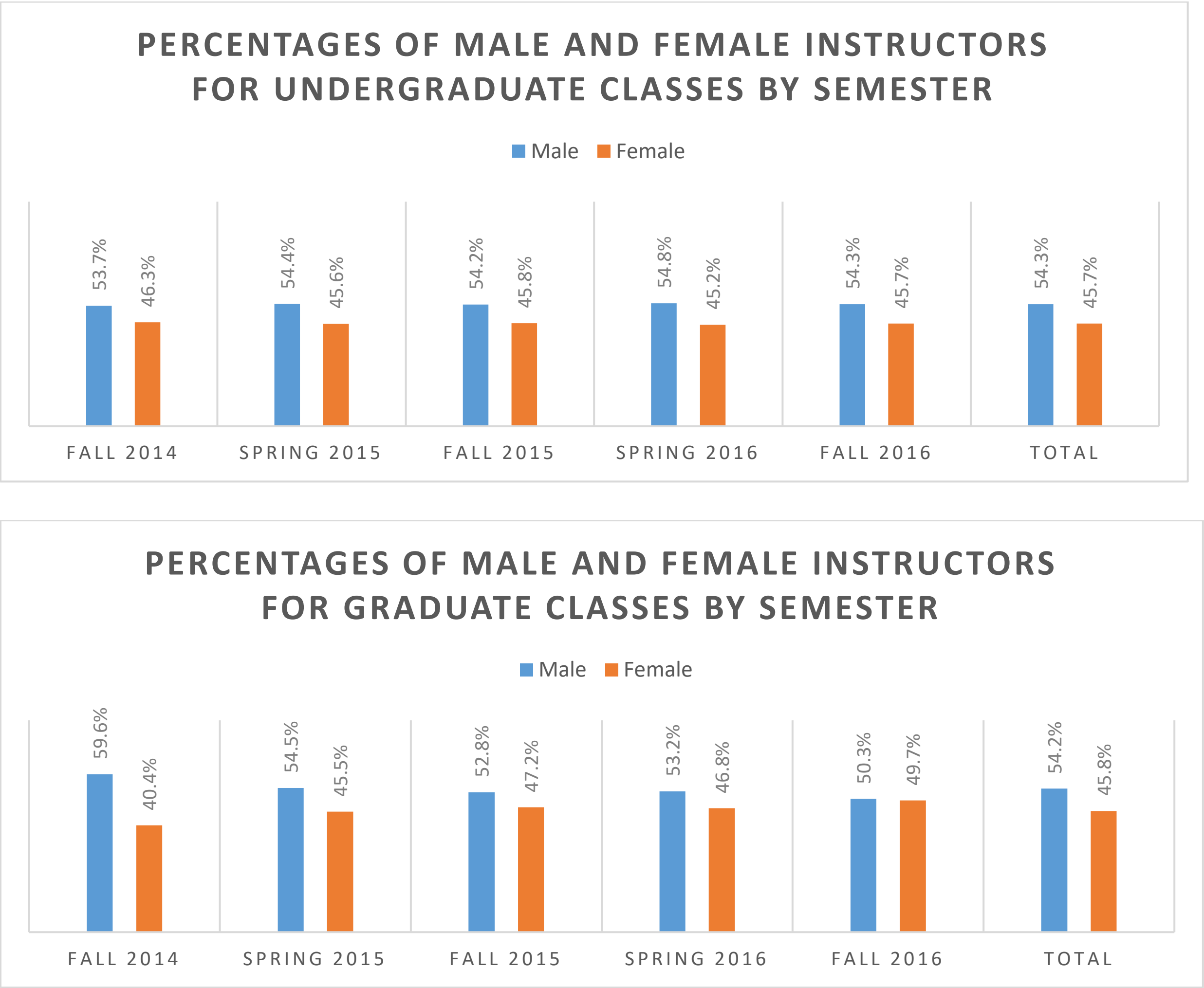


Figure 9. Percentages of male and female instructors for undergraduate and graduate classes by semester.

Correlations between student average ratings on ICE constructs (four key constructs and the total scale), instructor's sex and class average GPA are in Table 10. Separate analysis by semester was also conducted and the results were similar across semesters. A positive correlation between an ICE construct and the instructor's sex variable means that a female instructor had a higher value on the ICE construct; and a negative correlation means that a male instructor had a higher value.

Despite some statistically significant correlations between student ratings and instructor's sex, which could be due to the large sample sizes, none of these correlations reached a magnitude of 0.10. In addition, independent samples t-test results showed very small differences in average student ratings between classes taught by male and female instructors despite some statistically significant differences (see Table 11). The differences between ratings for male and female instructors ranged from -0.09 to 0.02. These results suggest that instructor's sex was not strongly related to student rating of teaching. However, this is not to say that there was no gender bias since we are not sure if teaching effectiveness is truly equal between male and female instructors.

Similarly, the correlations between student ratings and class average GPA were very small, despite some statistically significant ones. The highest correlation was between student average rating on Teaching Delivery and class average GPA for undergraduate classes at 0.154. These results suggest that class average GPA was not strongly related to student ratings of teaching.

Table 10. Correlations between ICE Constructs, Instructor's Sex and Class Average GPA

	Undergraduate Sample		Graduate Sample	
	Instructor's Sex	Class Average GPA	Instructor's Sex	Class Average GPA
Course Content and Structure _Mean	.057**(12284)	.090** (13192)	0.020 (2669)	0.008 (2864)
Teaching Delivery_Mean	.083** (12281)	.154** (13189)	-0.026 (2664)	.066** (2859)
Learning Environment_Mean	.076** (12281)	.115** (13189)	-0.015 (2664)	.047* (2859)
Assessment_Mean	.098** (12285)	.114** (13193)	.057** (2669)	0.018 (2864)
Total Scale_Mean	.084** (12285)	.133** (13193)	-0.001 (2669)	.044* (2864)
Course Content and Structure _SD	-.048** (12194)	-.046** (13098)	-0.02 (2628)	-0.024 (2823)
Teaching Delivery_SD	-.080** (12185)	-.140** (13089)	0.008 (2631)	-.039* (2825)
Learning Environment_SD	-.067** (12186)	-.099** (13090)	-0.009 (2632)	-0.022 (2826)
Assessment_SD	-.089** (12197)	-.084** (13101)	-.073** (2637)	-0.007 (2832)
Total Scale_SD	-.070** (12197)	-.092** (13101)	-0.013 (2637)	-0.033 (2832)

Note. Sample sizes are in parentheses. $^{*}p<.05$, $^{**}p<.01$

Table 11. T-Test Results for Differences Between Male and Female Instructors in Average Student Ratings

	Male Instructor		Female Instructor		
Undergraduate	Mean	SD	Mean	SD	t
F1: Course Content and Structure	4.35	0.32	4.39	0.33	-6.27***
F2: Teaching Delivery	4.38	0.44	4.45	0.42	-9.23***
F3: Learning Environment	4.41	0.39	4.47	0.39	-8.40***

F4: Assessment	4.19	0.48	4.28	0.47	-10.87***
Total Scale	4.36	0.38	4.42	0.38	-9.30***
	Male Instructor		Female Instructor		
Graduate	Mean	SD	Mean	SD	t
F1: Course Content and Structure	4.47	0.36	4.48	0.37	-1.051
F2: Teaching Delivery	4.51	0.42	4.49	0.47	1.319
F3: Learning Environment	4.53	0.39	4.52	0.44	0.745
F4: Assessment	4.28	0.53	4.34	0.55	-2.95**
Total Scale	4.48	0.38	4.48	0.42	0.028

Note: $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$

Another thought is that maybe student's gender matters when they rated a male or female instructor (**note**: in this report, we use gender and sex interchangeably although they are not the same. Also, for student gender, we only used Male and Female genders due to the relatively small number of students who chose other gender categories). For example, a male student might give a higher rating to a male instructor than to a female instructor; or a male instructor may receive higher ratings from male students than from female students.

Two sets of analyses were conducted. For the first set, we calculated correlations between instructor's sex and the four ICE key constructs, separately for when the student gender and instructor's sex were opposite and for when the student gender and instructor's sex were the same. A positive correlation would mean that the female instructor was rated higher on the construct. Results are in Table 12. For the undergraduate sample, when student gender and instructor's sex were opposite (i.e., a male student rating a female instructor or a female student rating a male instructor), female instructors received lower ratings on F1 "Course Content and Structure" and higher ratings on F4 "Assessment," compared to male instructors; when student gender and instructor's sex were the same, female instructors received higher ratings on all four constructs. For the graduate sample, the only statistically significant result (at alpha=.05) was that when student gender and instructor's sex were opposite, female instructors received higher ratings on F4 "Assessment," compared to male instructors. However, all correlations were very small (magnitude ranging from 0.000 to 0.071), suggesting that the instructor's sex was not strongly related to student rating of teaching when the student's gender is the same as or opposite to the instructor's sex.

For the second set of analysis, we calculated correlations between student gender and the four ICE key constructs, separately for male instructors and female instructors. A positive correlation would mean that female students gave a higher rating compared to male students. Results are in Table 13. For the undergraduate sample, when the instructor was male, female students gave higher ratings on F1 "Course Content and Structure" and lower ratings on F4 "Assessment," compared to male students; when the instructor was female, female students gave higher ratings on all four ICE constructs, compared to male students. For the graduate sample, at the statistically significance level of .05, when the instructor was male, female students gave lower ratings on F1 "Course Content and Structure," F2 "Teaching Delivery," and F4 "Assessment," compared to male students; when the instructor was female, male and female students gave similar ratings. However, even the statistically significant correlations were small (highest was in the magnitude of .048), suggesting that student gender was not strongly related to their rating for classes taught by a male instructor or a female instructor.

Table 12. Correlations between Instructor's Sex and Four ICE Constructs

When student gender and instructor's sex are opposite	F1	F2	F3	F4	Instructor's Sex
F1: Course Content and Structure		0.818^{***}	0.827^{***}	0.905^{***}	-0.026^{**}
F2: Teaching Delivery	0.849^{***}		0.946^{***}	0.894^{***}	-0.002
F3: Learning Environment	0.848^{***}	0.954^{***}		0.922^{***}	-0.004
F4: Assessment	0.862^{***}	0.859^{***}	0.894^{***}		0.032^{**}
Instructor's Sex	-0.016	-0.021	-0.016	0.039^{*}	
When student gender and instructor's sex are the same	F1	F2	F3	F4	Instructor's Sex
F1: Course Content and Structure		0.832^{***}	0.842^{***}	0.908^{***}	0.068^{***}
F2: Teaching Delivery	0.874^{***}		0.951^{***}	0.903^{***}	0.071^{***}
F3: Learning Environment	0.874^{***}	0.957^{***}		0.930^{***}	0.069^{***}
F4: Assessment	0.884^{***}	0.883^{***}	0.916^{***}		0.058^{***}
Instructor's Sex	-0.016	-0.033	-0.027	0.000	

Note: Correlations for the graduate sample are shaded, and Correlations for the undergraduate sample are not shaded. $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$

Table 13. Correlations between Student Gender and Four Key ICE Constructs

When instructor was male	F1	F2	F3	F4	Student Gender
F1: Course Content and Structure		0.819^{***}	0.83^{***}	0.907^{***}	0.014^{**}
F2: Teaching Delivery	0.856^{***}		0.947^{***}	0.895^{***}	0.000
F3: Learning Environment	0.853^{***}	0.951^{***}		0.921^{***}	0.004
F4: Assessment	0.864^{***}	0.852^{***}	0.891^{***}		-0.022^{***}
Student Gender	-0.025^{*}	-0.028^{*}	-0.023	-0.048^{***}	
When instructor was female	F1	F2	F3	F4	Student Gender
F1: Course Content and Structure		0.834^{***}	0.843^{***}	0.907^{***}	0.078^{***}
F2: Teaching Delivery	0.873^{***}		0.952^{***}	0.903^{***}	0.070^{***}
F3: Learning Environment	0.876^{***}	0.962^{***}		0.933^{***}	0.066^{***}
F4: Assessment	0.890^{***}	0.898^{***}	0.926^{***}		0.044^{***}
Student Gender	0.022	0.014	0.011	0.005	

Note: Correlations for the graduate sample are shaded, and Correlations for the undergraduate sample are not shaded. $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$

References

American Education Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing* Washington, DC: Author.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464-504. doi: 10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255.

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54*, 22-42. doi: http://dx.doi.org/10.1016/j.stueduc.2016.08.007