

A Renewed Validation Study of the University of Toronto's Cascaded Course Evaluation Framework

Study of Institutional Items

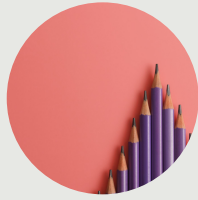
Evaluation & Assessment
Center for Teaching Support & Innovation

February 20, 2025



UNIVERSITY OF
TORONTO

CENTRE FOR TEACHING SUPPORT & INNOVATION



Published by

The Centre for Teaching Support & Innovation (CTSI)
University of Toronto

130 St. George Street
Robarts Library, 4th Floor
Toronto, ON M5S 3H1

Phone: (416) 946-3139

Email: ctsi.teaching@utoronto.ca

Website: www.teaching.utoronto.ca

Table of Contents

Executive Summary	5
Scope and Data Sample	5
What are Course Evaluation Surveys Measuring?	5
The Framework: Kane's Validity Framework	5
Key Findings of the Validation Study	6
Implications for Interpreting Course Evaluation Results	9
Conclusion	9
Section 1: Introduction	10
Background	10
Purpose: Why Conduct the Renewed Validation Study Now?	11
Objective	11
The Validity Framework Applied to Course Evaluations in this Study	11
Data Used in This Study	13
Section 2. Scoring	17
2.1 Survey Content, Design, and Administration	17
2.2. Item Responses and Scoring	21
2.3 Evidence of Comparability	23
2.4. Scoring Inference Summary	28
Section 3. Generalization	29
3.1 Response Rates and Representativeness of the Sample	29
3.2. Inter-Rater Reliability, Test-Retest Reliability, and Inter-Item Reliability	34
3.3. Generalizability Theory	36

3.4. Generalization Inference Summary _____	40
Section 4. Extrapolation _____	41
4.1. Multilevel Internal Structure (Dimensionality) _____	41
4.2. Extrapolation Inference Summary _____	43
Section 5. Implications _____	44
Section 6. Conclusion _____	46
Final Notes _____	47
References _____	48
Appendices _____	50
Appendix A: Institutional Items in the Cascaded Course Evaluation Framework _	50
Appendix B: Supplemental Response Rate Metrics _____	51

Executive Summary

Since 2012, the University of Toronto (U of T) has utilized a Cascaded Course Evaluation Framework (CCEF) to gather student feedback on courses¹. In 2018, the Centre for Teaching Support & Innovation (CTSI) conducted a Validation Study to assess the reliability and validity of the Institutional Composite Mean (ICM) used in this framework (CTSI, 2018a). The 2018 Validation Study analyzed 277,498 surveys from courses from 2015/16 to 2016/17 in four major undergraduate divisions, confirming the ICM's reliability and validity as a key metric in evaluating teaching and measuring student experience in their courses. This Renewed Validation Study revisits the validity and reliability of the CCEF in the context of current student demographics, evolving educational technology, the impact of the COVID-19 pandemic, and expanded use of the CCEF at the university.

Scope and Data Sample

The current Renewed Validation Study expands the scope of the 2018 study to include data from the 15 Faculties/schools using the centralized course evaluation system, including both undergraduate and graduate courses. This study considers the development of U of T survey items and uses data over the five years from 2018/19 through 2022/23, encompassing 967,817 completed surveys. This broader dataset allows for a more comprehensive analysis of the CCEF across different contexts within the university. In addition to investigating the ICM, this Renewed Validation Study investigates the performance of institutional items 1 through 6 (Ins 1 to Ins 6) and the response rates of institutional items 7 and 8 (Ins 7 to Ins 8; Appendix A).

What are Course Evaluation Surveys Measuring?

Key studies in the course evaluation literature indicate that course evaluations are designed to measure student-reported learning experiences and are not a measure of teaching quality, teaching effectiveness, or student satisfaction (Dyer & Donnelly-Hermosillo, 2024; Spooen et al., 2013). Following this research and the process of the development of the U of T course evaluation survey, it is proposed that the underlying construct measured by course evaluations within the U of T context is student-reported learning experiences of key teaching and learning priorities at the University of Toronto.

The Framework: Kane's Validity Framework

This study uses Kane's validity framework to guide the collection and evaluation of validity evidence for the CCEF's intended interpretations and use. Kane's framework provides a comprehensive perspective on the different facets of validity including the four inferences to validity (*Scoring*, *Generalization*, *Extrapolation*, and *Implications*) and ensures a thorough examination of the course evaluation

¹ Throughout this Renewed Validation Study, the term "courses" is used to refer to course-sections.

instrument's performance. In Kane's framework, validation is a process of accumulating evidence that enables us to state or refute a validity argument (Kane 2006).

In adapting Kane's framework to the course evaluation context, this study followed a model of a four-phase process. In the first phase, scoring is conceptualized as the process of survey respondents translating their individual experience in a course into responses to items contained in the course evaluation surveys. In the second phase, generalization, course evaluation results of one course at a point in time are generalized to a snapshot of possible responses of all students in the course over a universe of possible survey items. In the third phase, extrapolation, it is inferred that course evaluation results reflect students' actual experience within the course. The fourth phase, implications, is conceptualized as the possibility of applying the course evaluation results to inform decisions. Decisions can involve instructors' formative, iterative improvement (e.g., lesson plan revision) or high-stakes decisions about teaching (e.g., faculty hiring, promotion and merit assessment).

Key Findings of the Validation Study

This section provides evidence for the four levels of inference of Kane's argument-based approach to validation by a list of findings. The key findings and evidence are organized into the four key inferences of Kane's validity framework.

Scoring Inference: Do Survey Responses Reflect Students' Learning Experiences?

The findings from this study support the conclusion that students' survey response accurately reflect their experiences of the U of T teaching and learning priorities in a course.

1. **Survey Items Reflecting Teaching Priorities:** The institutional survey items were carefully designed to align with key teaching and learning priorities at the institutional level. This alignment was achieved in their development through extensive consultations with the stakeholder groups.
2. **Minimizing Irrelevant Influences:** Efforts were made to reduce any irrelevant factors influencing students' responses, by minimizing unclear wording in survey items and inconsistencies in survey administration.
3. **Clear and Meaningful Response Options:** The findings from a recent focus-group study at U of T (Gibbs et al., 2023) suggest students understood the response scales and found them relevant to their experiences.
4. **Active Participation:** Analysis of response patterns provided proxy evidence that students were engaged in providing feedback.
 - Only 6% of course evaluation surveys selected the same response consistently to all quantitative items.
 - Most submitted surveys included responses to the two open-ended questions (Ins 7 and Ins 8), with 81% of surveys responding to Ins 7 and 81% of surveys responding to Ins 8.
 - Nearly all submitted surveys included responses to Likert-scale items Ins 1 to Ins 6 (over 99% completion).

5. Inter-Item Relationships: The findings suggest a strong relationship between institutional items that make up the ICM.

- Items were strongly related to each other at the course level (Spearman's rho correlations range from .78 to .92).
- Items were strongly related to the mean of other ICM items at the course level (Spearman's rho correlations range from .89 to .93).

6. Stability of the ICM Scores Over Time: ICM scores have remained stable across the university from 2018/19 to 2022/23, showing that the measure is not overly influenced by temporary factors. Student-reported learning experiences from courses were also largely positive.

- The average ICM had a small increase from 4.0 to 4.1 when comparing the 2018 Validation Study to the Renewed Validation Study using comparable datasets.
- The average ICM had a small increase from 4.1 to 4.2 when comparing 2018/19 to 2022/23.

7. Context-Sensitive Variations: The data show that different instructional contexts (e.g., large courses vs. small courses) have distinct ICM profiles, which indicates the measure's sensitivity to contextual differences. In particular, variations in the ICM scores by course sizes are observed.

- Larger courses are associated with smaller values of the ICM (Spearman's rho = -0.40).
- While courses with multiple instructors and undergraduate-level courses have slightly smaller values of the ICM compared to single-instructor (Cohen's $d = .36$) and graduate-level courses (Cohen's $d = .36$), both effects become negligible when controlling for course size.

Generalization Inference: How Well Do Survey Results Capture the Overall Picture of Students' Experience in a Course?

The generalization inference extends the interpretation of individual student responses as a snapshot of a sample of possible responses from all students in the course. The findings support the following conclusions:

8. Precision for the Majority of Courses: The ICM is a precise measure for representing the overall picture of student-reported learning experiences in the majority of courses (particularly medium to large-size courses).

- The average course-level response rate for all courses was 40%.
- The average course-level response rate for undergraduate and graduate courses was 36% and 53%, respectively.
- 34% of evaluated courses provide somewhat precise to very precise estimates of the ICM (width of interval around the mean ≤ 0.49).
- 69% of evaluated courses provide general to very precise estimates of the ICM (width of interval around the mean ≤ 0.99).

9. Consistency Across Students and Time: There is strong agreement among students within a course (inter-rater reliability) and course ICM values remain stable when measured at different times by different students (test-retest reliability).

- Inter-rater reliability among students within the same course was good ($ICC(k) = 0.86$).
- Test-retest reliability for an instructor teaching the same course at different times was moderate ($ICC(k) = 0.72$).

10. Consistency Among Institutional Items: The inter-item reliability of the five institutional items that comprise the ICM is very strong, indicating that the core items reliably measure the same underlying construct.

- Inter-item reliability for institutional items Ins 1 to Ins 5 was strong (Cronbach's alpha = .92).
- The removal of any institutional item decreases Cronbach's alpha, suggesting each item positively contributes to the reliability of the ICM.

11. Generalizability in Representing Student Experiences: Generalizability-study variance decomposition results suggest that students, courses, and items contribute to variance in course evaluation scores in a manner aligned with the course evaluation literature (Chang & Hocevar, 2000; Curby et al., 2020; Spooren et al., 2014).

- The largest source of variance in institutional items was found in students nested within courses (53.0%), which supports the notion that course evaluations measure students' individualized learning experiences.
- A large source of variance in institutional items was contributed by courses (16.8%), suggesting that course evaluation results capture students' experiences across courses.
- 14 to 18 student responses (depending on course size) were sufficient to achieve a reliability of .80 for the ICM.

Extrapolation Inference: Do Survey Results Reflect Students' Actual Learning Experiences in the Course?

The extrapolation inference explores whether students' responses to course evaluation surveys reflect their learning experiences in a course. Findings using multilevel factor analysis suggest that the five institutional items comprising the ICM have a good internal structure and excellent internal consistency. In the U of T context, available data are limited linking course evaluation results to broader measures of students' learning experience.

12. Good Internal Structure: the first five institutional items (Ins 1 to Ins 5) have psychometric characteristics that justify combining them into the ICM.

- Findings from Multilevel Exploratory Factor Analysis indicate that the five items measure a single underlying construct (unidimensionality).
- Institutional items Ins 1 to Ins 5 are strongly related to the ICM (standardized course-level factor loadings all above .90).
- Institutional items demonstrate good item reliability (course-level item communalities are all above .60).
- The course-level ICM demonstrates excellent reliability above .90.

Implications Inference: Using Course Evaluation Results to Make Decisions

The implications inference involves the potential for course evaluation results to guide decisions about instructors' formative, iterative improvement and high-stakes decisions about teaching (e.g., faculty hiring, promotion and merit assessment). This study suggests the CCEF reflects student responses to teaching and learning priorities (scoring inference), an overall picture of students' experiences (generalization inference), and good internal structure of Ins 1 to Ins 5 (extrapolation inference), but the data and analyses considered have not been connected to decision-making (implication inference). Consequently, course evaluation results should not be used as the sole data source for high-stakes decisions. A multi-faceted approach to evaluating teaching incorporating various data sources and measures is necessary,

as recommended by the *University of Toronto Provostial Guidelines on the Student Evaluation of Teaching in Courses* (Office of the Vice President and Provost, 2022).

Implications for Interpreting Course Evaluation Results

Results from this study have been used to inform the development of a [Step-by-Step Guide to Reviewing Course Evaluations](#) (CTSI, 2025) for instructors and administrators. The interpretation guide recommends instructors take a reflective practitioner approach to consider the context of their course. It also encourages administrators to consider the implications of contextual factors, such as course size and multi-instructor courses, when interpreting course evaluation results. The guide details how the precision and reliability of course evaluation results vary as a function of response rate and the number of responses received. It provides guidance on the response rates necessary for reliable and precise results across various course sizes. Additionally, the guide provides updated typical values of the ICM, and institutional item endorsement rates grouped by course size and based on data from 2018/19 to 2022/23. General guidance for interpreting open-ended qualitative items (Ins 7 and Ins 8) is also included. For further details, please refer to the Step-by-Step Guide available on the CTSI website.

Conclusion

The Renewed Validation Study provides evidence supporting the use of CCEF as a tool for understanding student individualized learning experiences in courses at the University of Toronto. The findings from this study offer a basis for ongoing assessment and further analysis to enhance the application of the CCEF in the context of the University of Toronto.

Section 1: Introduction

Background

In 2012, the University of Toronto (U of T) began implementation of the Cascaded Course Evaluation Framework (CCEF) to collect student feedback about their learning experiences in courses. (See Appendix A for institutional items used in the CCEF, which are the focus of this report.) In 2018, a validation study was completed by researchers at the Centre for Teaching Support & Innovation (CTSI) to assess the reliability and validity of the Institutional Composite Mean (ICM²), a key metric in the CCEF (CTSI, 2018a).

The 2018 Validation Study provided evidence to support the construct validity of the ICM and recommendations for the interpretation of the ICM. The most significant impact of the study is that it established the reliability, aspects of validity, and generalizability of the use of ICM in U of T's teaching and learning contexts. It established grounds for the ICM as a key metric of course evaluation data and support for its use as one component of the evaluation of teaching effectiveness. The data used in the 2018 Validation Study included 277,498 completed evaluation surveys across two academic years from the single-instructor taught courses from the divisions with the largest number of undergraduate courses: Applied Science & Engineering (APSC), Faculty of Art & Science (FAS), University of Toronto Mississauga (UTM), and University of Toronto Scarborough (UTSC).

This Renewed Validation Study expands the scope of the 2018 Validation Study to include data from all 15 Faculties/schools using the centralized course evaluation system. This study considers the development of the institutional U of T survey items and uses responses from course evaluation surveys over the five years from 2018/19 through 2022/23, encompassing 967,817 completed surveys. This broader dataset allows for a more comprehensive analysis of the CCEF across different contexts within the university. Table 1 summarizes the differences in scope between the 2018 Validation Study and this Renewed Validation Study.

Table 1. A Comparison of the Data Used in the Two Validation Studies

	2018 Validation Study	This Renewed Validation Study
How many academic years of data?	2 years (2015/16 and 2016/17)	5 years (2018/19 to 2022/23)

² The ICM is the composite score of the average of five core institutional items (institutional questions Ins1 to Ins5).

COVID-19	Pre-COVID-19	During- and post-COVID-19
What are the types of courses included?	Single-instructor courses only	Single- and multi-instructor courses
Which divisions were included?	Undergraduate APSC, FAS, UTM, and UTSC courses	All 15 Faculties/schools that implemented CCEF
What are the types of analyses used?	Classical test theory, descriptive statistics, and exploratory factor analysis	Classical test theory, descriptive statistics, generalizability theory, and multilevel factor analysis.
What is the number of survey responses?	277,498 survey responses	967,817 survey responses

Note: Faculty of Applied Science & Engineering (APSC), Faculty of Art & Science (FAS), University of Toronto Mississauga (UTM), and University of Toronto Scarborough (UTSC)

Purpose: Why Conduct the Renewed Validation Study Now?

It has been over a decade since the university embarked on implementing the CCEF. In the past decade, the use of CCEF in divisions has grown. This study represents the ongoing work CTSI carries out to improve our institutional understanding of course evaluation metrics, applying contemporary approaches in the psychometrics and measurement field.

Objective

The main objective of this study is to build on the work of the 2018 Validation Study by examining a larger sample of all course evaluation surveys completed in the 5 years from 2018/19 through 2022/23. This study:

1. Applies a contemporary approach of psychometric analysis guided by Kane's validity framework.
2. Provides an overview of values of key metrics of course evaluations from 2018/19 to 2022/23, including response rate, ICM, and endorsement rate.
3. Informs practical recommendations for the interpretation of course evaluation results. Results from this study have been used to inform the development of [a Step-by-Step Guide to Reviewing Course Evaluations](#) (CTSI, 2025) for instructors and administrators.

The Validity Framework Applied to Course Evaluations in this Study

This Renewed Validation Study applies Kane's validity framework (Kane, 1992, 2006, 2013) to support the interpretation of validity evidence. Kane's framework provides a comprehensive perspective on the

different facets of validity including the four inferences to validity (*Scoring, Generalization, Extrapolation, and Implications*) and ensures a thorough examination of the course evaluation instrument's performance. In Kane's view, validation is a process of accumulating evidence that enables us to state or refute a validity argument.

Kane's framework can be adapted to the course evaluation framework following a four-phase process (Figure 1). Scoring, the first phase, is conceptualized as the process of student respondents translating their individual experiences in a course into responses to items contained in the course evaluation surveys. Generalization, the second phase, involves how course evaluation results of one course at a point in time are generalized to a snapshot of possible responses of all students in the course over a universe of possible survey items. Extrapolation, the third phase, infers that course evaluation results reflect students' actual experience within the course. Implications, the fourth phase, is conceptualized as the possibility of applying the course evaluation results to inform decisions. Decisions can involve instructors' formative, iterative improvement (e.g., lesson plan revision) or high-stakes decisions about teaching (e.g., faculty hiring, promotion and merit assessment).

The sections that follow are organized by findings for the four inferences to validity. Sections 2, 3, and 4 of this report examine the assumptions and evidence underlying the scoring, generalization, and extrapolation inferences. Implication inference is discussed in Section 5.

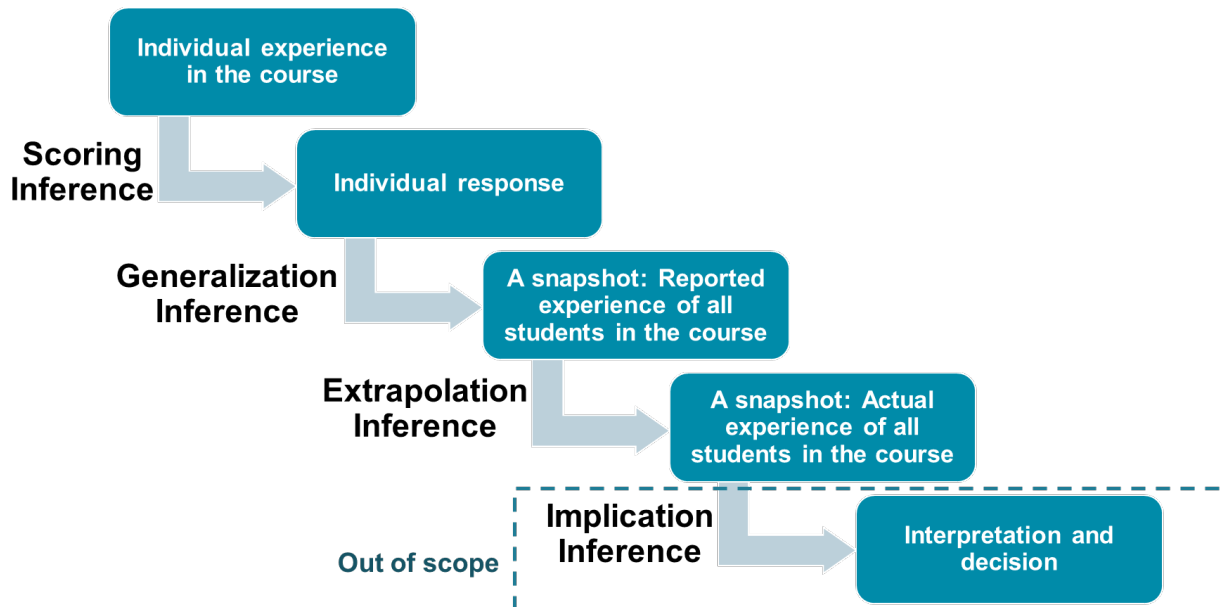


Figure 1. Key Elements in Kane's Argument-Based Approach to Validation in the Course Evaluation Context.

Data Used in This Study

This study used responses to course evaluation surveys received during the five years from the fall semester of 2018/19 to the summer semester of 2022/23 across 15 Faculties/schools³. The entire data set comprises 967,817 completed course evaluation surveys collected from 56,846 course-sections. Additional statistics about the data are provided in Figure 2.

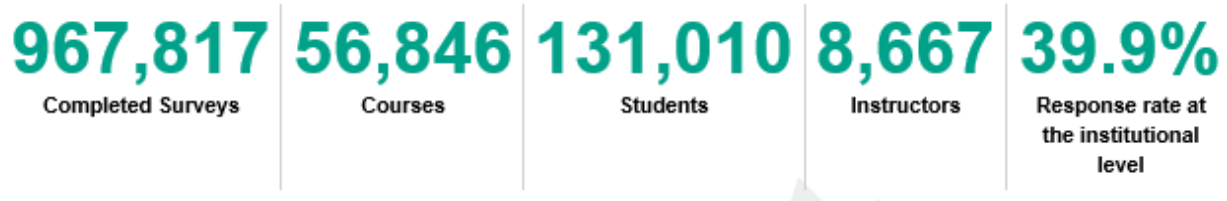


Figure 2. Descriptive Statistics About the Data.

Table 2 gives detailed information about the data considered in this study. Most of the completed surveys were from undergraduate courses (86.3%), St George Campus (64.1%), half-credit (89.4%), and single-instructor courses (90.3%). In terms of the course sizes, more than half of the surveys were from courses with 100 students or fewer (56.7%) and the remaining were from courses with 101-200 students and more than 200 students (20.8% and 22.6%, respectively). As the five-year period overlapped with the COVID-19 pandemic, about one-third of evaluations in the data sample were from courses taught in an online or hybrid environment⁴.

Undergraduate courses within the four divisions offering the largest number of undergraduate courses, Faculty of Art & Science (FAS), Faculty of Applied Science & Engineering (APSC), University of Toronto Mississauga (UTM), and University of Toronto Scarborough (UTSC) account for 41.4%, 7.0%, 18.1% and 16.5% of survey responses, respectively (Table 3).

Additionally, for direct comparison between the 2018 Validation Study and the Renewed Validation Study, comparisons were made using only undergraduate, single instructor, fall and winter data from the four largest University of Toronto divisions (APSC, FAS, UTM, and UTSC).

Three levels of aggregation were used in the current study. Survey-level data correspond to the observed survey results without aggregation. Course-instructor level data involve the aggregated results for a single instructor within a single course section. For analyses including multi-instructor courses, course-level data involve results for a single course section for the survey items that are related to student experience at the course level, aggregated across instructors.

³ 15 Faculties/schools include Faculty of Arts and Sciences, Faculty of Applied Science and Engineering, Dalla Lana School of Public Health, Faculty of Dentistry, Faculty of Information, Joseph L. Rotman School of Management, Temerty Faculty of Medicine, Faculty of Music, Lawrence S. Bloomberg Faculty of Nursing, Ontario Institute for Studies in Education, Leslie Dan Faculty of Pharmacy, Faculty of Kinesiology and Physical Education, Factor-Inwentash Faculty of Social Work, UTM, UTSC.

⁴ The delivery mode variable was determined from the ROSI codes. Delivery mode codes changed throughout the five academic years covered by this study (2018/19 to 2022/23). Online asynchronous, online synchronous, and hybrid courses were classified as “online and hybrid”, while in-person courses were classified as “in-person”.

Table 2. Characteristics of the Data Used in This Study.

Course Characteristic	Number of Completed Surveys		Number of Courses	
	Count	%	Count	%
Semester				
Fall Courses	444,951	46.0%	22,608	39.8%
Winter Courses	425,537	44.0%	27,186	47.8%
Summer Courses	97,329	10.1%	7,052	12.4%
Level of Study				
Undergraduate Courses	834,956	86.3%	44,019	77.4%
Graduate Courses	132,859	13.7%	12,827	22.6%
Campus				
St George	620,183	64.1%	38,111	67.0%
UTM	184,722	19.1%	9,733	17.1%
UTSC	162,910	16.8%	9,002	15.8%
Type of Courses				
Full credit	102,579	10.6%	6,853	12.1%
Half credit	865,236	89.4%	49,993	87.9%
Number of Instructors				
Single-Instructor Courses	874,025	90.3%	53,697	94.5%
Multiple-Instructor Courses	93,790	9.7%	3,149	5.5%
Course Size				
Very Small (1-25 students)	164,021	16.9%	24,968	43.9%
Small (26-50 students)	190,706	19.7%	14,596	25.7%

Medium (51-100 students)	193,818	20.0%	9,459	16.6%
Large (101-200 students)	200,995	20.8%	5,468	9.6%
Very Large (201+ students)	218,275	22.6%	2,355	4.1%
Delivery Mode				
Online or Hybrid Courses	323,284	33.4%	19,038	33.5%
In-Person Courses	644,443	66.6%	37,803	66.5%
Total	967,817	100%	56,846	100%

Table 3. Distribution of the Completed Surveys and Courses Included in this Study Across Divisions for Undergraduate and Graduate Courses.

Division		Number of Completed Surveys		Number of Courses	
		Count	%	Count	%
Undergraduate					
APSC	Faculty of Applied Science and Engineering	67,383	7.0%	2,943	5.2%
DENT	Faculty of Dentistry	3,962	0.4%	319	0.6%
FAS	Faculty of Arts and Sciences	400,679	41.4%	20,999	36.9%
FIS	Faculty of Information	695	0.1%	66	0.1%
KPE	Faculty of Kinesiology and Physical Education	9,075	0.9%	427	0.8%
MUSIC	Faculty of Music	5,497	0.6%	890	1.6%
NURS	Lawrence S. Bloomberg Faculty of Nursing	4,069	0.4%	227	0.4%
PHM	Leslie Dan Faculty of Pharmacy	8,815	0.9%	352	0.6%
UTM	University of Toronto Mississauga	175,349	18.1%	9,135	16.1%
UTSC	University of Toronto Scarborough	159,432	16.5%	8,661	15.2%

Total for Undergraduate Courses		834,956	86.3%	44,019	77.4%
Graduate					
APSC	Faculty of Applied Science and Engineering	15,776	1.6%	1,587	2.8%
DLSPH	Dalla Lana School of Public Health	2,503	0.3%	264	0.5%
FAS	Faculty of Arts and Sciences	29,133	3.0%	4,147	7.3%
FIS	Faculty of Information	13,783	1.4%	840	1.5%
MED	Temerty Faculty of Medicine	116	<0.1%	19	<0.1%
MGT	Joseph L. Rotman School of Management	8,540	0.9%	359	0.6%
NURS	Lawrence S. Bloomberg Faculty of Nursing	3,778	0.4%	319	0.6%
OISE	Ontario Institute for Studies in Education	37,404	3.9%	3,500	6.2%
PHM	Leslie Dan Faculty of Pharmacy	169	<0.1%	38	0.1%
SWK	Factor-Inwentash Faculty of Social Work	8,806	0.9%	815	1.4%
UTM	University of Toronto Mississauga	9,373	1.0%	598	1.1%
UTSC	University of Toronto Scarborough	3,478	0.4%	341	0.6%
Total for Graduate Courses		132,859	13.7%	12,827	22.6%
Total for All Courses		967,817	100%	56,846	100%

Section 2. Scoring

This section focuses on key elements related to the scoring inference in Kane’s validity framework in the context of course evaluations. The scoring inference concerns how course evaluation survey items are designed (i.e., question content and response options), how students respond, and the extent to which these items produce consistent scores.

2.1 Survey Content, Design, and Administration

This subsection considers evidence related to the survey design and administration with three findings relevant to scoring inference.

Finding 1: Survey Items Reflecting Teaching Priorities

In 2010, the course evaluation framework was developed by the Course Evaluation Working Group, co-chaired by the Vice Provosts for Faculty & Academic Life and for Students, with extensive input from University of Toronto central offices and Faculties. The group’s mandate was to review current course evaluation practices across the university and at peer institutions as well as to improve the practices in the University based on research and feedback. Following extensive consultations and reviews, the working group proposed a centrally supported online evaluation system, resulting in the creation of the CCEF. This framework includes course evaluation items at four levels: core institutional, division-selected, department-selected, and instructor-selected. This report addresses the core institutional items.

The institutional items were designed to reflect teaching and learning priorities identified through consultations with divisions and instructors. To validate these items, the working group conducted student focus groups and a pilot study involving 437 students from four courses. Feedback was gathered on item perception, content, and generalizability to ensure the items effectively captured key teaching and learning priorities at the institutional level.

The core institutional items (Table 4) reflect these priorities, capturing key aspects of the student learning experience. Items Ins 1 through Ins 5 use a response scale ranging from “Not at All” to “A Great Deal” to measure students’ perceptions of these learning experiences. Item Ins 6 asks students to rate the overall quality of their learning experience on a scale from “Poor” to “Excellent.” Additionally, two open-ended questions solicit qualitative feedback: one on the overall quality of instruction (Ins 7) and another on academic support mechanisms (Ins 8).

Table 4. The Mapping between Institutional Core Teaching and Learning Priorities and Institutional Core Items Used in U of T's Course Evaluations.

Institutional Core Teaching and Learning Priorities	Institutional Core Items
Students are engaged in their own learning.	Ins 1: "I found the course intellectually stimulating."
Students learn a great deal in each course.	Ins 2: "The course provided me with a deeper understanding of the subject matter."
Students report that their course and instructor offer an environment conducive to learning.	Ins 3: "The instructor created a course atmosphere that was conducive to my learning."
Students indicate that the methods of assessment in a course reflect and contribute to their learning.	Ins 4: "Course projects, assignments, tests, and/or exams improved my understanding of the course material." Ins 5: "Course projects, assignments, tests, and/or exams provided opportunity for me to demonstrate an understanding of the course material."
Students have an overall positive learning experience with the course.	Ins 6: "Overall, the quality of my learning experience was:"
Students have an overall positive learning experience with the instructor.	Ins 7. "Please comment on the overall quality of the instruction in this course." (Open-ended question)
Students note the availability of support for their learning both from instructors and from across the institution.	Ins 8. "Please comment on any assistance that was available to support your learning in this course." (Open-ended question)

Finding 2: Survey Design and Administration Minimizing Irrelevant Influences in Survey Responses

The University of Toronto takes deliberate steps to minimize irrelevant influences in course evaluation survey responses, ensuring that the results accurately reflect students' learning experiences. "Irrelevant influences" refer to factors or distractions that can distort evaluation results, leading to outcomes that reflect unrelated elements rather than the students' actual experiences (Valencia, 2020). To address this, U of T employs an evidence-informed approach to survey design and administration, focusing on reducing these influences throughout the process.

The following strategies are used to minimize irrelevant influences:

- **Clear and relevant questions:** Items are crafted to focus on areas students are qualified to evaluate, specifically their self-reported learning experiences.
- **Student-facing framing statements:** Messages to students guide them to provide constructive feedback while raising awareness of potential biases.
- **Survey security:** Measures ensure that only invited students can access and respond to the questionnaire.

Research supports that students' responses to course evaluation surveys provide valuable insights into their learning experiences (Simonson et al., 2021; Spoooren et al., 2013). As key stakeholders in the classroom, students have ample opportunities to observe and reflect on their experiences, enabling them to offer credible feedback (Hativa, 2013). However, students are not positioned to evaluate aspects outside their expertise, such as instructors' teaching quality or subject matter knowledge (Dyer & Donnelly-Hermosillo, 2024; Spoooren et al., 2013). Recognizing this, U of T's course evaluation items are intentionally designed to focus on students' learning experiences, avoiding assessments of instructors' teaching performance or expertise.

In Summer 2024, CTSI introduced new student-facing framing statements for course evaluations. These statements are strategically integrated into the course evaluation platform to guide students as they complete their evaluations. Developed informed by evidence in survey methodology, these messages aim to:

- Increase awareness of unconscious biases that may influence feedback.
- Reinforce the purpose of course evaluations and the importance of constructive feedback.
- Assure students of the confidentiality of their responses.
- Guide students to provide actionable and respectful feedback.
- Direct students to appropriate resources for concerns that are outside the scope of course evaluations or that require immediate support (Benton et al., 2014).

Robust security measures are built into U of T's course evaluation process to ensure data integrity. Students must log in using their institutional credentials to access and complete the surveys, ensuring that only eligible participants provide feedback so that responses represent the intended student population.

By incorporating clear and relevant questions, leveraging framing statements to guide constructive feedback, and maintaining strong survey security protocols, U of T reduces irrelevant influences in course evaluation data. This ensures that the results more accurately represent students' authentic learning experiences.

Finding 3: Clear and Meaningful Response Options

Analysis of the institutional item response options at the University of Toronto indicates that they are clear, interpretable, and meaningful for students.

The six institutional items use two distinct sets of five-point response scales:

- **Ins 1 to Ins 5:** Options range from "Not at all," "Somewhat," "Moderately," "Mostly," to "A Great Deal."

- **Ins 6:** Options range from “Poor,” “Fair,” “Good,” “Very Good,” to “Excellent.”

Focus groups and a pilot study conducted between 2010 and 2012 confirmed that students could easily interpret and use these response options.

On the survey platform, response choices are displayed in a vertical list, which studies suggest improves usability compared to horizontal layouts (Hu, 2020). Additionally, the 2018 Validation Study found that students responded appropriately to a change in the scale orientation from descending (e.g., “A Great Deal” to “Not at all”) to ascending (e.g., “Not at all” to “A Great Deal”), providing proxy evidence that students actively consider response options when completing the surveys (CTSI, 2018a).

Both sets of response scales were carefully designed to capture meaningful variations in student feedback. Results indicate that students can differentiate among the five response options, as evidenced by their patterns of answers. Analysis of item response frequencies (Table 5) demonstrates that students consistently use the full range of the scales. Key findings include:

- A significant proportion of respondents selected the most positive response option for each item, ranging from 36% (“Excellent” in Ins 6) to 52% (“A Great Deal” in Ins 3).
- The endorsement rate (i.e., the percentage of surveys selecting the top two positive options) ranged from 64% (“Excellent” and “Very Good” in Ins 6) to 78% (“A Great Deal” and “Mostly” in Ins 2). Items Ins 1 to Ins 5 all had endorsement rates above 70%.
- A small percentage of respondents selected the most negative response option, with values ranging from 3% to 6%.
- Items Ins 2 and Ins 3 had the highest rates of students selecting the most positive response option, at approximately 50%.

The distribution of responses across all options suggests that the scales are effective in capturing a range of student experiences while indicating that most students report positive learning experiences.

Table 5. Percentages of Each of the Five Answer Options in the Six Institutional Items.

Item	Number of Responses	% of "Not at all"	% of "Somewhat"	% of "Moderately"	% of "Mostly"	% of "A Great Deal"	% of Endorsement
Ins 1	965,652	3%	7%	15%	31%	43%	74%
Ins 2	963,989	3%	7%	12%	30%	48%	78%
Ins 3	1,093,178 ^a	5%	7%	12%	24%	52%	77%
Ins 4	964,538	4%	9%	15%	31%	40%	72%
Ins 5	964,883	4%	8%	15%	31%	42%	73%
	Number of Responses	% of "Poor"	% of "Fair"	% of "Good"	% of "Very Good"	% of "Excellent"	% of Endorsement
Ins 6	965,426	6%	10%	20%	28%	36%	64%

Note: Results are analyzed at the survey level.

^a Ins 3 has a higher number of responses because it is answered for each instructor within multi-instructor courses.

2.2. Item Responses and Scoring

This section considers the nature of student responses to items with two findings relevant to scoring inference.

Finding 4: Indirect Evidence of Active Participation

The nature of student responses — specifically the choices made for institutional items and the completion of open-ended questions — was analyzed as a proxy for student engagement with the survey. The investigation focused on two key indicators:

- **Variation in Responses to Likert-Scale Items:** Most students selected varied response options across survey items.
- **High Engagement with Open-Ended Items:** A majority of students who responded to the survey, included responses to the open-ended questions, with substantive comments provided.

Most surveys had varied responses to survey items. Only 6% of completed surveys across all courses, and just 3% in undergraduate courses, had uniform responding, where the same response option was selected for every item. In the surveys with uniform response, the majority of respondents chose the most positive options, with very few surveys selecting the most negative option consistently (Table 6).

This study has a slightly higher proportion of uniform responses (6%) compared to the 2018 Validation Study (2%). This increase is primarily due to the inclusion of graduate-level courses in the current study.

Table 6. Percentage of Uniform Responses by Response Option.

Metric	Response Option 1	Response Option 2	Response Option 3	Response Option 4	Response Option 5
Percent	2%	1%	10%	11%	77%
Count	1,216	434	5,773	6,336	46,058

Among those who responded to the survey, 81% provided answers to the open-ended items (Ins 7 and Ins 8). Course evaluation literature suggests that a typical response rate to open-ended questions is 30-70% (Boysen, 2016). This high-level of response rate at U of T provides indirect evidence of student engagement in the survey process.

Analysis of word counts reveals that most qualitative responses were substantive, with:

- 61% of responses to Ins 7 (comments on instruction quality) contained more than five words.

- 70% of responses to Ins 8 (feedback on academic support) exceeded five words.

In summary, the data suggest that students who complete course evaluations surveys engage actively, as reflected in their nuanced quantitative responses and high engagement with qualitative items.

Finding 5: Positive and Large Inter-Item Correlations

All items have strong correlations with the mean of other items that form the ICM (i.e., corrected item-total correlations) when aggregated to the course-instructor level. Spearman rank-order correlations ranged from 0.89 to 0.94. Additionally, the items also have strong inter-item relationships when aggregated at the course (for Ins 1, Ins 2, Ins 4, Ins 5, Ins 6) or course-instructor (for Ins 3) level (Table 7).

The ICM also has a strong association with the overall item measuring students' perception of their learning experience (Ins 6), with a Spearman rank-order correlation of 0.94. This finding suggests that students' aggregated responses to questions about institutional teaching and learning priorities (Ins 1 to Ins 5) are closely associated with their rating of overall learning experience (Ins 6).

The relationships between pairs of individual items themselves are also robust, with inter-item correlations ranging from 0.78 to 0.92. These high inter-item correlations indicate that students' responses are consistent across the survey and reflect related aspects of their learning experience.

The strong and positive relationships among items provide confidence in using the ICM as an aggregate measure of students' learning experiences in courses.

Table 7. Bivariate Correlation Coefficients.

Item	Corrected Item-Total Correlation	Ins 1	Ins 2	Ins 3	Ins 4	Ins 5	Ins 6
Ins 1. Intellectually stimulating	0.91	1					
Ins 2. Deeper understanding	0.93	0.89	1				
Ins 3. Atmosphere promotes learning (Instructor-specific)	0.89	0.78	0.79	1			
Ins 4. Components improve understanding	0.94	0.79	0.81	0.78	1		
Ins 5. Opportunity to demonstrate understanding	0.93	0.78	0.80	0.78	0.92	1	
	ICM	Ins 1	Ins 2	Ins 3	Ins 4	Ins 5	Ins 6
Ins 6. Overall learning experience	0.94	0.86	0.87	0.87	0.87	0.87	1

Note: All correlations were statistically significant with $p < .001$. In this table, data were aggregated to the course level for Ins 1, Ins 2, Ins 4, Ins 5, and Ins 6, and aggregated to the course-instructor level for Ins 3 and the ICM. Correlations are Spearman rank order correlations. The corrected item-total correlation is the correlation between an item (e.g., Ins 1) and the mean of all other items that compose the ICM (e.g., Ins 2, Ins 3, Ins 4, and Ins 5).

2.3 Evidence of Comparability

This section investigates the comparability of ICM scores over time and across course contexts was investigated, with two findings relevant to scoring inference.

Finding 6: Stability of ICM Scores Over Time

Figure 3 shows the distribution of the ICM across the 60,995⁵ instructor-course pairings in 2018/19 to 2022/23. The ICM ranged from 1.1 to 5.0 with a mean value of 4.2, a median of 4.3 and a standard deviation of 0.56. The distribution of ICM values indicates that higher values occur more commonly than lower values.

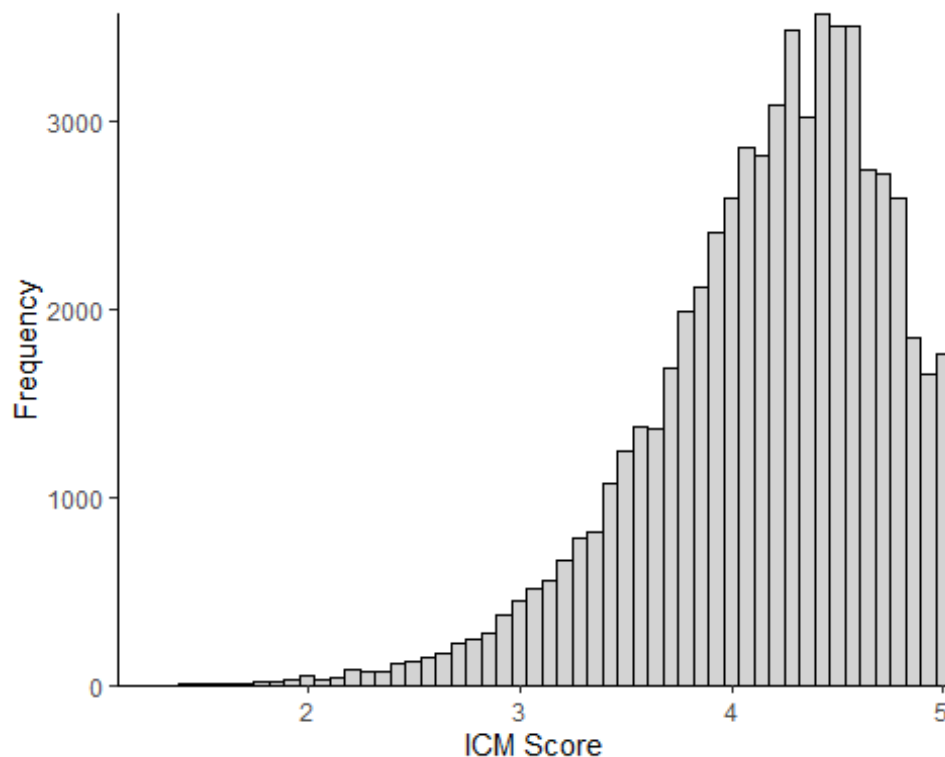


Figure 3. A Histogram of the ICMs at the Course-Instructor Level for the Whole Sample (Count=60,995).

Analysis of the ICM over time show that the mean of the ICM has remained relatively stable, with a slight increase observed over the last five years. The average ICM rose from 4.1 ($SD = 0.55$) in 2018/19 to 4.2

⁵ This number is slightly higher than the number of courses (Count = 56,846) shown in Tables 3 and 4, as the ICM is calculated at the instructor-course pairing level, that is, for each instructor for each course. 3,149 courses in the sample had more than one instructor included on the evaluation form.

($SD = .56$) in 2022/23. This increase corresponds to a small effect size (Cohen's $d = .22$), supporting the stability of the mean of the ICM during this period. These findings align with another study of course evaluations (Boysen, 2023), which reported that course evaluation results at a small private college in the USA were largely unaffected by the impact of the COVID-19 pandemic.

The distributions of ICMs were compared between the 2018 Validation Study and the current study (Figure 4). To ensure comparability, single-instructor courses, courses from fall and winter terms, and undergraduate courses from the four divisions with the most undergraduate courses (APSC, FAS, UTM, and UTSC) were used. The mean ICM increased slightly by 0.1 between studies ($M = 4.0$, $SD = .52$ for data from 2015/16 - 2016/17 and $M = 4.1$, $SD = .54$ for data from 2018/19 - 2022/23). The effect size of this difference was very small (Cohen's $d = .19$).

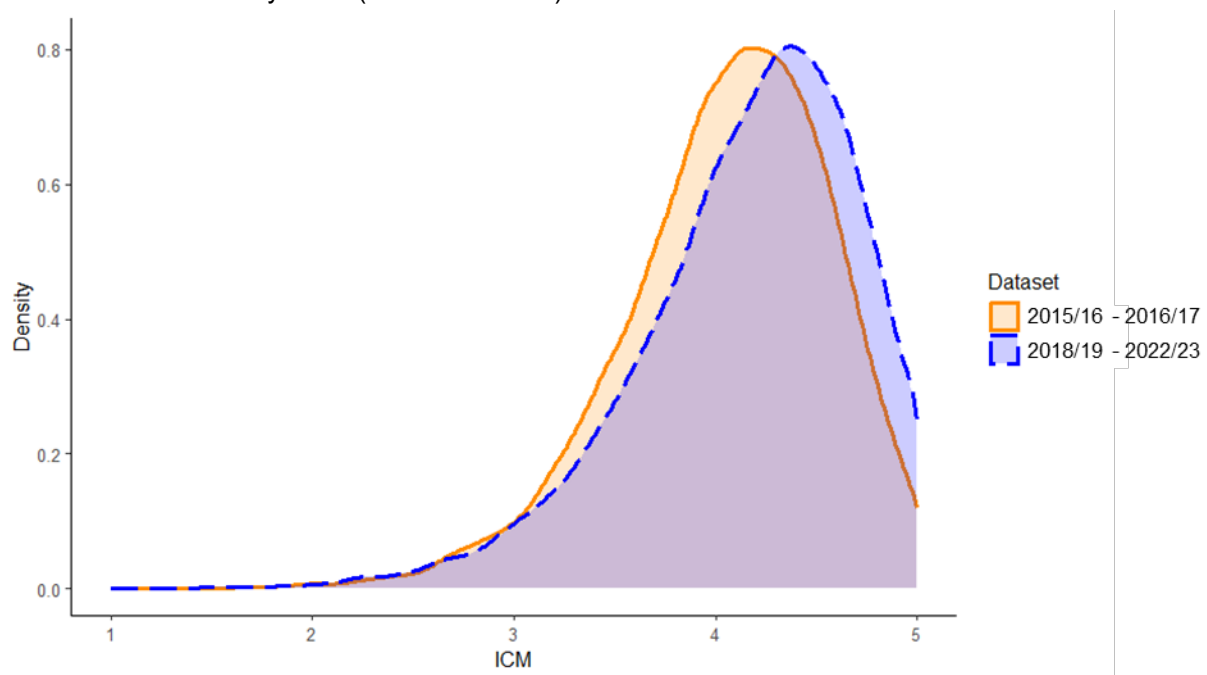


Figure 4. Density Plot of the ICMs in APSC, FAS, UTM, and UTSC Single-Instructor Undergraduate Courses for Two Time Periods (2015/16 - 2016/17 and 2018/19 - 2022/23).

Means for individual institutional items were compared between the 2018 Validation Study (2015/16 - 2016/17) and the current study (2018/19 - 2022/23) (Table 8). Results showed stability in institutional item means with only slight increases. Mean scores for most items increased by approximately 0.1, and median values for most items increased by 0.1 or 0.2. Cohen's d estimates for the comparison of item mean differences were all very small effects, ranging from $d = .05$ to $d = .11$.

Table 8. Comparison of Six Institutional Items between the 2018 and the Current Validation Studies.

Item	Number of Courses	Mean	Weighted Mean	Median	SD
2018 Validation Study (Undergraduate ASPC, FAS, UTM, UTSC)					
Ins 1	11,922	4.0	3.9	4.1	0.6
Ins 2	11,922	4.1	4.0	4.2	0.5
Ins 3	11,922	4.1	4.0	4.3	0.6
Ins 4	11,922	4.0	3.8	4.0	0.5
Ins 5	11,922	4.0	3.9	4.0	0.5
Ins 6	11,918	3.8	3.7	3.9	0.6
ICM	11,922	4.0	3.9	4.1	0.5
Renewed Validation Study (Undergraduate ASPC, FAS, UTM, UTSC)					
Ins 1	34,799	4.1	4.0	4.2	0.6
Ins 2	34,799	4.2	4.1	4.3	0.6
Ins 3	34,799	4.2	4.1	4.4	0.7
Ins 4	34,799	4.1	3.9	4.1	0.6
Ins 5	34,799	4.1	4.0	4.2	0.6
Ins 6	27,843	3.9	3.8	4.0	0.7
ICM	34,799	4.1	4.0	4.2	0.5

Note: Data in this table include undergraduate, single-instructor, fall and winter courses from the four divisions with the most undergraduate courses (APSC, FAS, UTM, and UTSC). The 2018 Validation Study includes data from Fall 2015 to Winter 2017. For this comparison, the current Validation Study includes data from Fall 2018 to Winter 2023.

Finding 7: Course Context Variations in the ICM Scores

The relationship between course size and ICM suggests larger courses tend to have lower ICM scores. Moderate negative relationships were observed between the ICM and course size (Spearman's rank-order correlation = -0.40, S.E. < .001, $p < .001$; Pearson correlation = -0.34, S.E. < .001, $p < .001$).

When grouping course sizes into categories, comparisons of ICM values across course size categories revealed that ICM means are moderately lower for larger courses ($\eta^2 = .10$) (Table 9). These results are similar to those found in the 2018 Validation Study (Table 10).

Table 11 displays the typical range of item endorsement rates across course size categories. Larger courses tend to have lower endorsement rates. For example, in courses with 1–25 students, the typical range of endorsement rates for Ins 1 is 62%–100%, compared to 48%–85% in courses with more than 200 students.

In addition to course size, this study also investigated whether other course characteristics have an impact on the ICM score, finding the following:

- **The semester the course was offered:** No meaningful differences were found.
- **Half-credit vs. full-credit courses:** No meaningful differences were found.
- **Online and hybrid vs. in-person courses:** No meaningful differences were found.
- **Undergraduate vs. graduate courses:** Undergraduate courses ($M = 4.1$, $SD = .55$) have slightly lower ICM scores than graduate courses ($M = 4.3$, $SD = .57$) with a small effect size (Cohen's $d = .36$). After controlling for course size using a general linear model, this effect becomes negligible (partial $\eta^2 < .001$).
- **Single-instructor vs. multi-instructor courses:** Courses with multiple instructors ($M = 4.0$, $SD = .54$) tend to have slightly lower ICM scores compared to single-instructor courses ($M = 4.2$, $SD = .56$) with a small effect size (Cohen's $d = .36$). This effect also becomes negligible after controlling for course size (partial $\eta^2 = .004$).

Table 9. ICM Summary Statistics by Course Size

Course Size	Number of Courses	ICM Mean	ICM Median	ICM SD	Typical ICM
Very Small (1-25 students)	26,616	4.3	4.3	0.55	3.8 - 4.9
Small (26-50 students)	15,431	4.1	4.1	0.53	3.6 - 4.6
Medium (51-100 students)	9,968	4.0	4.0	0.50	3.5 - 4.5
Large (101-200 students)	5,887	3.9	3.9	0.49	3.4 - 4.4
Very Large (201+ students)	3,093	3.8	3.9	0.48	3.3 - 4.3

Note: ICM refers to the institutional composite mean, an average of the first five institutional items for a course-section (Ins 1 to Ins 5). The mean and typical values in this table correspond to the average of ICMs over all course-sections taught in 2018/19 to 2022/23 (undergraduate & graduate) from all 15 faculties/schools using the central course evaluation system. The range of typical values corresponds to the 15th and 85th percentiles of the distribution of course-section ICMs.

Table 10. Comparison of ICM Values between the 2018 and the current Renewed Validation Studies

Course Size	Number of Courses	ICM Mean	ICM Weighted Mean	ICM SD	Typical ICM
2018 Validation Study (Undergraduate courses from ASPC, FAS, UTM, UTSC)					
Very Small (1-25 students)	4,041	4.3	4.3	0.54	3.8 - 4.8
Small (26-50 students)	3,216	4.1	4.1	0.48	3.6- 4.5
Medium (51-100 students)	2,343	3.9	3.9	0.47	3.4 - 4.4
Large (101-200 students)	1,694	3.8	3.8	0.42	3.4 - 4.3
Very Large (201+ students)	628	3.8	3.8	0.40	3.4 - 4.2
Renewed Validation Study (Undergraduate courses from ASPC, FAS, UTM, UTSC)					
Very Small (1-25 students)	6,345	4.3	4.4	0.55	3.8 - 4.8
Small (26-50 students)	4,790	4.1	4.2	0.50	3.6 - 4.6
Medium (51-100 students)	3,660	4.0	4.0	0.50	3.5 - 4.5
Large (101-200 students)	2,433	3.9	3.9	0.49	3.4 - 4.4
Very Large (201+ students)	878	3.8	3.8	0.47	3.3 - 4.3

Note: This table includes undergraduate, single-instructor, fall and winter courses from the four divisions with the most undergraduate courses (APSC, FAS, UTM, and UTSC). The 2018 Validation Study includes data from Fall 2015 to Winter 2017. The Renewed Validation Study includes data from Fall 2018 to Winter 2023.

Table 11. Typical Item Endorsement Percentages by Course Size

Course Size	Institutional Items (Ins) Typical Item Endorsement					
	Ins 1	Ins 2	Ins 3	Ins 4	Ins 5	Ins 6
Very Small (1-25 students)	62% - 100%	67% - 100%	67% - 100%	60% - 100%	62% - 100%	50% - 100%
Small (26-50 students)	56% - 97%	60% - 100%	57% - 100%	56% - 95%	57% - 100%	42% - 92%
Medium (51-100 students)	50% - 91%	57% - 93%	52% - 95%	50% - 90%	52% - 90%	36% - 85%
Large (101-200 students)	50% - 87%	56% - 91%	48% - 92%	48% - 85%	50% - 86%	33% - 79%
Very Large (201+ students)	48% - 85%	56% - 89%	47% - 90%	46% - 82%	48% - 84%	32% - 77%

Note: Values correspond to typical item endorsement (i.e., 15th & 85th percentiles of item endorsement rates) for institutional items (Ins 1 to Ins 6) for all course-sections using the central course evaluation framework from 2018/19 to 2022/23 academic years. Item endorsement is the percentage of students who selected the two most positive Likert-scale response options (i.e., 'Mostly' and 'A Great Deal' for Ins 1 to Ins 5, and 'Very Good' and 'Excellent' for Ins 6).

These results highlight course size as an important contextual factor for consideration when interpreting course evaluation data.

2.4. Summary of Scoring Inference

The above findings provide validity evidence supporting the scoring inference. Findings 1–3 demonstrate the robustness of the item development process and survey design. Finding 4 suggests that students provide effortful and intentional responses. Findings 6 and 7 show that the ICM remained stable over time and reflected variations based on course size. Collectively, these findings offer evidence that students translate their individual course experiences into meaningful survey responses.

Section 3. Generalization

This section focuses on key elements related to the generalization inference in Kane’s validity framework in the context of course evaluations. The generalization inference extends the interpretation of course evaluation results for a single course at a given time to represent a broader snapshot of all students’ possible responses across a range of survey items.

3.1 Response Rates and Representativeness of the Sample

Response rates to course evaluation surveys provide insights regarding the precision of course evaluation results and are summarized here, with one finding relative to generalization inference.

Finding 8: Precision for the Majority of Courses

Analysis of response rates and the sampling margin of error for course evaluations provides insights into the precision of the ICM. Applying the response rate guidelines from the 2018 Validation Study (CTSI, 2018a), 34.4% of courses in the current study have at least a “somewhat precise estimate” of the ICM, while 69.2% of courses have at least a “general estimate” of the ICM.

During the period 2018/19 to 2022/23, the average course-level response rate at the institutional level was 39.9% with typical values ranging from 20.0% to 62.5% (Table 12). Course-level response rates have decreased from 45.0% in 2018/19 to 37.8% in 2022/23, with the lowest response rate of 36.5% in the academic year 2021/22.

Table 12. Course-Level Response Rate (RR) by Academic Year (2018/19 to 2022/23)

Academic Year	Number of Courses	RR Mean	RR Median	RR SD	Typical RR
2018/19	10,560	45.0%	40.7%	0.2	25.0%-66.7%
2019/20	10,528	40.3%	36.0%	0.2	21.4%-62.5%
2020/21	12,499	40.6%	36.1%	0.2	21.4%-62.5%
2021/22	11,290	36.5%	31.2%	0.2	17.6%-58.8%
2022/23	11,969	37.8%	33.3%	0.2	17.6%-60.0%
Total	56,846	39.9%	35.3%	0.2	20.0%-62.5%

Note: RR: response rate. SD: standard deviation. Typical values correspond to the 15th and 85th percentiles of the distribution of response rates across courses.

Response rates also differ with course size. Smaller courses tend to achieve higher response rates compared to larger courses (Table 13). Response rates for other course characteristics (e.g., semester, level of study, number of course instructors, division) are provided in Appendix B.

Table 13. Response Rate (RR) by Course Size.

Course Size	Number of Courses	RR Mean	RR Median	RR SD	Typical RR
Very Small (1-25 students)	24,968	48.8%	45.5%	0.2	26.3% - 72.7%
Small (26-50 students)	14,596	37.4%	34.4%	0.2	20.0% - 56.2%
Medium (51-100 students)	9,459	30.6%	27.4%	0.2	16.8% - 44.8%
Large (101-200 students)	5,468	27.6%	25.6%	0.1	16.4% - 38.0%
Very Large (201+ students)	2,355	27.9%	25.5%	0.1	16.9% - 38.1%
Total	56,846	39.9%	35.3%	0.2	20.0% - 62.5%

Note: RR: response rate. SD: standard deviation. Typical values correspond to the 15th and 85th percentiles of the distribution of response rates across courses.

The average response rate varies across divisions (Figure 5). For undergraduate courses, FIS courses had the highest response rate of 46% and Dentistry had the lowest of 16%. For graduate courses, Pharmacy had the highest response rate of 58% and UTM had the lowest of 45%.

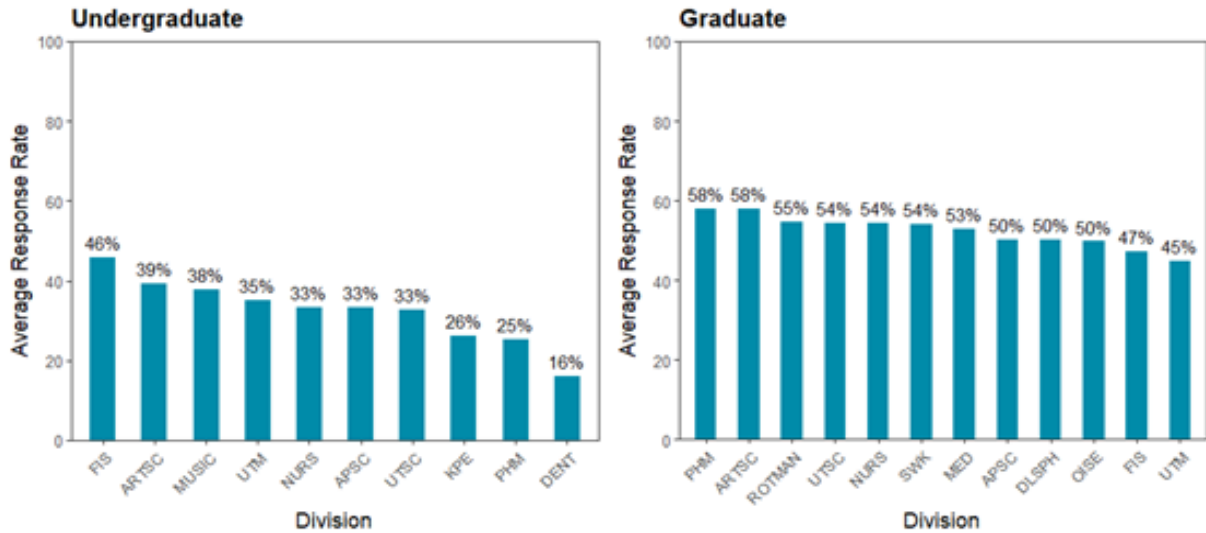


Figure 5. Average Course-Level Response Rates Across Undergraduate and Graduate Divisions.

At the student level, 79.7% of students responded to at least one course evaluation survey they were asked to complete, while 7.4% of students responded to every course evaluation survey they were asked to complete. 23.6% of students completed at least half of their course evaluations.

The sampling margin of error provides a measure of the precision of the ICM. The 2018 Validation Study used the sampling margin of error of the ICM to provide precision interpretation guidelines. Response rate interpretation thresholds were based on the width of a 95% confidence interval with a finite population correction (Equation 1 in James et al., 2015, p. 1129) and a standard deviation of 1.0. A standard deviation of 1.0 was used to align with the 2018 Validation Study. The table provided in the 2018 Validation study is reproduced below for reference (Table 14).

Table 14. 2018 Validation Study: Response Rate Needed for Precise Estimate of ICM

Width of Interval Around the ICM	Interpretation	Course Size				
		Very Small (1-25)	Small (26-50)	Medium (51-100)	Large (101-200)	Very Large (201+)
< ±0.1	Very precise estimate	90% - 100%	80% - 100%	80% - 100%	60% - 100%	50% - 100%
± 0.1 - ±0.2	Precise estimate	80% - 89%	70% - 79%	70% - 79%	50% - 59%	40% - 49%
±0.3 - ±0.5	Somewhat precise estimate	70% - 79%	50% - 69%	40% - 69%	20% - 49%	11% - 39%
±0.6 - ±1.0	General estimate	60% - 69%	20% - 49%	10% - 39%	10% - 19%	10%
> ±1.0	Very general estimate	0% - 59%	0% - 19%	0% - 9%	0% - 9%	0% - 9%

Note: Guidelines are based on the width of a 95% confidence interval, a standard deviation of 1.0, and correction for a finite population (CTSI, 2018a; James et al., 2015). Slight revisions were made to the 2018 table to ensure each cell represents a non-overlapping range of response rates. Course size refers to the number of students enrolled in the course section.

The number and percentage of courses within each precision category are given in Table 15. Most courses (69%) had response rates high enough (and thus the margin of errors low enough) to support at least a “general” level of precision. Approximately 34.4% of course sections had a response rate high enough to allow for a “somewhat” to “very precise” estimate of the ICM.

Table 15. Number and Percentage of Course Sections by Width of Interval Around the ICM

Width of Interval Around the ICM	Interpretation	Number of Courses	Percent of Courses
< 0.10	Very precise estimate	2,759	5.1%
0.10 - 0.19	Precise estimate	2,074	3.9%
0.20 - 0.49	Somewhat precise estimate	13,657	25.4%
0.50 - 0.99	General estimate	18,522	34.5%
≤ 1.00	Very general estimate	16,560	30.8%

The percentage of courses that fell within each level of precision within each course size category is given in Table 16. The percentage of courses with at least a general estimate increases as course size increases from 59.3% (1-25 students) to 67.9% (26-50 students) to 77.9% (101-200 students) to 99.7% (201+ students).

Table 16. Percentage of Courses within the Level of Precision of ICM by Course Size.

Width of Interval around the ICM	Interpretation	Course Size				
		Very Small (1-25)	Small (26-50)	Medium (51-100)	Large (101-200)	Very Large (201+)
< 0.10	Very precise estimate	10.7%	1.0%	0.4%	0.4%	1.5%
0.10 - 0.19	Precise estimate	4.0%	3.9%	2.5%	2.9%	10.9%
0.20 - 0.49	Somewhat precise estimate	18.8%	23.6%	26.3%	42.9%	70.2%
0.50 - 0.99	General estimate	25.3%	38.4%	48.7%	47.5%	17.1%
> 1.00	Very general estimate	40.7%	33.0%	22.1%	6.3%	0.2%

Note: Course size refers to the number of students enrolled in the course section.

3.2. Inter-Rater Reliability, Test-Retest Reliability, and Inter-Item Reliability

Consistency of the ICM was evaluated using inter-rater reliability, test-retest reliability, and inter-item reliability metrics, with two findings relevant to generalization inference.

Finding 9: Consistency across Students and Time

Inter-rater and test-retest reliability statistics provide information about the consistency of the ICM across students and over time. Reliability coefficients typically range between 0 and 1, with values above .70 or .80 indicative of adequate reliability. Inter-rater reliability measures the similarity of the ICM among student raters within a given course. Test-retest reliability measures the consistency of the ICM across occasions based on ratings from different students.

Inter-rater reliability statistics indicate that course ICM values generally exhibit good inter-rater reliability, with an overall coefficient of $ICC(k) = 0.86$. As shown in Table 17, inter-rater reliability varies with course size, with greater inter-rater reliability in courses with more students.

- Courses with fewer than 26 students had moderate inter-rater reliability ($ICC(k) = 0.74$).
- Courses with 26–100 students had good reliability ($ICC(k) = 0.81–0.85$).
- Courses with more than 100 students had excellent reliability ($ICC(k) > 0.90$).

Table 17. Inter-Rater Reliability ICC(k) for Whole Sample and Grouped by Course Size.

Total	Course Size				
	Very Small (1-25)	Small (26-50)	Medium (51-100)	Large (101-200)	Very Large (201+)
0.86	0.74	0.81	0.85	0.91	0.96

Note: Inter-rater reliability coefficients were calculated with an intraclass correlation based on a one-way random effect, absolute agreement, and multiple-rater model.

Test-retest reliability statistics indicate good reliability for the same instructor teaching different courses $ICC(k) = 0.80$, and moderate reliability for the same instructor teaching the same course across occasions $ICC(k) = 0.72$ (Table 18). Moderate consistency was also found for the same course across occasions regardless of instructor $ICC(k) = 0.73$, and the same students across different courses $ICC(k) = 0.73$, suggesting that students exhibit some stable tendencies in how they complete course evaluation surveys.

Table 18. Test-Retest Reliability Across Instructors, Courses, and Students

Object of Measurement	Description	ICC(k)
Instructor	Consistency of a given instructor's course ICM across different courses between 2018/19 - 2022/23 academic years	0.80
Course & Instructor	Consistency of ICM values for a given instructor teaching the same course across course-sections and semesters between 2018/19 - 2022/23 academic years	0.72
Course	Consistency of a given course's ICM across different instructors, course sections, and semesters between 2018/19 - 2022/23 academic years	0.73
Student	Consistency of a given student's ratings across different courses between 2018/19 - 2022/23 academic years	0.73

Finding 10: Consistency Among Institutional Items

Cronbach's alpha provides an indication of the internal consistency of survey items. Cronbach's alpha for the five institutional items that comprise the ICM (Ins 1 to Ins 5) is excellent (Cronbach's alpha = 0.92).

Together with the results of multilevel factor analysis in Section 6.1, this result suggests that the items work cohesively to measure an underlying construct.

Analysis of corrected item-total correlations confirms that each institutional item (Ins 1 to Ins 5) contributes positively to internal consistency. Removing any one item would reduce the overall Cronbach's alpha, further supporting the collective value of these items (Table 19).

Table 19. Results of Single-Item Analysis of Internal Consistency for Items Comprising the ICM

Item	Corrected Item-Total Correlation	Cronbach's Alpha if removed	Cronbach's Alpha S.E.	Difference in Cronbach's Alpha if Item Removed
Ins 1	0.787	0.909	< .001	-0.015
Ins 2	0.827	0.901	< .001	-0.023
Ins 3	0.761	0.914	< .001	-0.010
Ins 4	0.823	0.902	< .001	-0.022
Ins 5	0.805	0.906	< .001	-0.018

Note. The corrected item-total correlation is the correlation between an item (e.g., Ins 1) and the mean of all other items that compose the ICM (e.g., Ins 2, Ins 3, Ins 4, and Ins 5).

3.3. Generalizability Theory

Generalizability theory was used to investigate sources of variance in course evaluation scores and the predicted reliability of the ICM based on the number of student responses, with one finding relevant to generalization inference.

Finding 11: Generalizability in Representing Student Experiences

Generalizability (G) Theory, originally developed to address reliability, also supports validity studies by analyzing how various factors (facets) contribute to score variability. It extends classical test theory by replacing the concept of "true score" with "universe score" and using advanced statistical methods to disentangle specific sources of error. A G (generalizability) study estimates the contribution of various facets to score variance, while a D (decision) study uses these estimates to assess the reliability and accuracy of measurements under different scenarios. This framework helps identify construct-relevant facets and eliminate construct-irrelevant variance, supporting construct validity. For example, G theory can verify whether scores across different measurement conditions (e.g., courses or items) consistently and accurately reflect the intended construct.

In this study, G theory was applied to investigate sources of variance in course evaluation responses (G study) and determine the number of responses needed for reliable ICM values (D study) (see Figure 6 for the design). The course-section was defined as the object of measurement to align with the reporting of course evaluations. Results show that variation is primarily due to students within a course (53.0%) and

course-section (16.8%). D study findings indicate that 14–18 student responses per course evaluation survey are sufficient for reliable ICM values.

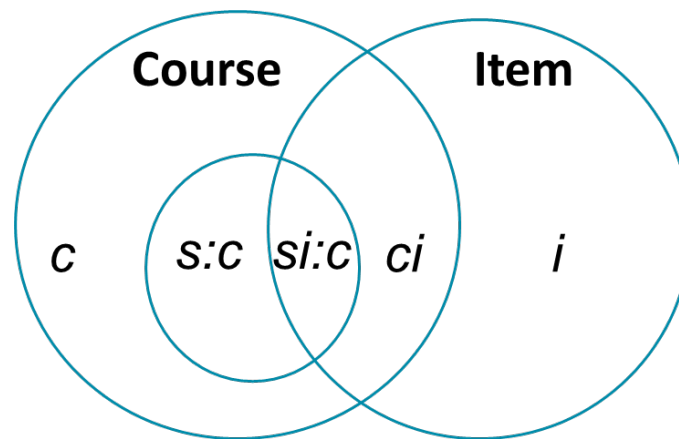


Figure 6. Generalizability Study Design

Notes: c: Course. s: Student. i: item. s:c: student nested within the course. ci: the interaction of course and item. si:c: the interaction of student and item nested within a course. This design coincides with the (i:p) x h design (Brennan, 2001, p. 56). Due to nesting, s:c refers to the confounded combination of s and sc, while si:c refers to the confounded combination of si, sic, and random error.

G theory variance estimates and the percentage of variance for each facet (Table 20) suggest that:

- **Student nested within a course:** The substantial amount of variance attributable to students nested within course-sections (53.0%) suggests individual student ratings differ when averaging over items. This finding supports the notion that course evaluations measure students' individualized learning experience.
- **Course:** A large amount of variance (16.8%) was due to differences among course-section. Averaging over items and students (nested within courses), course-sections differ systematically in their ratings. This variance indicates meaningful differences in scores across course-sections.
- **Course*Item:** The small item-by-course variance (2.2%) suggests that the differences among responses to items were similar across courses.
- **Item:** Item variance was minimal (0.5%), suggesting that items were similarly rated on average, supporting internal consistency.
- **Unexplained Variance:** The unexplained variance component is a confounded combination of the item-by-student interaction, the three-way course-by-student-by-item interaction, and unmeasured variation. The substantial unexplained variance (27.5%) suggests that a large amount of variation is due to these factors.

Table 20. G Study Results

Source of Variance	Variance Percent	Interpretation
Student nested within a course (s:c)	53.0%	Variation among students within courses
Course (c)	16.8%	Object of measurement

Course*Item (<i>ci</i>)	2.2%	How differences between items vary across courses
Item (<i>i</i>)	0.5%	Variation among items
Unexplained Variance	27.5%	Unexplained variance

Note. *c*: Course. *s*: Student. *i*: item. *s:c*: student nested within a course. *ci*: the interaction of course and item. Unexplained variance: the confounded combination of *si:c*: the interaction of student and item nested within a course and the error term. This design coincides with the (*i:p*) x *h* design (Brennan, 2001, p. 56). G-theory analysis is based on a random sample of 200 instructors, which includes 946 single-instructor courses, 19814 students, and 24778 surveys.

A D study is performed when the results of G studies are used to forecast the reliability coefficient that may be obtained for alternative hypothetical conditions, such as different numbers of student responses. In this case, a D study was conducted to estimate the number of student responses for a dependability coefficient of 0.8 for the mean of the five items that comprise the ICM. The D study model design used parallels the G study model design, using (*s:c*) x *i* with all facets treated as random. Dependability coefficients were estimated separately for various course sizes (5-25, 26-50, 51-100, 101-200, and 201). Results are similar across course sizes, with slightly larger student response thresholds for larger courses (Table 21). The number of responses for estimated reliability of .80 is 14 for courses with 5-50 students, 15 for courses with 51-200 students, and 18 for courses with 201+ students. Because larger courses have more students, the percentage response rate thresholds for adequate reliability decrease dramatically for larger course sizes.

Table 21. Sufficient Number of Responses (*n*) across Course Sizes for Reliability of the ICM

Predicted dependability coefficient ^a	Reliability interpretation	Course Size				
		5-25	26-50	51-100	101-200	201+
0.90 - 1.00	Very precise estimate	NA ^b	n = 34 ^b	n = 42	n = 42	n = 53
0.80 - 0.89	Precise estimate	n = 14	n = 14	n = 15	n = 15	n = 18
0.70 - 0.79	General estimate	n = 8	n = 8	n = 9	n = 9	n = 10
< 0.70	Very general estimate	n < 8	n < 8	n < 9	n < 9	n < 10

^a The predicted dependability coefficient, phi, is an estimate of reliability. Phi is based on absolute error, providing a more conservative reliability estimate than using relative error. An absolute error also aligns more closely with how course evaluations are interpreted for decision-making.

^b For smaller course sizes of 5-50 students, a dependability coefficient of .90 or higher requires 34 or more student responses. This is not possible for courses with less than 34 students.

The 2018 Validation Study's response rate thresholds can be combined with the results from the current generalizability theory D study to guide circumstances where the ICM values provide both good precision and reliability (Table 22). While courses with less than 25 students infrequently meet these student response rate thresholds, courses with more than 100 students often meet these response rate thresholds.

Table 22. Response Thresholds for Precise and Reliable Estimates of the ICM

Course Size	Sufficient Response Rate and Number of Responses	Percentage of Courses in 2018/19 to 2022/23 that Met the Threshold
Very Small (1-25 students)	Response rate of 70% or higher and at least 14 completed surveys	4.4%
Small (26-50 students)	Response rate of 50% or higher and at least 14 completed surveys	22.3%
Medium (51-100 students)	Response rate of 40% or higher and at least 15 completed surveys	21.0%
Large (101-200 students)	Response rate of 20% or higher and at least 15 completed surveys	73.0%
Very Large (201+ students)	Response rate of 10% or higher and at least 18 completed surveys	99.2%

Note. The sufficient response rate corresponds to an estimated width of the interval around the mean of ≤ 0.49 based on the findings of the 2018 Validation Study. The sufficient number of responses corresponds to an estimated ICM reliability coefficient of ≥ 0.80 based on this Renewed Validation Study's D-study results.

3.4. Generalization Inference Summary

The findings in this section provide validity evidence that course evaluation results for a single course at a specific time can be generalized to represent a broader snapshot of students' potential responses across various survey items. Finding 8 shows that many courses exhibit acceptable sampling error, Findings 9-10 indicate adequate inter-rater reliability, test-retest reliability, and internal consistency, and Finding 11 provides response rates for sufficient precision and reliability based on sampling error and generalizability theory results.

Section 4. Extrapolation

This section addresses extrapolation inference in Kane’s validity framework in the context of course evaluations. Extrapolation inference suggests that course evaluation results reflect students’ actual experience within the course. Multilevel factor analysis is used to explore the structure of the five institutional items comprising the ICM. Course-level unidimensionality supports the combination of the first five institutional items into the ICM. Course evaluation literature suggests course evaluation questionnaires measure student-reported learning experiences in the courses instead of a measurement of teaching quality, teaching effectiveness or student satisfaction (Dyer & Donnelly-Hermosillo, 2024; Spooren et al., 2013). Because available data are limited linking course evaluation results to broader measures of students’ learning experience, further extrapolation is beyond the scope of this study.

4.1. Multilevel Internal Structure (Dimensionality)

Multilevel factor analysis was used to explore the structure of institutional items Ins 1 to Ins 5, with one finding relevant to generalization inference. A unidimensional structure supports combining the items into the ICM at the course level (Stapleton et al., 2016). To be combined, institutional items Ins 1 to Ins 5 should measure a single underlying construct and demonstrate good item performance at the course level (AERA et al., 2014; van der Meulen et al., 2019).

Finding 12: Good Internal Structure

Multilevel Exploratory Factor Analysis (EFA) was conducted on a subsample of the data. Multilevel EFA results found evidence of unidimensionality at the course level based on a scree plot, the percent of item variance explained, parameter estimates, and the interpretability of the factor structure. The unidimensional model explained 91.0% of item variance at the course level. Multilevel EFA parameter estimates of the unidimensional model demonstrated large, standardized factor loadings and large item communalities at the course level. These results indicate that it is appropriate to combine the items into the ICM at the course level (i.e., course-level unidimensionality).

Multilevel Confirmatory Factor Analysis (CFA) was conducted on an additional subsample of the data. Comparison of different models suggests a 1-factor model without residual correlations fits the data at the course-level (the between-level). For surveys within courses (the within-level), a 1-factor model with residual correlations between Ins 4 and Ins 5, and between Ins 1 and Ins 2 best fit the data. The final model demonstrated good model fit. Course-level parameter estimates for the final multilevel CFA model demonstrate excellent item performance (Table 23). Items were strongly related to the ICM (standardized factor loadings ranging from 0.92 to 0.98). Item thresholds indicate that items capture an acceptably large range of the latent construct. Large, statistically significant item communalities show excellent course-level item reliability ranging from 0.85 to 0.94. Standardized factor loadings and item communality estimates surpassed recommended $>.40$ and $>.50$ cutoffs, respectively.

Table 23. Multilevel CFA Course-Level Item Parameter Estimates.

Item	Loading		Communality		Threshold 1		Threshold 2		Threshold 3		Threshold 4	
	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.
Ins 1	0.98	0.001	0.90	0.002	-4.00	0.01	-2.79	0.01	-1.54	0.01	0.15	0.01
Ins 2	0.97	0.001	0.94	0.002	-4.86	0.01	-3.40	0.01	-2.09	0.01	-0.12	0.01
Ins 3	0.92	0.002	0.85	0.003	-4.17	0.01	-3.03	0.01	-1.97	0.01	-0.45	0.01
Ins 4	0.97	0.001	0.94	0.002	-3.83	0.01	-2.58	0.01	-1.43	0.01	0.24	0.01
Ins 5	0.96	0.001	0.92	0.002	-3.70	0.01	-2.51	0.01	-1.44	0.01	0.13	0.01

Note: Loadings correspond to standardized factor loadings. Thresholds correspond to unstandardized thresholds using a probit link function. Factor loadings and communality estimates were statistically significant $p < .001$. The reported results are at the course-level (between-level). Data used in the multilevel factor results only included single-instructor courses.

Model-based internal consistency calculations of multilevel Cronbach's alpha and multilevel McDonald's unidimensional omega indicate excellent internal consistency at the course level (Table 24). In particular, the course-level reliability estimates (Cronbach's alpha = .977, McDonald's omega = .982) provide evidence for the internal consistency of the ICM (Stapleton et al., 2016), providing validity evidence supporting the aggregation of items into the ICM.

Table 24. Multilevel CFA Course-Level Internal Consistency Estimates.

Level	Parameter	Est.	S.E.	p	95% CIs
Course	Alpha	0.977	<.001	<.001	[.976, .978]
	Omega	0.982	<.001	<.001	[.981, .983]

Alpha: Cronbach's alpha (Cronbach, 1951). Omega: McDonald's unidimensional omega (McNeish, 2018). The reported results are at the course-level (between-level). Analysis for multilevel internal consistency was carried out only on single-instructor courses.

Multilevel CFA demonstrated adequate model fit for a unidimensional model and good item performance of the items that compose the course-level ICM. Multilevel internal consistency estimates demonstrated excellent scale reliability at the course level (Cronbach's alpha = 0.98, McDonald's omega = 0.98). These results provide validity evidence supporting the aggregation of items into the ICM metric.

4.2. Extrapolation Inference Summary

The finding in this section provides validity evidence supporting the extrapolation inference. Finding 12 shows that the five institutional items comprising the ICM have good internal structure and excellent internal consistency. Additional investigation of the relationships between the ICM and student learning experiences would require data sources beyond the responses to course evaluation surveys and is out of the scope of this current study.

Section 5. Implications

In the context of course evaluations, implication inference involves how course evaluation results are interpreted and used by stakeholders (Boysen 2015; Linse, 2017) such as informing decisions about iterative course improvement and decisions related to the evaluation of teaching (e.g., faculty hiring, promotion and merit assessment) (Chapelle & Voss, 2021; Cook et al., 2015). This section describes initiatives undertaken at U of T to support instructors and administrators in interpreting course evaluation results

While an empirical investigation into implication inference was beyond the scope of this study, U of T provides several resources to support instructors and administrators in using evidence-informed practices when interpreting course evaluation results (Dyer & Donnelly-Hermosillo, 2024; Spooren et al., 2013). These resources emphasize that course evaluation results should not be used as the sole data source for the assessment of teaching. A multi-faceted approach to evaluating teaching incorporating various data sources and measures is recommended by the *University of Toronto Provostial Guidelines on the Student Evaluation of Teaching in Courses* (Office of the Vice President and Provost, 2022).

The [University of Toronto Course Evaluation Interpretation Guidelines for Academic Administrators](#) (CTSI, 2018b) highlight key elements for effectively interpreting course evaluation reports for teaching assessment. These recommendations are based on ongoing work regarding validity of U of T course evaluation data (e.g., CTSI, 2018a) and a review of course evaluation literature. The 2018 Course Evaluation Interpretation Guidelines for Academic Administrators emphasize the following:

- **Course Evaluations as One Component of Teaching Evaluation:** Triangulation of multiple data sources (e.g., peer review of teaching, teaching dossier, course materials, student artifacts) is required to make valid inferences for evaluating teaching (Harrison et al., 2020).
- **University of Toronto Institutional Composite Mean (ICM):** This metric is valid and reliable for inferring student-perceived learning experiences, but it is somewhat impacted by contextual factors (e.g., course size, division) and sampling error. CTSI provides resources to interpret precision based on course size and response rate (CTSI 2018a).
- **Student-Reported Learning Experience:** Course evaluation scores represent students' reported learning experiences. The University of Toronto's course evaluation surveys only include questions regarding course elements about which students can adequately report their experience.
- **Interpreting Distribution of Responses:** In order to have a complete picture of the range of student experiences in a course, it is preferable to interpret the distribution of responses for a given course rather than a single summary statistic such as the mean.
- **Contextual Factors:** Factors such as multi-instructor courses, response rates, and course size should be considered when interpreting U of T course evaluations. Comparisons to departmental or divisional averages should be made with caution given contextual differences across courses.
- **Interpreting Open-Ended Student Comments:** It is important to focus on areas of noteworthy student consensus, rather than individual responses.

- **Discouraging Instructor Ranking:** Ranking instructors based on mean scores, especially at the decimal level, is discouraged due to the lack of precision of course evaluation values and the impact of contextual variables.

In the Winter 2025 term, new course evaluation reports were introduced. This redesign used user-experience principles and course evaluation literature reflecting recommended practices to guide interpretation. The results of this study have been used to inform the development of a [Step-by-Step Guide to Reviewing Course Evaluations](#) (CTSI, 2025) for instructors and administrators. This guide is consistent with other U of T documentation (CTSI, 2018b; Office of the Vice Provost and President, 2022) emphasizing the importance of using multiple data sources and considering contextual factors when interpreting course evaluation results. It also provides updated guidance on how the precision and reliability of results are related to the number of responses received. Typical ICM values and item endorsement rates, grouped by course sizes, are provided reflecting the findings in this study. General guidance for interpreting open-ended qualitative items (Ins 7 and Ins 8) is also included.

The *University of Toronto's Provostial Guidelines on the Student Evaluation of Teaching in Courses* (Office of the Vice Provost and President, 2022) underscore the University's commitment to supporting the valid interpretation and use of course evaluation data. CTSI provides educational materials and consultation to academic administrators and instructors to facilitate meaningful interpretation and decisions using course evaluation data. CTSI collaborates with academic administrators and instructors to use course evaluation results to support teaching development and effective pedagogical practices. Furthermore, CTSI engages in regular research, development, monitoring, and analysis of course evaluation data, updating interpretation guidelines regularly.

Section 6. Conclusion

This Renewed Validation Study analyzed data from all 15 Faculties/schools using the centralized course evaluations system over five academic years (2018/19 - 2022/23) to investigate the evidence for the validity of the quantitative institutional items. Utilizing a contemporary approach to psychometric analysis, this study advances the understanding of validity inferences in the context of U of T course evaluation surveys and processes, guided by Kane's argument-based validation framework. Both theoretical and empirical evidence were used to support validity inferences from twelve findings, as summarized in Figure 7.

The empirical evidence cited in this report can be integrated to form a validity argument for the ICM and institutional items Ins 1 through Ins 6.

- Validity evidence for the scoring inference links student experiences to the quantitative scores provided on the course evaluation survey (Findings 1-7). Findings 1-3 demonstrated the adequacy of the item development process and survey design, Finding 4 demonstrated student effortful responding, and Findings 6 and 7 demonstrated how the ICM remained stable across time and varied by course characteristics. Collectively, these claims provide evidence to support the conclusion that student respondents adequately translate their individual experiences in the course into survey responses.
- Validity evidence was also found supporting the interpretation of the ICM as a snapshot of possible responses from students in the course (generalization inference, Findings 8-11). Finding 8 showed that many courses exhibit acceptable sampling error. Findings 9-10 demonstrated adequate inter-rater reliability, test-retest reliability, and internal consistency. And Finding 11 provided recommended response rates based on sampling error and generalizability theory results. Together, these claims support the precision of ICM scores, suggesting the ICM adequately represents an aggregated snapshot of student responses within the course. These informed the development of updated response rate recommendations (CTSI, 2025) to approximate the level of precision of the ICM under various circumstances.
- While available data are limited in linking course evaluation results to broader measures of students' learning experience, multilevel factor analysis suggests the items demonstrate a good internal structure (extrapolation inference, Finding 12). Finding 12 demonstrated a good internal structure of the ICM. Investigating the relationship between ICM scores and variables beyond student-reported learning experience as captured through course evaluations (e.g., instructor-reported student learning experience, observer-reported student learning experience, student artifacts, and course materials) would support the claim that the ICM reflects students' authentic learning experience and is beyond the scope of the current study.
- Evidence for the implication inference, showing how course evaluation scores are used to make decisions, was beyond the scope of the current study. Nonetheless, various resources and consultation services are provided to support best practices in interpreting and using course evaluations (e.g., CTSI, 2018b; CTSI, 2025).

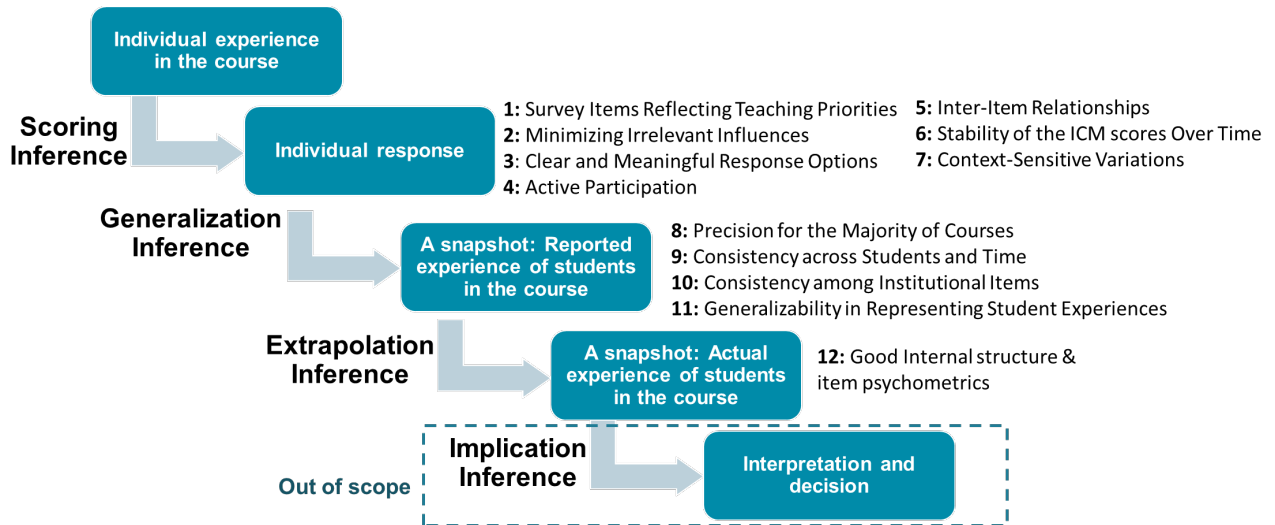


Figure 7. The Renewed Validation Study Validity Argument

Final Notes

This validation study is part of an ongoing institutional effort to support the quality of the Cascaded Course Evaluation Framework at the University of Toronto. This study draws upon input from diverse institutional stakeholders (e.g., the Course Evaluation Advisory Group) and experts who provided guidance around the key questions to examine the framework's effectiveness. Along with the 2018 Validation Study, this Renewed Validation Study exemplifies the University of Toronto and CTSI's dedication to continuous analysis and education concerning the quality and interpretation of course evaluation data.

Course evaluations provide snapshots of student perspectives on their learning experiences at the course level. Most experts on teaching evaluation advise that no individual method offers a complete picture of an instructor's teaching effectiveness. Multiple and diverse measures, taken on multiple occasions, are recommended to provide a comprehensive view of teaching effectiveness. Additionally, contextual factors such as course size, course level, and division are related to course evaluation responses. Therefore, student perspectives for a particular instructor or course should be interpreted as a snapshot, not as complete information on the teaching effectiveness of that instructor.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf
- Benton, S. L., & Cashin, W. E. (2014). Student Ratings of Instruction in College and University Courses. In *Higher Education: Handbook of Theory and Research* (Vol. 29, pp. 279–326). Springer Netherlands.
https://doi.org/10.1007/978-94-017-8005-6_7
- Boysen, G. A. (2015). Uses and Misuses of Student Evaluations of Teaching: The Interpretation of Differences in Teaching Evaluation Means Irrespective of Statistical Information. *Teaching of Psychology*, 42(2), 109–118.
<https://doi.org/10.1177/0098628315569922>
- Boysen, G. A. (2016). Using student evaluations to improve teaching: Evidence-based recommendations. *Scholarship of Teaching and Learning in Psychology*, 2(4), 273–284.
<https://doi.org/10.1037/stl0000069>
- Boysen, G. A. (2023). Student evaluations of teaching during the COVID-19 pandemic. *Scholarship of Teaching and Learning in Psychology*, 9(3), 254–263. <https://doi.org/10.1037/stl0000222>
- Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag Publishing. <https://doi.org/10.1007/978-1-4757-3456-0>
- Centre for Teaching Support & Innovation. (2018). *University of Toronto's Cascaded Course Evaluation Framework: Validation Study of the Institutional Composite Mean (ICM)*. Toronto, ON: Centre for Teaching Support & Innovation, University of Toronto. https://teaching.utoronto.ca/wp-content/uploads/Validation-Study_CTSI-September-2018.pdf
- Centre for Teaching Support and Innovation (2018). *University of Toronto Course Evaluation Interpretation Guidelines for Academic Administrators*. Toronto, ON: Centre for Teaching Support & Innovation, University of Toronto. https://teaching.utoronto.ca/wp-content/uploads/CE_Interpretation-Guidelines_Final_Oct.1.2018.pdf
- Centre for Teaching Support and Innovation (CTSI). (2025). *A Step-by-Step Guide to Reviewing Course Evaluations*. Toronto: Centre for Teaching Support & Innovation, University of Toronto.
<https://teaching.utoronto.ca/course-evaluations/for-instructors-and-administrators/a-step-by-step-guide-to-reviewing-course-evaluations/>
- Chang, L., & Hocevar, D. (2000). Models of Generalizability Theory in Analyzing Existing Faculty Evaluation Data. *Applied Measurement in Education*, 13(3), 255–275. https://doi.org/10.1207/S15324818AME1303_3
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2011). *Building a Validity Argument for the Test of English as a Foreign Language* (1st ed.). Routledge. <https://doi.org/10.4324/9780203937891>
- Chapelle, C. A., & Voss, E. (Eds.). (2021). *Validity Argument in Language Testing: Case Studies of Validation Research* (1st ed.). Cambridge University Press. <https://doi.org/10.1017/9781108669849>
- Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: a practical guide to Kane's framework. *Medical Education*, 49(6), 560–575.
<https://doi.org/10.1111/medu.12678>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
<https://doi.org/10.1007/BF02310555>
- Curby, T., McKnight, P., Alexander, L., & Erchov, S. (2020). Sources of variance in end-of-course student evaluations. *Assessment & Evaluation in Higher Education*, 45(1), 44–53.
<https://doi.org/10.1080/02602938.2019.1607249>
- Dyer, K. D., & Donnelly-Hermosillo, D. (2024). Student Ratings of Instruction: Updating Measures to Reflect Recent Scholarship. *Research in Higher Education*. <https://doi.org/10.1007/s11162-024-09804-8>

- Gibbs, A., McCloy, C., Tu, Y., Lau, C., Quibrantar, S. M., Hutrya, M., Pham, J. (2023, July 30 – August 2). *Highlighting the Student Voice: Co-developing & Piloting Research Protocols to Explore Student Perspectives and Experiences with Course Evaluations*. Bluenotes Global 2023, Louisville, KY, United States. <https://www.bluenotesgroup.com/wp-content/uploads/2023/08/Highlighting-the-Student-Voice-Co-developing-Piloting-Research-Protocols-to-Explore-Student-Perspectives-and-Experiences-with-Course-Evaluations.pdf>
- Harrison, R., Meyer, L., Rawstorne, P., Razee, H., Chitkara, U., Mears, S., & Balasooriya, C. (2020). Evaluating and enhancing quality in higher education teaching practice: a meta- review. *Studies in Higher Education*, 47(1), 80–96. <https://doi.org/10.1080/03075079.2020.1730315>
- Hativa, N. (2013). *Student ratings of instruction: A practical approach to designing, operating, and reporting*. Oron Publications.
- Hu, J. (2020). Horizontal or vertical? The effects of visual orientation of categorical response options on survey responses in web surveys. *Social Science Computer Review*, 38(6), 779-792. <https://doi.org/10.1177/0894439319834296>
- James, D. E., Schraw, G., & Kuch, F. (2015). Using the sampling margin of error to assess the interpretative validity of student evaluations of teaching: Assessment & Evaluation in Higher Education. *Assessment & Evaluation in Higher Education*, 40(8), 1123–1141. <https://doi.org/10.1080/02602938.2014.972338>
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527–535. <https://doi.org/10.1037/0033-2909.112.3.527>
- Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education/Praeger.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.1200>
- Linse (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in Educational Evaluation*, 54, 94–106. <https://doi.org/10.1016/j.stueduc.2016.12.004>.
- McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, 23(3), 412–433. <https://doi.org/10.1037/met0000144>
- Messick S. (1989). Validity. In Linn R. (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.
- Messick S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741–749. <https://doi.org/10.1037/0003-066X.50.9.741>
- Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the Validity of Student Evaluation of Teaching The State of the Art. *Review of Educational Research*, 83(4), 598–642. <https://doi.org/10.3102/0034654313496870>
- Spooren, P., Mortelmans, D., & Christiaens, W. (2014). Assessing the validity and reliability of a quick scan for student's evaluation of teaching. Results from confirmatory factor analysis and G Theory. *Studies in Educational Evaluation*, 43, 88–94. <https://doi.org/10.1016/j.stueduc.2014.03.001>
- Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics*, 41(5), 481–520. <https://doi.org/10.3102/1076998616646200>
- Valencia, E. (2020) Acquiescence, instructor's gender bias and validity of student evaluation of teaching, *Assessment & Evaluation in Higher Education*, 45:4, 483-495, <https://doi.org/10.1080/02602938.2019.1666085>
- van der Meulen, M. W., Smirnova, A., Heeneman, S., Oude Egbrink, M. G. A., Van Der Vleuten, C. P. M., & Lombarts, K. M. J. M. H. (2019). Exploring Validity Evidence Associated With Questionnaire-Based Tools for Assessing the Professional Performance of Physicians: A Systematic Review. *Academic Medicine*, 94(9), 1384–1397. <https://doi.org/10.1097/ACM.0000000000002767>

Appendices

Appendix A: Institutional Items in the Cascaded Course Evaluation Framework

Cascaded Course Evaluation Framework

The Cascaded Course Evaluation Framework (CCEF) applies a cascaded assessment structure that acknowledges the need for both broad-based and granular assessment across the various levels of the institution. The CCEF includes four levels of questions: core institutional, division-selected, department-selected, and instructor-selected. Institutional items are the subject of this report.

The Five Core Items

The five key teaching and learning priorities and their respective items include:

- Students are engaged: Ins 1, “I found the course intellectually stimulating.”
- Students gain knowledge: Ins 2, “The course provided me with a deeper understanding of the subject matter.”
- Atmosphere promotes learning: Ins 3, “The instructor created a course atmosphere that was conducive to my learning.”
- Components improve understanding: Ins 4, “Course projects, assignments, tests, and/or exams improved my understanding of the course material.”
- Students have an opportunity to demonstrate understanding: Ins 5, “Course projects, assignments, tests, and/or exams provided opportunity for me to demonstrate an understanding of the course material.”

Response options for the five core institutional items range from “Not At All,” “Somewhat,” “Moderately,” “Mostly,” to “A Great Deal.”

The Institutional Composite Mean

Responses to the five core items are averaged together to create a single “Institutional Composite Mean” (ICM). The ICM (which ranges from 1.0 to 5.0) reflects the extent to which all five institutional priorities were part of the students’ learning experience within a given course.

Overall Learning Experience

A sixth institutional rating scale item assesses students’ perceptions of their overall learning experience in a course.

- Overall learning experience in a course: Ins 6, “Overall, the quality of my learning experience in this course was:”

Response options for institutional item Ins 6 range from “Poor,” “Fair,” “Good,” “Very Good,” to “Excellent.”

Qualitative Feedback

The last two institutional items allow students the opportunity to make qualitative comments in response to two open-ended prompts:

- Ins 7, “Please comment on the overall quality of the instruction of this course.”
- Ins 8, “Please comment on any assistance that was available to support your learning in the course.”

Appendix B: Supplemental Response Rate Metrics

Response rate metrics were compared based on a variety of course characteristics (Table B1). Descriptive statistics illustrated that courses with a single instructor tend to have a higher response rate than courses with multiple instructors. The higher number of the instructors in a course, the lower the response rates were. On average, graduate courses had a much higher response rate as compared to undergraduate courses. There was no noticeable difference in the average response rates between online and in-person courses. Response rates across divisions are reported in Table B2.

Table B1. Descriptive Statistics of Response Rates (RR) Across Course Categories.

Course Category	Number of Courses	RR Mean	RR Median	RR SD	Typical RR
Semester					
Fall Courses	22,608	42.7%	38.4%	20.34	22.8% - 66.7%
Winter Courses	27,186	38.9%	33.9%	20.35	19.5% - 60.7%
Summer Courses	7,052	35.1%	29.7%	20.23	16.2% - 57.1%
Level of Study					
Undergraduate	44,019	36.2%	32.0%	0.2	18.9% - 55.4%
Graduate	12,827	52.8%	50.0%	0.2	30.0% - 75.8%
Type of Courses					
Full-Credit	6,853	37.8%	33.3	21.01	17.9% - 61.1%
Half-Credit	49,993	40.2%	35.7	20.40	20.4% - 62.5%
Number of Instructors					
Single-Instructor	53,697	40.1%	35.6%	0.2	20.0% - 62.5%
2 Instructors	2461	38.1%	34.1%	0.2	18.8% - 58.8%
3 Instructors	390	33.6%	29.8%	0.2	17.8% - 50.0%
4+ Instructors	298	28.9%	26.8%	0.1	16.7% - 41.0%

Delivery Mode

Online and Hybrid Courses	19,038	40.7%	36.4%	0.2	20.5% - 63.6%
In-Person Courses	37,803	38.4%	33.3%	0.2	19.5% - 60.0%

Table B2. Descriptive Statistics of Response Rates (RR) Across Divisions.

Division	Number of Courses	RR Mean	RR Median	RR SD	Typical RR
Undergraduate					
APSC	2,943	33.3%	29.5%	17.3%	17.9% - 49.5%
DENT	319	16.0%	10.8%	15.6%	4.2% - 27.2%
FAS	20,999	39.2%	35.0%	18.8%	21.4% - 60.0%
FIS	66	45.7%	42.1%	21.4%	24.8% - 71.5%
KPE	427	26.1%	22.1%	15.8%	12.5% - 39.3%
MUSIC	890	37.9%	33.3%	18.8%	20.0% - 57.1%
NURS	227	33.4%	29.3%	17.6%	17.9% - 48.5%
PHM	352	25.4%	20.0%	15.9%	12.0% - 42.0%
UTM	9,135	35.0%	30.4%	18.3%	18.2% - 53.8%
UTSC	8,661	32.7%	28.6%	17.3%	17.1% - 50.0%
All UG Courses	44,019	36.2%	32.0%	20.0%	18.9% - 55.4%
Graduate					
APSC (Grad)	1,587	50.2%	48.1%	21.3%	28.6% - 74.2%
DLSPH (Grad)	264	50.2%	46.3%	21.0%	29.3% - 72.1%

FAS (Grad)	4,147	57.9%	55.6%	21.6%	35.0% - 81.8%
FIS (Grad)	840	47.3%	47.4%	17.3%	28.6% - 64.7%
MED (Grad)	19	52.9%	50.0%	21.7%	31.3% - 77.5%
ROTMAN (Grad)	359	54.5%	52.6%	22.6%	30.7% - 82.8%
NURS (Grad)	319	54.3%	51.5%	17.7%	37.0% - 76.5%
OISE (Grad)	3,500	49.7%	48.0%	20.7%	27.6% - 72.7%
PHM (Grad)	38	57.9%	60.0%	17.5%	37.5% - 70.4%
SWK (SGS)	815	54.1%	53.8%	19.6%	33.3% - 75.0%
UTM (SGS)	598	44.9%	39.4%	24.8%	20.5% - 76.0%
UTSC (SGS)	341	54.4%	52.2%	23.2%	30.6% - 80.0%
All Grad Courses	12,827	52.8%	50.0%	20.0%	30.0% - 75.8%