University of Toronto's Cascaded Course Evaluation Framework: Validation Study of the Institutional Composite Mean (ICM)

Centre for Teaching Support & Innovation, 2018

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

Please cite this publication in the following format:

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.

Academic Leads:

Susan McCahan, Vice-Provost, Academic Programs and Vice-Provost, Innovations in Undergraduate Education; Professor, Department of Mechanical and Industrial Engineering, University of Toronto

Carol Rolheiser, Director, Centre for Teaching Support & Innovation (CTSI); Professor, Department of Curriculum, Teaching and Learning, University of Toronto

Contributors:

Kosha Bramesfeld, Data Analyst, Course Evaluations, CTSI
Gregory Hum, Assistant Director, Teaching Assessment, CTSI

Acknowledgements:

Megan Burnett, Associate Director, CTSI
Diane Horton, Acting Director (January-June, 2018), CTSI
Katharine Lauder, Course Evaluation Assistant, CTSI
Tara Wells, Operations Coordinator, Course Evaluations, CTSI

Table of Contents

EXECUTIVE SUMMARY

Study Overview

Since 2012, the University of Toronto has progressively implemented an evidence-based centralized Cascaded Course Evaluation Framework (CCEF) for collecting feedback data from students. This paper reports the results of a validation study that examined the reliability and validity of the institutional items of the CCEF. The validation study used data from 277,498 completed evaluation surveys collected across two academic years (2015/2016 and 2016/2017) from 11,919 single-instructor undergraduate course sections from 118 academic units across the four largest undergraduate divisions at the University of Toronto (Faculty of Applied Science & Engineering (FASE), Art & Science (ARTSC), University of Toronto, Mississauga (UTM), and University of Toronto Scarborough (UTSC)). It is important to note that the generalizability of the results contained here is not yet conclusively determined for other divisions at the University of Toronto or time periods outside of this sampling frame. Further analyses are planned and ongoing.

The validation study focused on assessing the reliability and validity of the Institutional Composite Mean (ICM). The ICM represents the average of five core institutional items that are included on all course evaluation surveys that use the Cascaded Course Evaluation Framework. These five items are intended to capture five key teaching and learning priorities at the University of Toronto. These five priorities, and their respective items, are listed below:

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

Key Findings of the Validation Study

1. Response Rates

A. Across course sections, the average course evaluation response rate was 42%.
B. Students were more likely to submit their surveys in the afternoon or evening.
C. Larger courses were associated with smaller response rates.
D. The response rates were comparable with other online surveys of student engagement.
E. Response rates were high enough to allow for general-levels of meaningful inference.
F. Student-faculty interaction, not student dissatisfaction, predicted higher response rates.
G. Response rates were not associated with survey length, fatigue, or alphabetical order.
H. Lower response rates did not meaningfully disadvantage instructors.

2. Student Response Patterns

A. Students rated an average of 99% of the rating scale items presented to them.
B. Students did not engage in wide-spread yea-saying, nay-saying, or neutral responding.
C. Students were responsive to shifting scale options.
D. Students favoured the upper end of the rating scale.
E. Rates of endorsement were within recommended levels.

3. Reliability

A. **Interrater reliability**. Students within a single course exhibited high enough agreement and reliability in their ratings of the institutional items to justify aggregating these ratings to the course-section level for interpretation.
B. **Internal consistency**. The five items of the ICM exhibited high enough internal consistency to justify averaging the items into an Institutional Composite Mean (ICM).
C. **Test-retest reliability**. ICM scores were most stable when considering the same instructor teaching the same course over time.

4. Construct Validity

A. **Student engagement**: The ICM was more strongly correlated with indicators of course-created engagement than with students' prior interest in the topic or class attendance.
B. **Knowledge gains**: The ICM was more strongly correlated with students' perceived opportunities to gain knowledge than with their expected grade performance.
C. **Learning atmosphere**: The ICM was more strongly correlated with quality of instruction indicators than with course support factors.
D. **Quality of assessment:** The ICM was more strongly correlated with the quality and fairness of assessment than with the perceived workload of the course.

5. Dimensionality

A. The ICM is more reliable and stable than the institutional items considered individually.
B. The ICM exhibits stronger construct validity than any given institutional item.
C. The ICM is better at differentiating between course sections than any individual item.
D. The ICM is more appropriately used for summative purposes than individual items.

6. Contextual Analysis

A. Larger course sizes were moderately associated with lower ICM scores.
B. Course level predicted ICM scores, but mainly due to course size differences.
C. Only trivial differences in ICM scores emerged between the four academic divisions.
D. ICM scores differed between academic units, but mostly due to course size.
E. ICM differences between course formats were trivial, and mostly due to course size.
F. ICM scores were not associated with course length or the course term.
G. ICM scores were not associated with students' full time status or year of study.

7. Demographic Analysis

A. No gender differences emerged on response rates or institutional item ratings.
B. ICM scores were not associated with faculty rank, age, or seniority.

8. Interpretability of ICM scores

A. ICM scores fell along the full continuum of possible scores (1.0 to 5.0).
B. ICM scores were skewed towards the upper end of the scale (M = 4.0, S = 0.52).
C. ICM scores exhibited discrimination ability across the full range of scale options.
D. ICM scores are especially diagnostic at the upper and lower ends of the scale.
E. Larger course sizes were associated with lower ICM scores, r = -0.41.
F. Scores between 3.4 and 4.8 reflect a 'typical' student experience.

9. Generalizability

A. The ICM exhibits identical reliability and validity patterns across academic divisions studied.
B. The ICM is generalizable to graduate-level courses.
C. The ICM is generalizable to dual-instructor courses, but the evaluation context differs.

Implications for Interpretation of the ICM

Response rates

ICM scores will be most meaningful when response rates are 50% or higher for small courses (< 50 students) and 20% or higher for larger courses (> 100 students).

Table 1

Response Rate Needed to Make Meaningful Inference

		Course Size				
Interval around the mean	**Recommended interpretation of the quality of the mean estimate**	1-25	26-50	51-100	101-200	200+
< ±0.1	Very precise estimate	>90%	>80%	>80%	>60%	>50%
< ±0.2	Precise estimate	>80%	>70%	>70%	>50%	>40%
< ±0.5	Somewhat precise estimate	>70%	>50%	>40%	>20%	>10%
< ±1.0	General estimate	>60%	>20%	>10%	>10%	>10%
> 1.0+	Very general estimate	< 30%	<10%	<5%	<3%	<1%

Note. Guidelines are based on a 95% confidence interval around the mean with margin of errors ranging from ±0.1 to ±1.0, a standard deviation of 1.0, and correction for the use of a finite population.

ICM Interpretation

The table below describes the "range of typicality" (i.e., the middle 70%) for any given course size. Scores within this range reflect a 'typical' collective student experience as measured by the ICM. Scores outside of this range are 'atypical' in that they reflect the bottom 15% of ICM scores and the top 15% of ICM scores. Importantly, however, atypically low scores do not, necessarily, indicate poor teaching, nor do atypically high scores, necessarily, indicate exemplary teaching. ICM scores can be influenced by a number of factors, some of which are outside of the control of the instructor. With that said, an atypical ICM score may warrant further investigation. See the section below for additional sources of evidence that can be used to contextualize ICM scores.

Table 2

Range of Typical ICM Scores for Each Course Size Category

Course size	*M*	Typical (middle 70%)	Lower than typical (bottom 15%)	Higher than typical (top 15%)
1-25	4.3	3.7 and 4.8	≤ 3.6	≥ 4.9
26-50	4.0	3.6 and 4.5	≤ 3.5	≥ 4.6
51-100	3.9	3.4 and 4.4	≤ 3.3	≥ 4.5
101-200	3.9	3.4 and 4.3	≤ 3.3	≥ 4.4
201+	3.8	3.4 and 4.2	≤ 3.3	≥ 4.3

ICM scores in a larger context

Course evaluation scores should always be interpreted within the larger teaching and learning context. Possible sources of evidence that can be used to contextualize ICM scores include (but are not limited to) an instructor's narrative explanation of their teaching contexts, course context variables, students' written comments, classroom observation, course materials, and/or other supporting documents.

When interpreting ICM scores the results of the validation study suggest that the following contextual factors may be of particularly high importance for interpreting ICM scores:

- **Specific division/department**. Although differences were small and mostly explained by differing course sizes, ICM scores varied from division to division and from department to department. As such, ICM scores should be interpreted within the context of specific divisions and departments. It is important to note that these observed differences do not necessarily indicate relative quality of teaching or learning experiences between divisions/departments.
- **The size of the course**. Larger course sizes are associated with lower ICM scores. As such, course size should be taken into consideration when interpreting ICM scores.
- **Single instructor versus dual/multi-instructor courses.** Somewhat similar ICM values emerged between single-instructor and dual-instructor courses. The items were also psychometrically similar when it came to the factor structure. However, analyses suggest that students used somewhat different criterion to rate single-instructor versus dual-instructor courses, especially when it came to institutional item 3. Whether a course has multiple instructors should be taken into consideration when interpreting ICM scores.

INTRODUCTION

The University of Toronto's course evaluation website notes that, "An essential component of our commitment to teaching excellence is the regular evaluation of courses by students." (https://courseevaluations.utoronto.ca/). Since 2012, the University of Toronto has progressively implemented an evidence-based centralized Cascaded Course Evaluation Framework (CCEF) for collecting feedback data from students. This paper reports the results of a validation study that examined the reliability and validity of the CCEF, especially in relation to the five core institutional items that make up the Institutional Composite Mean (ICM).

The Cascaded Course Evaluation Framework (CCEF)

The University of Toronto's Cascaded Course Evaluation Framework (CCEF) provides students with an opportunity to provide feedback on institutional, divisional, departmental, and instructor-identified teaching priorities. The core institutional items are included in all course evaluation surveys that use the CCEF. These items are particularly useful for understanding students' learning experiences across the University of Toronto, as they reflect five key teaching and learning priorities for the institution. The image below provides an overview of the Cascaded Course Evaluation Framework (CCEF), along with a summary of the five institutional priorities and their respective institutional course evaluation items.

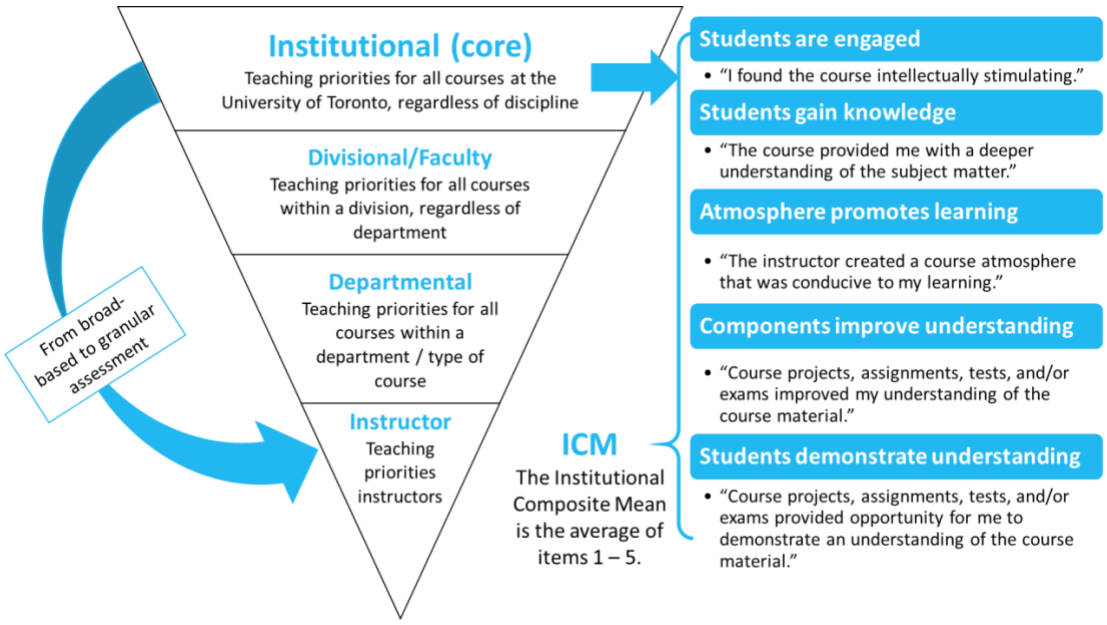


***Figure 1*. Overview of the University of Toronto's Cascaded Course Evaluation Framework (CCEF)**

The Institutional Items

The Five Core Items

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

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

The five core institutional items are rated on a 1 (Not at All) to 5 (A Great Deal) scale.

The Institutional Composite Mean

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. Item 6 is measured on a 1 (*Poor*) to 5 (*Excellent*) scale:

- **Overall learning experience in a course**: Item 6, "Overall, the quality of my learning experience in this course was: excellent (5), very good (4), good (3), fair (2), or poor (1)."

Qualitative Feedback

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

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

This current report focuses on evaluating the validity of the five core rating scale items that make up the Institutional Composite Mean (ICM) of the CCEF. The sixth institutional item was included in the analyses for comparison purposes. The two qualitative items (items 7 & 8) are not included in this validation study.

THE VALIDATION PROCESS

Validating the ICM

Validity refers to the extent to which a measurement tool assesses what it is intended to measure and can be used for its intended purpose(s). The validity of a tool cannot be determined by a single indicator; nor can any measurement tool be considered "valid" or "not valid" in a dichotomous sense. Rather, the establishment of validity is a process that involves the collection of data that supports (or refutes) the utility of a tool within specific intended uses and/or contexts (AERA/APA/NCME, 2014).

The Policy on the Student Evaluation of Teaching in Courses (2011) at U of T notes that:

> Course evaluations are part of an overall teaching and program evaluation framework that includes regular peer review, instructor self-assessment, cyclical program review and other forms of assessment, as appropriate. As part of this framework, course evaluations are a particularly useful tool for providing students with an opportunity to provide feedback on their own learning experiences.

The U of T Course Evaluations Website goes on to clarify:

> At the University of Toronto, course evaluations are conducted to collect formative data for instructors to improve their teaching, to provide summative data for administrative purposes (such as annual merit, tenure, and promotion review) and for program and curriculum review, and to provide members of the University community, including students, with information about teaching and courses at the university.

Given the stated purpose of the University of Toronto's Cascaded Course Evaluation Framework (CCEF), the current validation study was conducted to examine the utility of using the institutional items as an indicator of students' experiences with institutional teaching and learning priorities for formative and summative purposes.

Specifically, the current study sought to establish the extent to which the ICM:

1. was associated with acceptable completion rates (Response Rates),
2. produced meaningful student response patterns (Student Response Patterns),
3. was reliable across raters, items, and course-instructor pairings (Reliability),
4. was consistent with identified institutional priorities (Construct Validity),
5. reflected a unidimensional construct of student experience (Dimensionality),
6. needed to be contextualized within specific learning contexts (Contextual Analysis),
7. was not biased based on faculty characteristics (Bias Analysis),
8. allowed for meaningful interpretation of the ICM scores (Interpretability of ICM Scores),
9. could be used across teaching and learning contexts (Generalizability).

Data and Inclusion Criterion

Except for where otherwise noted, the validation study focused on all single-instructor courses evaluated in the fall and winter terms of two academic years (2015/2016 and 2016/2017) within the four largest undergraduate divisions at the University of Toronto (FASE, ARTSC, UTM, and UTSC). The final sample included 277,498 completed evaluation surveys collected from 11,919 single-instructor undergraduate course sections across 118 academic departments and units. The sample represents more than 75% of the all of the course evaluation surveys, and nearly 85% of all the undergraduate surveys, collected during the two-year time period. It is important to note that the generalizability of the results contained here is not yet conclusively determined for other divisions at the University of Toronto or time periods outside of this sampling frame. Further analyses are planned and ongoing.

Detecting meaningful effects

When working with numerical data, indicators of **statistical significance** are commonly used to examine the presence of an "effect" within the data. An "effect" might include a difference between groups and/or a specific association pattern between variables. Statistical significance indicates if an effect can be detected. It does not indicate the magnitude of the effect (nor does it indicate its theoretical or practical significance). Indeed, even trivial effects can be statistically significant when sample sizes are large enough. **Effect size** is a better indicator of the magnitude of an effect. In this study, an effect was considered meaningful only if it was (a) statistically significant **and** (b) associated with a meaningful effect size (a small effect or greater). The table below summarizes the effect size indicators used to examine the magnitude of the effects reported within this validation study.

Table 3

Effect Size Indicators

Effect size indicator		No effect	Small effect	Med. effect	Large effect
d	Cohen's *d* reports the standardized difference between two group means.	< .20	.20-.49	.50-.80	≥ .80
η^2	Eta-squared (η^2) is the magnitude of difference between two or more group means.	< .01	.01-.08	.09-.25	≥ .25
r	A correlation coefficient reports the magnitude of the association between two variables.	< .10	.10-.29	.30-.49	≥ .50
R^2	The coefficient of determination (R-squared) is the proportion of variance shared between two or more variables.	< .01	.01-.08	.09-.25	≥ .25

FINDINGS

Overview

This section provides the results of analyses that examined the extent to which the ICM:

1. was associated with acceptable completion rates (Response Rates),
2. produced meaningful student response patterns (Student Response Patterns),
3. was reliable across raters, items, and course-instructor pairings (Reliability),
4. was consistent with identified institutional priorities (Construct Validity),
5. reflected a unidimensional construct of student experience (Dimensionality),
6. needed to be contextualized within specific learning contexts (Contextual Analysis),
7. was not biased based on faculty characteristics (Demographic Analysis),
8. allowed for meaningful interpretation of the ICM scores (Interpretability of ICM Scores),
9. could be used across teaching and learning contexts (Generalizability).

1. Response Rates

A. Across course sections, the average response rate was 42%.
B. Students were more likely to submit their evaluation survey in the afternoon or evening.
C. Larger courses were associated with smaller response rates.
D. The response rates were comparable with other online surveys of student engagement.
E. Response rates were high enough to allow for general-levels of meaningful inference.
F. Student-faculty interaction, not student dissatisfaction, predicted higher response rates.
G. Response rates were not associated with survey length, fatigue, or alphabetical order.
H. Lower response rates did not meaningfully disadvantage instructors.

A. Across course sections, the average response rate was 42%

Across the two academic years, and the four undergraduate divisions, 68% of the students invited to complete a course evaluation survey completed at least one evaluation survey. In total, 36% of invited surveys were completed and submitted. At the course-section level this resulted in response rates that varied between 5% and 100% (see Figure 2 below), with an average course response rate of 43% (*S* = 17%).

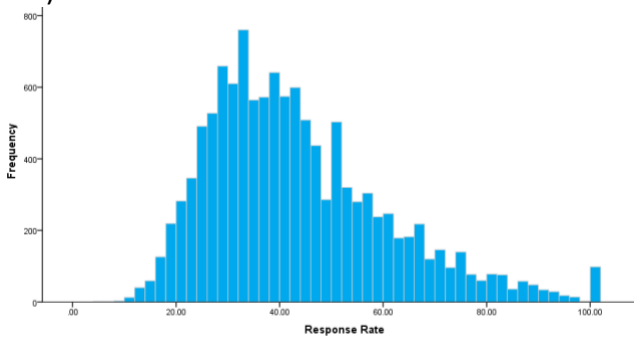


Figure 2. Spread course section level response rates

B. Students were more likely to submit their evaluation survey in the afternoon and evening

As illustrated in the figure below, students were far more likely to submit their course evaluation surveys in the afternoon and evening than in the morning or overnight. Importantly, however, ICM scores did not differ based on the time of submission.

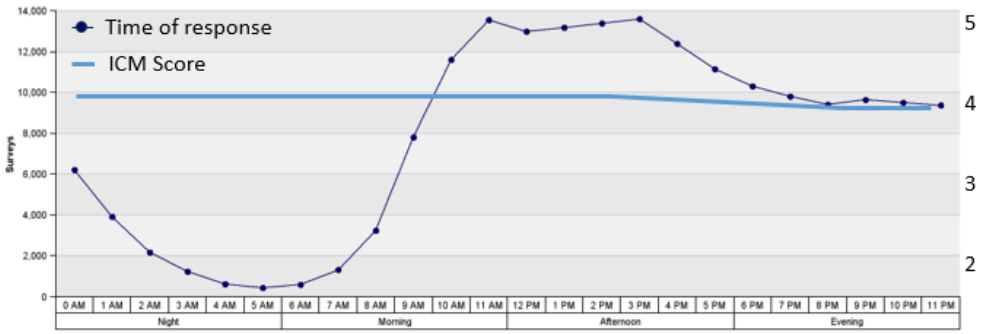


***Figure 3*. Time of day influences when surveys are submitted, but not actual ICM ratings**

C. Larger courses were associated with lower response rates

Course size was moderately, and negatively associated with response rates. In general, response rates were higher for smaller enrollment courses than for larger enrollment courses, Spearman's rho, r = -.39 (medium effect) [1].

Table 4

Average Course-Section Response Rates by Course Size

1-25	26-50	51-100	101-200	200+
50%	44%	38%	34%	32%

D. The response rates were comparable with other online surveys of student engagement

The average course response rates associated with the U of T Cascaded Course Evaluation Framework are very consistent with the response rates found with online course evaluation frameworks (Goos & Salomons, 2017), surveys of student engagement (NSSE, 2016), and other forms of online survey research (Cook et al., 2000; Shih & Fan, 2008; 2009).

For example, the National Survey of Student Engagement (NSSE) is a survey of student engagement used across the United States and Canada. The University of Toronto participates in the NSSE survey, along with 15 other research-intensive Canadian Universities that are used as comparators (U15). In 2011, the average response rate for the U15 and Ontario Universities was **32%**. The response rate for the University of Toronto was slightly higher at **40%** (University of Toronto NSSE Report, 2012). These response rates are comparable to the average course section response rates of the CCEF.

[1] Spearman rho is a statistical test used to examine the association between two variables that are measured at the ordinal level or higher. It is interpreted the same as the correlation coefficient *r*.

Furthermore, a 42% average response rate is very consistent with the average response rates found with online survey research, in general. For example, Cook et al. (2000) engaged in a meta-analytic examination of 68 online surveys (across numerous research and evaluation contexts). Across these studies, the researchers found an average response rate of **39.6%**. Similarly, in two meta-analytic reviews of 39 and 34 online surveys, Shih and Fan (2008; 2009) found average online response rates of **34%** and **33%**. The response rates associated with the University of Toronto's Cascaded Course Evaluation Framework are highly consistent with those response rates.

E. Response rates were high enough to allow for general-levels of meaningful inference

Higher response rates ensure more accurate inference

Course evaluation data are often simplified down to a summary statistic, such as the ICM. When this is the case, the summary statistic is thought to be a representation of a collective whole. In our current context, for example, the ICM is thought to reflect students' collective experiences with the institutional teaching and learning priorities. When response rates are 100%, one can trust that this summary statistic captures the "true" experiences of the collective, as the data represents the voice of everyone in that collective. When response rates are lower than 100%, then one must use the data that one has to make an estimate of the collective experience. This estimation process is subject to measurement error. Consequently, the more data points that one has, the more confident one can be about the estimate.

The minimum response rate required depends on interpretation goals

The minimum response rate required to use course evaluation data to make meaningful inferences depends on one's interpretation goals. If the goal is to make a general estimate of the collective experience of students in a course, then meaningful inference can be made across a broad range of responses rates (e.g., in formative evaluation). If the goal is to make a very precise estimate of the collective experience of students, then a larger response rate will be required (e.g., in summative evaluation). Because larger courses result in more data than smaller courses, even at comparable response rates, it is easier to achieve precise levels of estimation with larger courses than with smaller courses.

The table below shows the response rates that would be required to make what we have opted to label "very precise", "precise", "somewhat precise", or more "general estimates" about students' collective experiences in a course. Please note that the response rate required to achieve a certain level of interpretation varies based on the size of the course, with smaller courses requiring larger response rates to achieve the same level of interpretation as larger courses.

Table 5

Response Rate Needed to Make Meaningful Inference

		Course Size				
Interval around the mean	**Recommended interpretation of the quality of the mean estimate**	1-25	26-50	51-100	101-200	200+
< ±0.1	Very precise estimate	>90%	>80%	>80%	>60%	>50%
< ±0.2	Precise estimate	>80%	>70%	>70%	>50%	>40%
< ±0.5	Somewhat precise estimate	>70%	>50%	>40%	>20%	>10%
< ±1.0	General estimate	>60%	>20%	>10%	>10%	>10%
> 1.0+	Very general estimate	< 30%	<10%	<5%	<3%	<1%

Note. Guidelines are based on a 95% confidence interval around the mean with margin of errors ranging from ±0.1 to ±1.0, a standard deviation of 1.0, and correction for the use of a finite population.

Example. In a course with 75 students, a response rate near 80% would allow for a "very precise" estimate of the collective experiences of the students in the course. If the ICM was 4.0, one could feel confident that the key institutional teaching and learning priorities were "mostly" a part of the students' classroom experience. On the other hand, if the response rate was closer to 30%, then a more general estimate would be appropriate. In this case, an ICM value of 4.0 would indicate that the key institutional teaching and learning priorities were "moderately" to "a great deal" a part of the average students' classroom experience.

Response rates were high enough to allow for at least general-level inference

The table below summarizes the percentage of course sections within the sample that fell within each interpretation category. **Almost all of the courses (96%) had response rates high enough to allow for at least a "general" level of inference**. The majority of course sections (68%) had a response rate high enough to allow for a "somewhat" to "very precise" estimate of the students' collective experience in the course. Only 4% of course sections had response rates so low as to render the course evaluation results "very general" or "non-diagnostic". Overall, the ICM can be considered a general indicator of where students' collective experiences fall on the 5-point scale.

Table 6

Percent of Course Sections Falling into Each Interpretation Category

Interpretation		Percent
< ±0.1	Very precise estimate	3%
< ±0.2	Precise estimate	25%
< ±0.5	Somewhat precise estimate	38%
< ±1.0	General estimate	29%
> ±1.0	Very general estimate	4%

F. Student-faculty interaction, not student dissatisfaction, predicted higher response rates

To examine if students' perceptions predicted responses rates, 207 division, unit, and instructor course evaluation items were grouped into 27 composite variables (see Section 4 Findings: Construct Validity). Five variables meaningfully correlated with response rates ($r > .30$). Response rates were higher when students perceived the instructor to be available to students, $r = .35$, and concerned about student learning, $r = .30$. Response rates were also positively correlated with perceptions of quality assessments, $r = .33$, and feedback, $r = .32$. Courses that included more collaborative interaction also had higher response rates, $r = .30$. Importantly, response rates were not meaningfully correlated with collective perceptions of expected grades, $r = .11$ (very small effect), workload, $r = .08$ (no effect), or attendance rates, $r = .05$. These results suggest that higher student-faculty interaction, not student dissatisfaction, predicted response rates.

G. Response rates were not associated with survey length, fatigue, or alphabetical order

Survey length does not lower response rates

The University of Toronto's Cascaded Course Evaluation Framework allows for a maximum of 20 items. Each course evaluation survey that was part of this study contained 9 to 19 rating scale items. Of these, 0 to 3 items were instructor-selected items included for formative purposes only. **There was no correlation between response rates and the number of institutional items (6 items), division items (3 to 7 items), and department/unit items (0 to 8 items) pre-populated on the survey**, Spearman rho, $r = .04$ (no effect).

There was, however, a small positive correlation between the number of instructor-selected items (0 to 3 items) and response rates, $r = .15$ (small effect). Specifically, faculty who added three instructor-selected items had, on average, a 5% higher response rate than faculty who did not add any instructor-selected items. The presence of a **positive** correlation provides support **against** the assumption that faculty will suffer a response rate penalty if they choose to add instructor-selected items to their course evaluation surveys. Neither the total length of the survey used at the University of Toronto, nor the addition of instructor-selected items, lowered response rates.

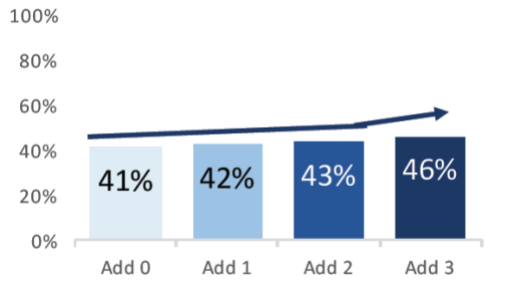


Figure 4. Faculty who added instructor items had higher response rates.

Survey fatigue does not lower response rates

A special analysis of the five-year response rate trends for UTSC and UTM undergraduate students revealed no evidence that students experienced survey fatigue as a consequence of being invited to complete multiple course evaluation surveys. Indeed, the correlation between the number of

invitations received and response rates was weak, but **positive**, Spearman's rho, *r* = 0.13 (UTSC, small effect), *r* = 0.12 (UTM, small effect). If anything, the likelihood of responding **increased**, rather than decreased, the longer the students were evaluating courses using the University of Toronto evaluation system (although this effect was small). There was also no meaningful correlation between year of study and actual ICM scores (Spearman's rho, *r* = 0.08, no effect).

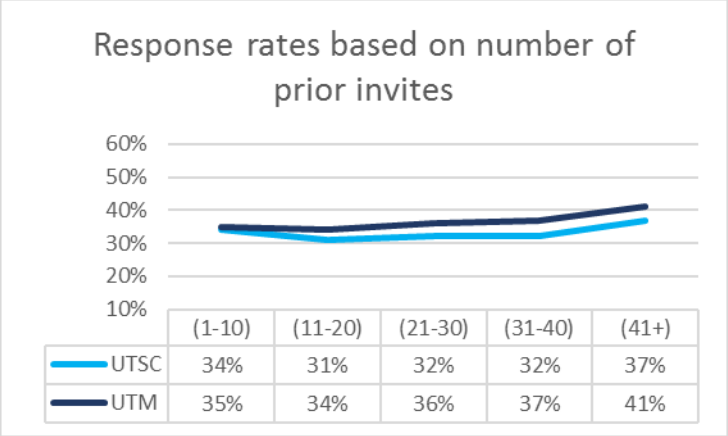


Figure 5. Response rates based on the number of prior invitations

The alphabetical order of the survey did not hurt response rates

The alphabetical order in which a course evaluation survey was listed was not associated with response rates or ICM scores.

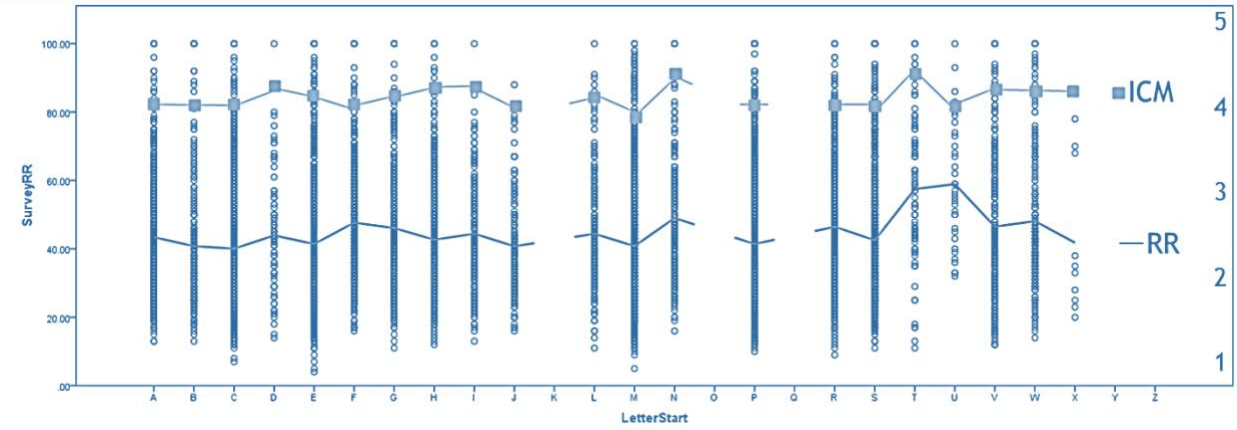


Figure 6. Average response rate and ICM score based on alphabetical letter

H. Lower response rates did not meaningfully disadvantage instructors

Across course sizes, the correlation between response rates and ICM scores was small (see the table below). When controlling for course size, there was less than a 0.1 difference in ICM scores between course sections with high response rates (ICM, *M* = 4.0) and course sections with low response rates (ICM, *M* = 3.9). Instructors with low response rates do not appear to be at a meaningful disadvantage relative to peers with higher response rates.

Table 7

Correlation Between Response Rates and ICM Scores by Course Size

	1-25 students	26-50 students	51-100 students	101-200 students	200+ students
Spearman's rho, *r* =	.14 (small)	.15 (small)	.17 (small)	.15 (small)	.15 (small)

2. Student Response Patterns

A. Students rated an average of 99% of the rating scale items presented to them.
B. Students did not engage in wide-spread yea-saying, nay-saying, or neutral responding.
C. Students were responsive to shifting scale options.
D. Students favoured the upper end of the rating scale.
E. Rates of endorsement were within recommended levels.

A. Students rated an average of 99% of the rating scale items presented to them

Each course evaluation survey contained 9 to 19 rating scale items. Respondents who opted to complete an evaluation survey tended to complete all 9 to 19 rating scale items. Indeed, students rated an average of 99% of the rating scale items presented to them in their course evaluation surveys. The number of rating scale items present in the survey was not at all correlated with the completion rate (*r's* = .001 to -.03, no effects).

B. Students do not engage in wide-spread yea-saying, nay-saying, or neutral responding

Contrary to fears that respondents engage in mindless "down the line" responding, only 2% of respondents gave the same uniform response across all of the rating scale items. The other 98% of respondents showed at least some nuance in their ratings. Even when considering just the six institutional rating scale items (which focus on similar teaching and learning priorities), only 25% of respondents gave the same rating to all six items. The other 75% of respondents showed nuance in their ratings. When students did engage in uniform responding, they were far more likely to engage in yea-saying (assigning all "5"s) than in nay-saying (all "1"s) or neutral-responding (all "3"s).

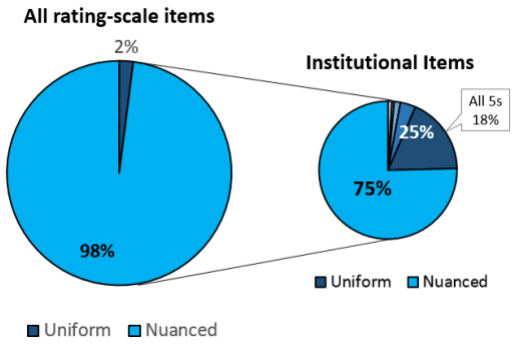


Figure 7. Uniform versus nuanced responding across the items

C. Students are responsive to shifting scale options

Follow-up analyses with the ARTSC division items demonstrated that students were responsive to shifting scale options. Indeed, even though the ARTSC division items used a different scale orientation

from the institutional items in 2015/2016, and underwent a shift in scale orientations between 2015/2016 and 2016/2017, respondents chose each scale option by the same frequency and gave the same ratings to the items, regardless of the orientation used. These analyses suggest that respondents appropriately adjusted their ratings according to the scale, regardless of its orientation.

D. Students favour the upper end of the rating scale

Across the six institutional items, respondents were far more likely to give a response at the upper end of the scale relative to the bottom end of the scale. Indeed, 69% of the time respondents gave a rating of "4" or "5" to one of the institutional items (*M* = 3.9, *S* = 1.0, *mdn* = 4.0). This provides evidence against the belief that only disgruntled students complete the course evaluation survey.

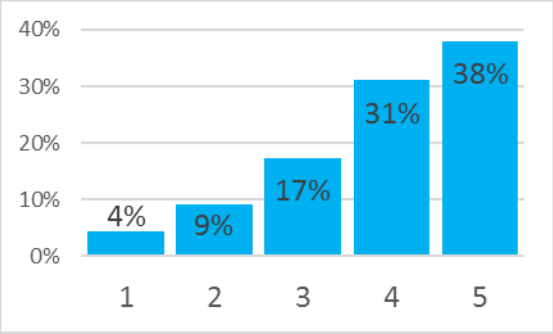


Figure 8. Percentage of students assigning each score

E. Rates of endorsement were within recommended levels

Because the response patterns were skewed (with respondents favouring the upper end of the scale relative to the bottom end of the scale), it was important to examine if the skewed distribution resulted in a restriction in the range of responses. Streiner and Norman urge the reconsideration of any item whose rate of endorsement is outside of 20% to 80%. Using the method recommended by Nulty (2008), the rate of endorsement was calculated by combining scores of "4" and "5" on the 5-point scale. Across the six items, the rate of endorsement ranged from 53% to 72%. These rates of endorsement were well within Streiner and Norman's recommended rate of endorsement (20% to 80%). In addition, they were highly consistent with the 70% rate of endorsement found in other course evaluation surveys (Nulty, 2008; Zumrawi, Bates, & Schroeder, 2014). The standard deviation of 1.0 indicates that, on average, any given rating was within ±1.0 points from the mean of 3.9. On a 5-point scale this indicates moderate variability in the responses. Extreme restriction of range does not appear to be a problem with the institutional items of the CCEF.

3. Reliability

A. **Interrater reliability**. Students within a single course exhibited high enough reliability in their ratings of the institutional items to justify aggregating these ratings to the course-section level for interpretation.
B. **Internal consistency**. The five items of the ICM exhibited high enough internal consistency to justify averaging the items into an Institutional Composite Mean (ICM).
C. **Test-retest reliability**. ICM scores were most stable when considering the same instructor teaching the same course over time.

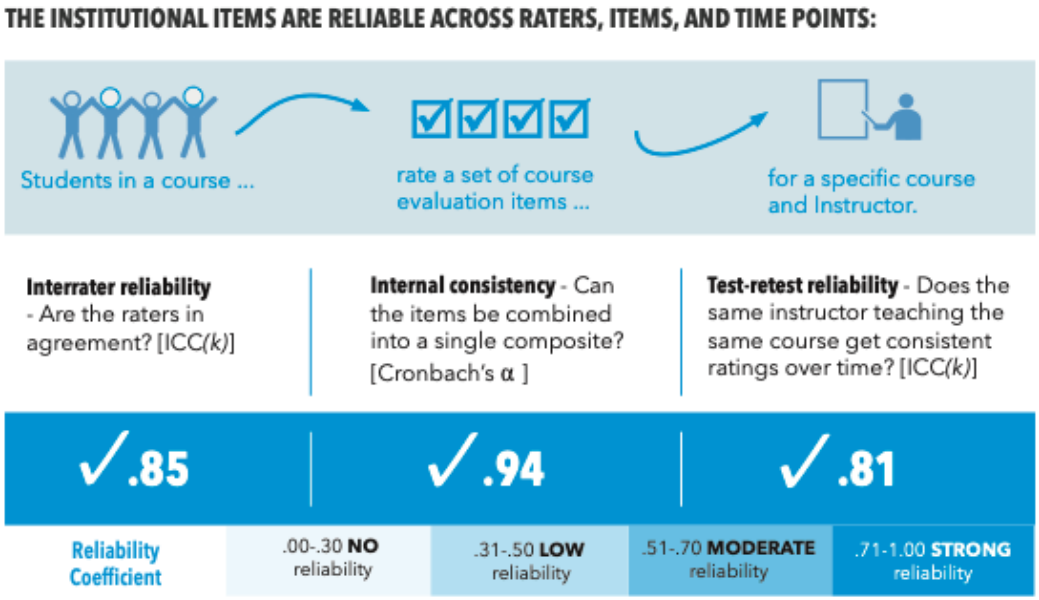


Figure 9. Infographic of the reliability of the ICM across rates, items, and time points

Measuring Reliability

Reliability refers to the stability of a measurement over multiple raters, items, time points, and/or other repetitions of measurement. Most reliability coefficients vary between 0 and 1, with a score of 0.00 indicating a complete lack of reliability and a score of 1.00 indicating perfect reliability. Benchmarks vary, but reliability coefficients are typically considered acceptable/good starting at around 0.70 or higher (LeBreton & Sentor, 2008).

A. Students exhibited strong interrater reliability

Interrater reliability examines the extent to which raters assessing the same target show agreement in their ratings. Across all four divisions, students exhibited acceptable agreement in their course-section ratings of the six institutional items ($r_{wg} \geq 0.72$) and very strong agreement when the first five items were considered together as an Institutional Composite Mean (ICM, $r_{wg} \geq 0.92$). Students' absolute agreement in their ratings of the institutional items (ICC(*k*) ≥ 0.78) and the ICM (ICC(*k*) ≥ 0.85) were high enough to differentiate between different course-sections. These results provide strong support for the aggregation and interpretation of data at the section-level of analysis.

B. The five items of the ICM exhibited high internal consistency

Internal consistency examines the extent to which different items are internally consistent enough to justify averaging them together to create a single composite score. The five items that make up the ICM were all highly correlated with one another, *r*'s ≥ 0.77. These five items also exhibited high item-total correlations with the ICM, *r*'s $\geq .91$. A factor analysis using principle axis factoring demonstrated that the five items of the ICM loaded on to a single factor explaining more than 83% of the variance. Factor loadings for each item exceeded .80. The five items also exhibited very strong internal consistency, with Cronbach's $\alpha \geq .94$. These results suggest that the first five items were internally consistent enough to be averaged together into a single Institutional Composite Mean (ICM).

Table 8

Correlations Between Ratings on the Six Institutional Items, The ICM, And Response Rates

Student Engagement	I1	I2	I3	I4	I5	I6	ICM	RR
1. Intellectually stimulating	----	0.89	0.80	0.78	0.77	0.86	0.92	0.25
2. Deeper understanding	0.89	----	0.81	0.81	0.80	0.88	0.93	0.25
3. Learning atmosphere	0.80	0.81	----	0.78	0.77	0.89	0.91	0.28
4. Components, understanding	0.78	0.81	0.78	----	0.93	0.85	0.93	0.24
5. Components, demonstrate	0.77	0.80	0.77	0.93	----	0.85	0.92	0.25
6. Overall learning experience	0.86	0.88	0.89	0.85	0.85	----	0.94	0.27

Note. I1 – 16 = institutional items; ICM = Institutional Composite Mean; RR = response rate.

Table 9

Factor Analysis Results

	%	I1	I2	I3	I4	I5	α
Factor analysis results	**86%**	0.90	0.93	0.87	0.93	0.91	**0.96**

Note. % = percent of variance explained, I1 – 16 = institutional items; α = Cronbach's alpha.

C. The ICM exhibited strong test-retest reliability across specific course-instructor pairings

Test-retest reliability examines the extent to which the ratings for a single target stay stable over multiple measurements or time periods. When the same student rated different courses, ICM scores exhibited only moderate stability across ratings, ICC(*k*) = .63. In contrast, the stability of ICM scores for the same instructor teaching across multiple sections or terms was good, ICC(*k*) = .75, as was the stability of ICM scores for the same course topic taught across multiple sections or terms, ICC(*k*) = .72. Importantly, however, ICM scores were most stable when ratings were considered across multiple offerings of the same course topic being taught by the same course instructor over multiple sections or time periods, ICC(k) = .81. These findings suggest that ICM scores produce reliable differentiation between specific course-instructor pairings and can be interpreted as reflecting an assessment of a specific course-instructor combination.

4. Construct Validity

A. **Student engagement**: The ICM was more strongly correlated with indicators of course-created engagement than with students' prior interest in the topic or class attendance.
B. **Knowledge gains**: The ICM was more strongly correlated with students' perceived opportunities to gain knowledge than with their expected grade performance.
C. **Learning atmosphere**: The ICM was more strongly correlated with quality of instruction indicators than with course support factors.
D. **Quality of assessment:** The ICM was more strongly correlated with the quality and fairness of assessment than with the perceived workload of the course.

Defining construct validity

Construct validity assesses the extent to which an item (or a group of items) successfully measures the construct for which it was intended to measure. At the University of Toronto, the ICM is meant to

capture the extent to which a course included all five institutional teaching and learning priorities: (1) students are engaged, (2) students gain knowledge, (3) the atmosphere promotes learning, (4) course components improve understanding, and (5) course components provide opportunity to demonstrate understanding.

A key question that arises is whether students' experiences with these teaching and learning priorities contribute to the bigger picture understanding of the quality of instruction. In other words, can we assume that ratings on these items are associated with actions that are within the control of the teacher or are they associated with factors that are largely beyond the control of the instructor? To establish construct validity, one can examine if items meant to measure one construct are predictably related to theoretically similar constructs (**convergent validity**) and predictably unassociated with theoretically distinct constructs (**discriminant validity**). If the institutional items are associated with the quality of instruction**,** then ratings on the institutional items should exhibit convergent validity with factors consistent with quality instruction and discriminant validity with factors outside of the control of the instructor.

Assessing construct validity

To examine the construct validity of the institutional items, 207 division, unit, and instructor-selected items were grouped together to create 27 composite variables that captured various aspects of student engagement, knowledge gains, quality of instruction, instructional approaches, course assessment, and course supports. Spearman rho correlations were used to examine the association between the ICM and each composite variable to examine patterns of convergent validity and discriminant validity.[2]

A. Student engagement

The ICM was highly correlated with course-specific measures of engagement, including students' perceptions that the course was intellectually engaging (r = 0.86), students' levels of interest after taking the course (r = 0.91), students' willingness to recommend the course to others (r = 0.88), and whether the instructor generated enthusiasm for the topic (r = 0.79). In contrast, the institutional items were only weakly associated with students' pre-existing interest in the topic (r = .14) and their reported attendance (r = 0.25). The ICM is more strongly associated with course-created engagement than students' prior interest in the topic or rates of attendance. Furthermore, on the whole, students were more likely to report a greater interest in the course at the time of completing the survey (M = 3.7) relative to students' reported interest in the course at the time of registration (M = 3.4), suggesting the most courses at the University of Toronto are successful at piquing students' interests in the topic, Cohen's d = 0.71 (medium effect). The ICM appears to be convergent with the University of Toronto's teaching and learning priority that "students are engaged".

[2] For ease of interpretation, all effect sizes are reported as correlation coefficients (r), regardless of the type of analysis to examine the association (e.g. Spearman correlation for ranked variables, ANOVA for grouped variables, etc).

Table 10

Convergent and Discriminant Validity Patterns Around Student Engagement

Student Engagement	#	*N*	I1	I2	I3	I4	I5	I6	ICM	RR
Intellectual engagement	5	6,000	**0.86**	**0.84**	**0.77**	**0.76**	**0.75**	**0.85**	**0.86**	0.02
Interest at end of course	1	488	**0.89**	**0.88**	**0.82**	**0.84**	**0.83**	**0.91**	**0.91**	0.20
Would recommend course	1	10,934	**0.82**	**0.82**	**0.83**	**0.80**	**0.80**	**0.90**	**0.88**	0.24
Generates enthusiasm	2	7,345	**0.73**	**0.72**	**0.85**	0.67	0.67	**0.79**	**0.79**	0.29
Pre-existing interest in topic	1	811	0.22	0.16	0.07	0.09	0.08	0.14	0.14	0.02
Reported attendance	3	204	0.23	0.23	0.23	0.19	0.18	0.23	0.25	0.05

Note. I1 – 16 = institutional items; ICM = Institutional Composite Mean; RR = response rate.

B. Students gain knowledge

Course evaluations are not meant to be an indicator of actual student learning (Marsh, 2007; Spooren, Brockx, & Mortelmans, 2013). However, students' can assess if they had opportunity to gain knowledge as a consequence of the course (Marsh, 2007; Spooren et al., 2013). A challenge with designing effective course evaluations is separating out students' perceptions of their opportunity to gain knowledge from their satisfaction/dissatisfaction with their performance in the course. Indeed, students can be unhappy with their grade even when a course offered numerous opportunities for knowledge gain.

Consistent with the premise that the ICM is convergent with opportunities for students to gain knowledge the ICM was strongly correlated with students' perceptions of their overall learning experience (r = 0.94). The ICM was also convergent with perceptions that the course helped students engage in higher order thinking (r = 0.87), covered a breadth of information (r = 0.76), and connected to the larger curriculum (r = 0.73). In contrast, the ICM was only moderately correlated with students' expected grade in the course and not all correlated with the perceived workload. These results suggest that the ICM is more reflective of students' opportunity to gain knowledge, rather than with their satisfaction/dissatisfaction with their performance or the workload of the course.

Table 11

Convergent and Discriminant Validity Patterns Around Knowledge Gains

Knowledge gains	#	N	I1	I2	I3	I4	I5	I6	ICM	RR
Higher order learning	8	2,503	**0.82**	**0.82**	**0.79**	**0.81**	**0.81**	**0.85**	**0.87**	0.24
Breadth of information	7	2,614	0.68	**0.72**	**0.76**	0.66	0.67	**0.75**	**0.76**	0.25
Connects to curriculum	5	1,549	0.66	0.66	**0.73**	0.64	0.65	**0.74**	**0.73**	0.14
Expected grade	1	17,772	0.25	0.27	0.26	0.31	0.32	0.37	0.33	0.11
Perceived workload	1	11,119	0.10	0.05	-0.04	0.03	-0.01	-0.04	0.03	0.08

Note. I1 – 16 = institutional items; ICM = Institutional Composite Mean; RR = response rate.

C. Learning atmosphere

The ICM was correlated with students' perceptions of the clarity of instruction (r = 0.85) and the extent to which the instructor promoted learning (r = 0.83), was available to students (r = 0.79), and

demonstrated respect for students (r = 0.69). In contrast, the ICM exhibited more moderate correlations with students' perceptions of the lab and tutorial (r = 0.59), classroom resources (r = 0.55), the quality of teaching assistants (r = 0.55), and specific instructional approaches (r's ranged from 0.50 to 0.67). These results suggest that ICM scores are more strongly correlated with factors that are within the control of the instructor (i.e., clarity of instruction, promoting learning, and demonstrating respect) than outside of the control of the instructor (i.e., quality of the lab, resources, and TA). The ICM appears to be convergent with the University of Toronto's teaching and learning priority that instructors create a safe and effective "learning atmosphere".

Table 12

Convergent and Discriminant Validity Patterns Around Learning Atmosphere

Atmosphere	#	N	I1	I2	I3	I4	I5	I6	ICM	RR
Clarity of instruction	10	6,958	**0.74**	**0.79**	**0.89**	**0.73**	**0.73**	**0.85**	**0.85**	0.25
Promotes learning	6	3,135	**0.73**	**0.76**	**0.85**	**0.72**	**0.75**	**0.83**	**0.83**	0.30
Available to students	3	2,220	0.65	0.67	**0.80**	0.69	**0.71**	**0.75**	**0.76**	0.35
Respects students	7	1,671	0.63	0.61	**0.72**	0.60	0.62	0.67	0.69	0.23

Note. I1 – 16 = institutional items; ICM = Institutional Composite Mean; RR = response rate.

Table 13

Convergent and Discriminant Validity Patterns Around Specific Instructional Approaches

Approaches	#	N	I1	I2	I3	I4	I5	I6	ICM	RR
Discussion / interaction	9	1,977	0.62	0.60	**0.70**	0.58	0.61	0.68	0.67	0.30
Communication skills	29	1,775	0.60	0.58	0.60	0.63	0.63	0.64	0.66	0.25
Research skills	27	157	0.50	0.58	0.64	0.62	0.62	0.59	0.64	0.14
Use of technology	31	431	0.58	0.57	0.66	0.57	0.57	0.65	0.64	0.24
Active learning strategies	10	344	0.50	0.53	0.62	0.57	0.56	0.57	0.62	0.14
Professional practice	6	1,108	0.39	0.45	0.52	0.42	0.48	0.54	0.50	0.05

Note. I1 – 16 = institutional items; ICM = Institutional Composite Mean; RR = response rate.

Table 14

Convergent and Discriminant Validity Patterns Around Course Supports

Course Supports	#	N	I1	I2	I3	I4	I5	I6	ICM	RR
Lab and tutorial	5	375	0.45	0.48	0.51	0.61	0.60	0.61	0.59	0.17
Resources (space, text)	4	261	0.48	0.52	0.49	0.53	0.55	0.52	0.55	0.24
Teaching assistant(s)	15	549	0.46	0.46	0.54	0.53	0.51	0.52	0.55	-0.04

Note. I1 – 16 = institutional items; ICM = Institutional Composite Mean; RR = response rate.

D. Quality of assessment

The ICM was correlated with items that captured the extent to which course components facilitated learning and improved understanding (r = 0.80) and course assessments were fair (r = 0.75). In contrast, the ICM was not correlated at all with the perceived workload of the course (r = 0.03). Quality and perceived fairness of assessment were stronger predictors of ICM scores than the perceived

workload of the course. The ICM appears to be convergent with the University of Toronto's teaching and learning priority that "course components improve understanding" and "course components provide opportunity to demonstrate understanding".

Table 15

Convergent and Discriminant Validity Patterns Around Course Assessment

Course Assessment	#	N	I1	I2	I3	I4	I5	I6	ICM	RR
Learn from components	4	1,305	0.69	**0.70**	**0.74**	**0.81**	**0.81**	**0.78**	**0.80**	0.33
Fairness of assessment	6	2,441	0.61	0.65	**0.70**	**0.74**	**0.77**	**0.73**	**0.75**	0.32
Perceived workload	1	11,119	0.10	0.05	-0.04	0.03	-0.01	-0.04	0.03	0.08

Note. I1 – 16 = institutional items; ICM = Institutional Composite Mean; RR = response rate.

5. Dimensionality

A. The ICM is more reliable and stable than the institutional items considered individually.
B. The ICM exhibits stronger construct validity than any given institutional item.
C. The ICM is better at differentiating between course sections than any individual item.
D. The ICM is more appropriately used for summative purposes than individual items.

Are the core institutional items measuring one construct or multiple constructs?

The five core institutional items of the CCEF were written to capture five teaching and learning priorities at the University of Toronto. These teaching and learning priorities are similar to one another in that all five capture students' learning experiences. However, each item also taps into slightly different aspects of the learning experience. This raises the question: is there utility in examining the core institutional items separate from one another or should the core institutional items always be considered collectively within the form of the Institutional Composite Mean (ICM)?

A. The ICM is more reliable and stable than the institutional items considered individually

An examination of the reliability analyses from earlier (see Reliability) certainly suggest that the five items considered together (as the ICM) produce a more stable measure, as the ICM was associated with higher interrater reliability and test-retest reliability than any given item considered individually. In addition, the five items of the Institutional Composite Mean (ICM) were highly correlated with one another and exhibited very strong internal consistency. These results suggest that it may be advantageous to examine the ICM as a unidimensional construct.

B. The ICM exhibits stronger construct validity than any given institutional item

Another way to approach the issue is to examine unique construct validity patterns. Table 16 summarizes five of the top findings from the construct validity analysis. The results suggest some differentiation between the items in their key correlation patterns. However, there is also a lot of overlap in the correlational patterns between the items. In contrast, the ICM is consistently correlated with all five of the key construct validity variables.

Table 16

Correlation Between Institutional Items and Key Construct Variables

Construct variable	I1 engagement	I2 knowledge	I3 atmosphere	I4 components	I5 demonstrate	ICM
Intellectual engagement	**0.86**	0.84	0.77	0.76	0.75	**0.86**
Higher order learning	0.82	**0.82**	0.79	0.81	0.81	**0.87**
Clarity of instruction	0.74	0.79	**0.89**	0.73	0.73	**0.85**
Learn from components	0.69	0.70	0.74	**0.81**	0.81	**0.80**
Fairness of assessment	0.61	0.65	0.70	0.74	**0.77**	**0.75**

Note. I1 – 16 = institutional items; ICM = Institutional Composite Mean; RR = response rate.

C. The ICM is better at differentiating between course sections than any individual item

Profile analyses (Marsh & Bailey, 1993) and reliability analyses (Morley, 2009) were used to examine if course sections could be differentiated from one another based on their unique rating patterns across the five items. The analyses revealed moderate effect sizes ($\eta^2 \geq .10$) and moderate consistency in the pattern of ratings across the institutional items, ICC(*C,k*) = 0.61. However, there was relatively low absolute agreement in the pattern of ratings across the items, ICC(*A,k*) = 0.42. These results indicate that a unique pattern of findings could be detected across the five institutional items for different instructor-course section pairings, but the pattern was not strong enough to recommend making high-stakes decisions based on these differences. In comparison, students exhibited high absolute agreement when the institutional items were considered together as a single Institutional Composite Mean (ICM), ICC(*k*) $\geq$ 0.85.

D. The ICM is more appropriately used for summative purposes than individual items

These results suggest that mean differences between the core institutional items can be considered for low-stakes formative purposes to inform the improvement of teaching and learning. However, when course evaluation scores are to be used as a piece of evidence to inform high-stakes decision making, it is better to interpret the institutional items holistically by using the Institutional Composite Mean (ICM). Relative to the use of any given item individually, the ICM is a more reliable, stable, and diagnostic indicator of students' experiences with the institutional teaching and learning priorities.

6. Contextual analysis

A. Larger course sizes were moderately associated with lower ICM scores.
B. Course level predicted ICM scores, but mainly due to course size differences.
C. ICM differences between academic division were trivial, and mostly due to course size.
D. ICM scores differed between academic units, but mostly due to course size.
E. ICM differences between course formats were trivial, and mostly due to course size.
F. ICM scores were not associated with course length or the course term.
G. ICM scores were not associated with students' full time status or year of study.

A. Larger course sizes were associated with lower ICM scores

Larger course sizes were associated with lower ICM scores, r = -0.41 (moderate effect size). On average, smaller courses had ICM scores 0.5 points higher than large enrollment courses (200+ students). Course size should be taken into consideration when interpreting ICM scores.

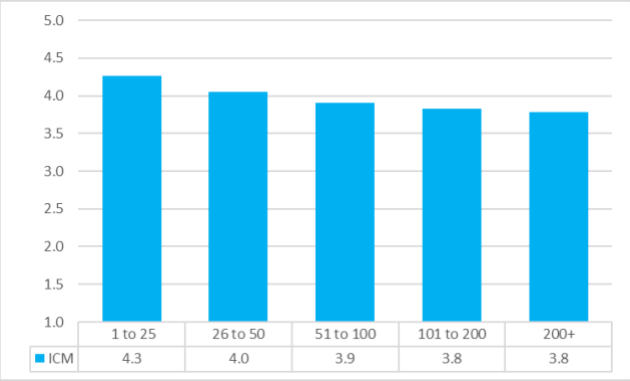


Figure 10. Course size and ICM scores

B. Course level predicted ICM scores, but mainly due to course size differences

Course level was positively correlated with ICM scores, η^2 = .04 (small effect). At the most extreme, course evaluation scores for 400/500-level courses (M = 4.24, S = 0.52) were 0.3 points higher than course evaluation scores for 200-level courses (M = 3.91, S = 0.49). Importantly, however, those differences were almost entirely explained by differences in course size. As the level of the course gets higher, the size of the course also gets smaller, r = -.42. When class size is taken into consideration, the association between course-level and ICM scores are even smaller, η^2 = .02 (very small effect). These results suggest that course level is not a strong correlate of ICM scores once course size is taken into consideration.

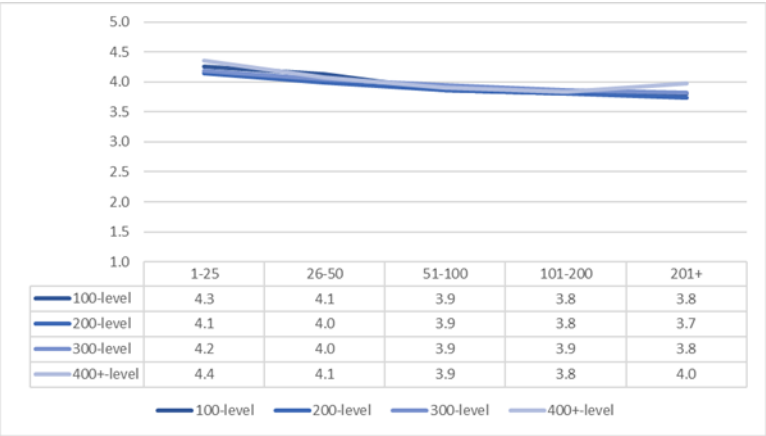


Figure 11. Course level, course size, and ICM scores

C. ICM differences between academic divisions were trivial, and mostly due to course size

There were only trivial differences in ICM scores between the four academic divisions, η^2 = .02 (small effect): FASE (M = 3.87, S = 0.50), ARTSC (M = 4.10, S = 0.51), UTM (M = 4.01, S = 0.52), UTSC (M = 4.02, S = 0.52). Once course size was taken into consideration, these differences became even smaller, η^2 = .01 (very small effect).

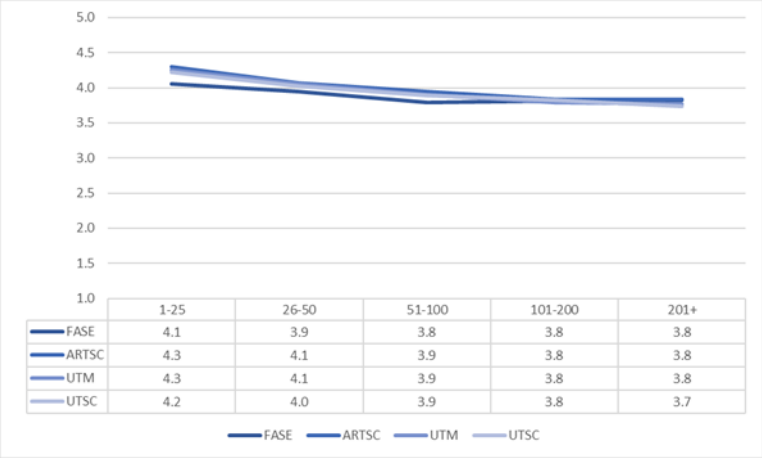
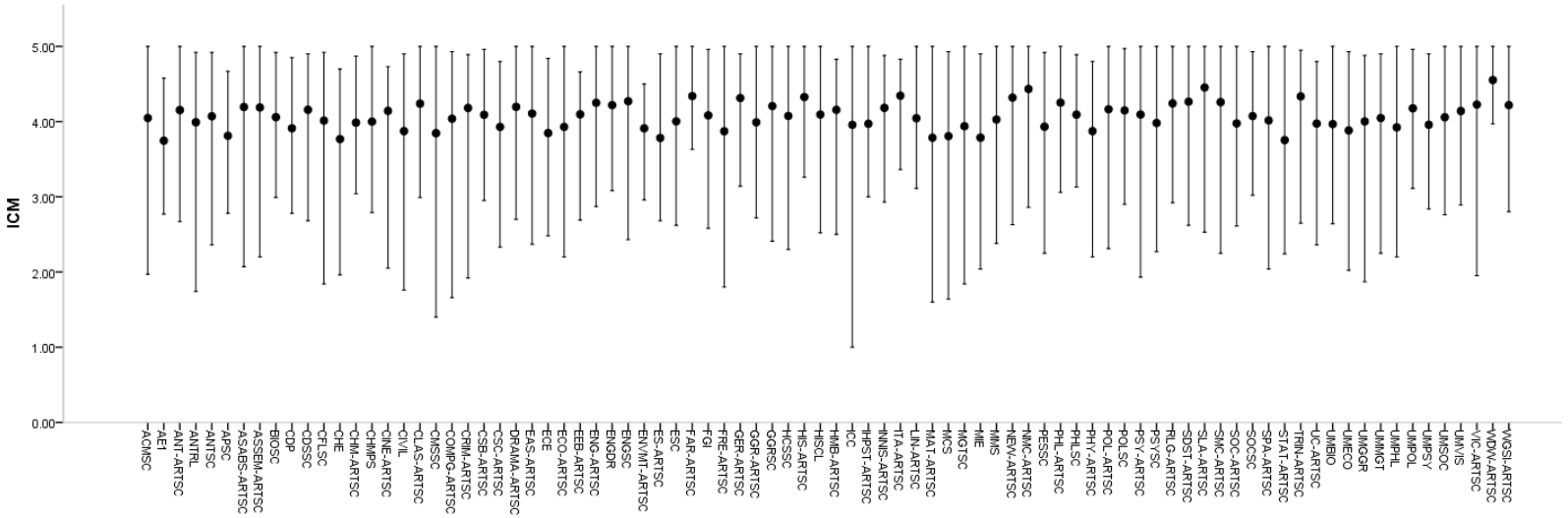


Figure 12. Division, course size, and ICM scores

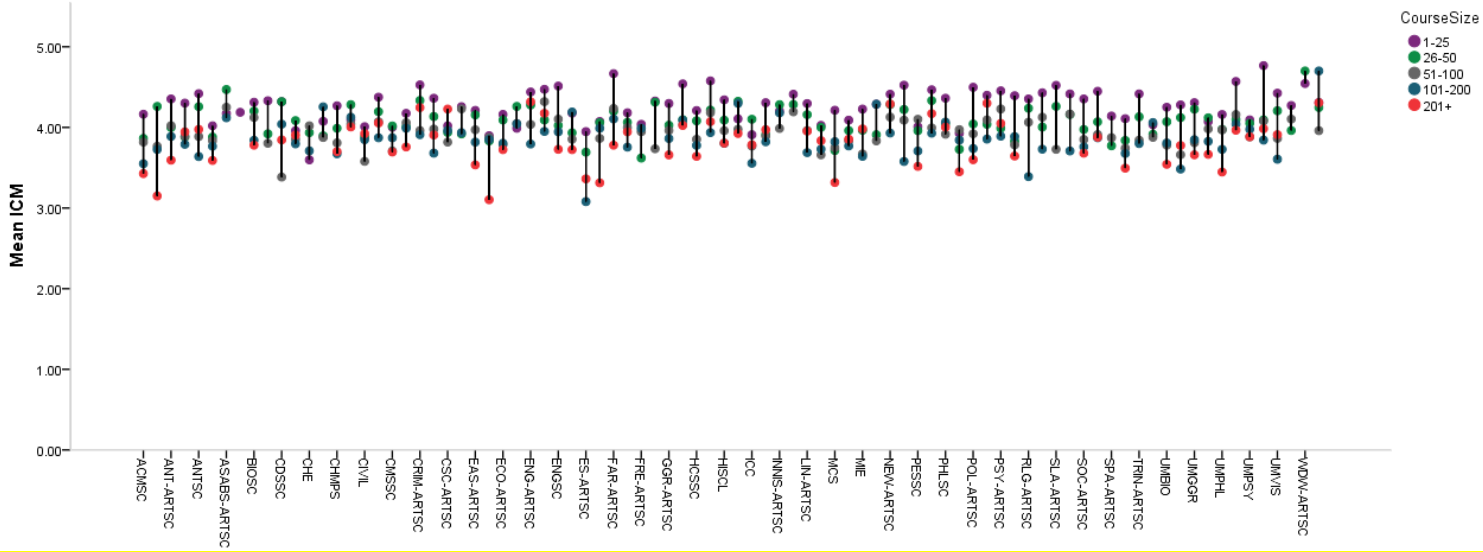
D. ICM scores differed between academic units, but mostly due to course size

The sample represented 118 academic units and departments. Each unit/department evaluated anywhere from 1 to 513 course sections in the two-year period under consideration. Units that evaluated more than 30 course sections were included in the analysis (N = 87). Overall, ICM scores differed somewhat between academic units (η^2 = 0.10, moderate effect). However, some of this variation was explained by course size, as ICM differences between academic units exhibited only trivial differences once course size was taken into consideration (η^2 = 0.04, small effect). Although the differences are small, academic unit should be considered an important contextual variable when interpreting ICM scores.



Note. The lines represent the range of ICM values within an academic unit. The circles indicate the ICM average for that unit.

Figure 13. ICM differences by academic unit



Note. The circles represent the ICM average for any given course size within a particular academic unit. The lines connect the means across the five course size categories.

Figure 14. ICM differences by academic unit and course size.

E. ICM differences between course formats were trivial, and mostly due to course size

The vast majority of evaluated course sections were labeled as "lecture" courses in the registration system (*N* = 11,442, 96%). The remaining courses could be identified as evening courses (*N* = 249), practicums (*N* = 79), tutorials (*N* = 31), and web-option courses (*N* = 118). It is important to note that "lecture" is the default label. As such, courses labeled 'lecture' courses could actually be evening, practicum, tutorial, or web-option courses (or some other format). Given this, the course format analyses should be interpreted with caution.

In general, ICM differences by course format tended to be trivial (η^2 = .005, no effect) and these differences became even smaller once course size was taken into consideration (η^2 = .002, no effect). If anything, tutorial courses received slightly lower ICM scores than other course types of similar size. Overall, however, different course types of similar sizes received similar ICM scores.

Table 17

ICM Scores by Course Type and Course Size

Course size	Lecture	Evening	Practicum	Tutorial	Web-option
1-25	4.3	4.1	4.2	3.7	----
26-50	4.0	4.0	4.0	----	----
51-100	3.9	3.9	3.8	----	3.8
100-200	3.8	3.8	----	----	3.7
201+	3.8	3.6	----	----	3.7
Overall	**4.1**	**4.0**	**4.0**	**3.7**	**3.7**

Note. ICM averages for a category are included in this table only if there were at least 10 course sections evaluated within that category.

F. ICM scores were not associated with course length or the course term

Half year courses (*N* = 10,501, 88%) meet for either the fall term (*N* = 5,128) or winter term (*N* = 5,373). Full year courses (*N* = 1,418, 12%) meet for both the fall and winter term. Full year courses are typically evaluated in the winter term (*N* = 1,225), although a few full year courses are evaluated in the fall term (*N* = 193). Neither the length of the course (half year or full year, $\eta^2 < .001$, no effect) nor the term of the course (fall or winter, η^2 = .001, no effect) was associated with ICM scores.

G. ICM scores were not associated with students' full time status or year of study

Our course evaluation system registers whether or not a student submitting an evaluation is a full time or part time student and their year of study. The current sample included 277,498 surveys submitted by 54,108 students. Student characteristics were analyzed at the survey-level of analysis using multilevel modeling (these analyses accounted for the fact that students were nested within specific course sections).

Full time/part time status

Of the 54,108 students that submitted course evaluation surveys in the two-year period under consideration, 93% of them were registered as full time students. Only trivial differences were found

between the average ICM scores of full-time students (M = 3.9, S = 0.97) versus part-time students (M = 4.1, S = 0.94), η^2 = .001 (no effect).

Year of study

Only trivial differences were found between the average ICM scores of first year students (M = 3.9, S = 0.93), second year students (M = 3.9, S = 0.99), third year students (M = 4.0, S = 0.99), fourth year students (M = 4.0, S = 0.97) and fifth year students and beyond (M = 4.0, S = 1.01), η^2 = .001 (no effect).

7. Demographic Analysis

A. No gender differences emerged on response rates or institutional item ratings.
B. ICM scores were not associated with faculty rank, age, or seniority.

A. No gender differences emerged on response rates or institutional item ratings

Gender bias is a recognized and acknowledged issue in the academy (undergraduate and graduate student assessment, faculty teaching assessment, faculty research assessment, etc.; see, for example, Eagan & Garvey, 2015). Course evaluation bias is most likely to arise in situations that utilize ambiguously or poorly worded survey questions. If students are asked to make judgements about domains that they cannot accurately assess, students will be more likely to fall back on gender stereotypes to make these assessments (Marsh, 2007).

From its earliest inception, the University of Toronto's Cascaded Course Evaluation Framework was designed to create a responsive and evidence-based approach to course evaluations explicitly designed to minimize the impact of this type of rating bias. In doing so, the course evaluations team took care to focus on students' experiences with specific teaching and learning priorities and to avoid questions known to be biased. For instance, questions related to instructor personality traits and/or domain knowledge are not used.

Adopting an evidence-based approach to survey item creation was an important first step in reducing gender bias. The monitoring of data for gender bias is another important step. For privacy reasons, the course evaluations team does not record instructor characteristics. However, the University of Toronto's Business Intelligence (UTBI) data warehouse allows a limited group of authorized users to generate anonymized tables[3] summarizing course evaluation results by faculty gender, rank, age, and years since faculty appointment (seniority).

The table below summarizes the aggregated mean averages for female versus male instructors for each of the institutional items and the ICM drawn from the UTBI data warehouse using Cognos. No systematic gender differences emerged based on survey response rates, ratings on the six institutional items, or the ICM.

[3] **Please note**: The method that is used to aggregate data in Cognos (the tool used to query the UTBI data warehouse) is restricted to the "survey level of analysis" (each survey is considered the unit of analysis). It was therefore not possible to aggregate to the "course section level" as was done in some of the prior analyses in this document.

Table 18

No Gender Differences Emerge at The Institutional Level

Gender	RR	I1	I2	I3	I4	I5	I6	ICM
Female instructor	39%	3.9	4.0	4.0	3.8	3.8	3.6	3.9
Male instructor	38%	3.9	4.0	4.0	3.8	3.8	3.6	3.9

Note. I1 – 16 = institutional items; ICM = Institutional Composite Mean.

B. ICM scores were not associated with faculty rank, age, or seniority

ICM scores did not differ based on faculty rank, age, or seniority.

Table 19

ICM Scores Did Not Differ Based On Faculty Rank, Age, or Seniority.

Rank	ICM	Age	ICM	Seniority	ICM
Assistant Professor, Teaching Stream	4.0	≤ 30 years old	3.9	0-5 years	4.0
Associate Professor, Teaching Stream	3.9	31-40 years old	3.9	6-10 years	3.9
Lecturer/Senior Lecturer	3.9	≥ 71 years old	3.8		
Assistant Professor	4.0	41-50 years old	3.9	11-20 years	3.9
Associate Professor	3.9	51-60 years old	3.9	21-30 years	3.8
Professor	3.9	61-70 years old	3.8	30+ years	3.8

Note. ICM = Institutional Composite Mean.

8. Interpretability of ICM Scores

A. ICM scores fell along the full continuum of possible scores (1.0 to 5.0).
B. ICM scores were skewed towards the upper end of the scale (*M* = 4.0, *S* = 0.52).
C. ICM scores exhibited discrimination ability across the full range of scale options.
D. ICM scores are especially diagnostic at the upper and lower ends of the scale.
E. Larger course sizes are associated with lower ICM scores, *r* = -0.41.
F. Scores between 3.4 and 4.8 reflect a 'typical' student experience.

A. ICM scores fall along the full continuum of possible scores

The ICM will be most useful if it successfully differentiates students' experiences with different instructors and courses. If all course sections receive the same ICM score, the ICM will have no informational value. Range restriction occurs when an instrument fails to achieve enough variability to yield meaningful interpretation at the desired level of analysis (in this case, the section-level of analysis). Fortunately, the ICM scores fell along the full continuum of possible scores (1.0 to 5.0) with mean ICM score of 4.0 and standard deviation of 0.52. This indicates that, on average, the ICM scores of any given course section deviated from the grand mean of 4.0 by approximately ½ of an ICM point (i.e., 0.5).

B. ICM scores are skewed towards the upper end of the scale

Figure 15 shows the distribution of ICM scores across course sections. ICM scores fell along the full continuum of possible scores (1.0 to 5.0), but the majority of the scores were skewed towards the upper end of the scale (*M* = 4.0, *S* = 0.52, *mdn* = 4.1).

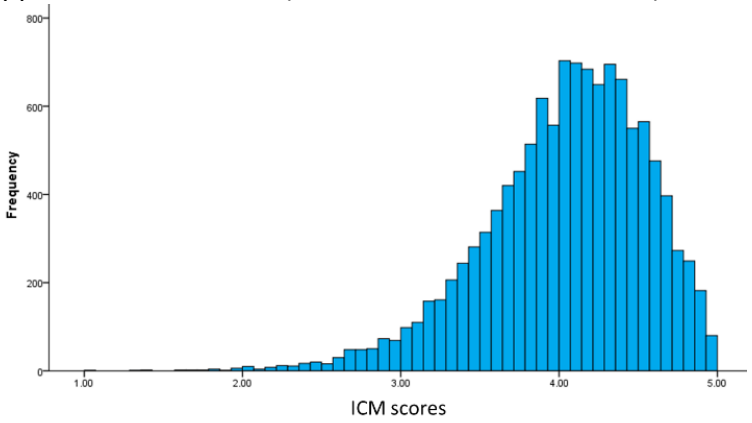


Figure 15. Spread of ICM scores

C. ICM scores exhibited discrimination ability across the full range of scale options

Because the ICM scores were skewed towards the upper end of the scale, it was important to examine the potential impact of a ceiling effect. A ceiling effect occurs when a large proportion of scores "max-out" at the upper end of the scale resulting in a loss of discrimination ability. **Discrimination ability**[4] is the ability to use a scale to differentiate between different entities that are being measured.

Discrimination ability was examined by grouping course sections into deciles (i.e. 10 equal sized groups created by rank ordering scores from low to high). For all six institutional items, and the ICM, there was meaningful differentiation across the decile groups (each decile group could be statistically and meaningfully differentiated from the decile below it, *p*'s < .05, Cohen's *d* effect size > 0.20). Importantly, there was no evidence that a ceiling effect resulted in a loss of discrimination ability at the upper end of the scale. If anything, there was slightly more differentiation at the upper end of the scale relative to the middle of the scale, as evidenced by the inverted "S" shape pattern of the percentile score plot.

[4] **Discrimination** ability is different from **discriminant** validity. Discrimination ability focuses on the ability of an item (or set of items) to differentiate amongst the different entities that are being measured (in our case students' experiences with different course-sections). In contrast, discriminant validity is the extent to which items meant to capture two theoretically distinct constructs can be differentiated from one another.

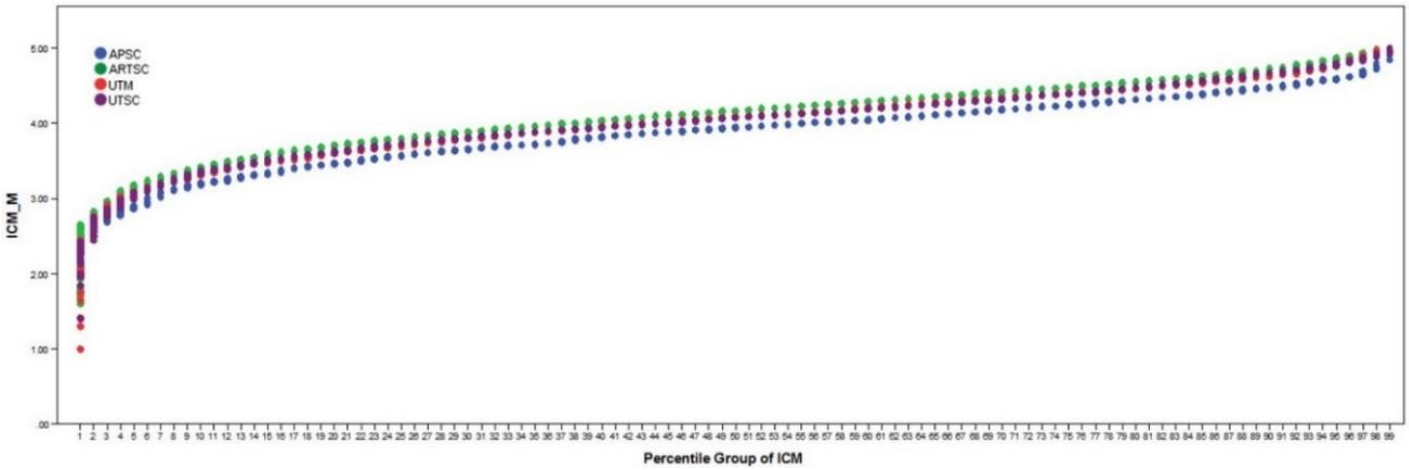


Figure 16. Percentile-score plot showing levels of discrimination ability across the 5-point scale

D. ICM scores are especially diagnostic at the upper and lower ends of the scale

The inverted "S" shaped pattern of the percentile score plot demonstrates that ICM scores have very high levels of discrimination ability at the bottom of the scale (between scores of 1 and 3), moderate levels of discrimination ability in the middle of the scale (between scores 3.0 and 4.5), and somewhat higher levels of discrimination ability at the very upper end of the scale (above 4.5). As such, the ICM may be particularly diagnostic when scores are lower than 3.0 or higher than 4.5.

Scores lower than 3.0, in particular, are "out of the norm" and warrant further investigation. **Importantly, however, scores lower than 3.0 do not, necessarily, indicate problematic teaching, poor student experience, or low learning outcomes**. Low evaluation scores could arise for any number of reasons, including factors that may be completely outside of the control of the instructor. Instructors should always be given an opportunity to investigate and contextualize potential reasons for low course evaluation scores.

E. Larger course sizes are associated with smaller ICM scores

Larger course sizes were associated with lower ICM scores, r = -0.41 (moderate effect size). On average, smaller courses had ICM scores 0.5 points higher than large enrollment courses (200+ students). Course size should be taken into consideration when interpreting ICM scores.

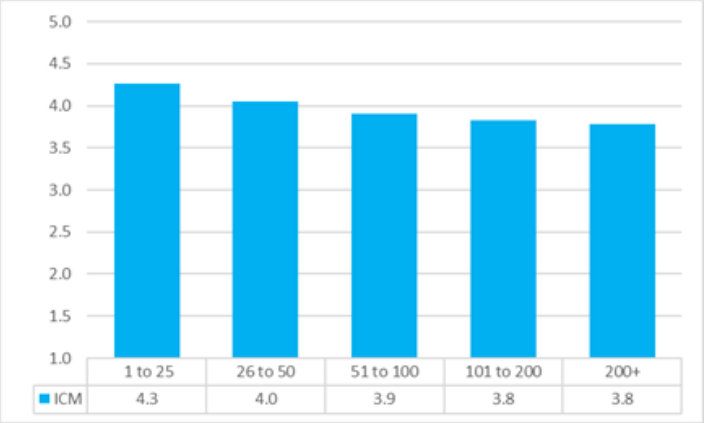


Figure 17. Course size and ICM scores

F. Scores between 3.4 and 4.8 reflect a 'typical' collective student experience

Overall, 70% of ICM scores were between 3.5 and 4.6. Only 15% of scores were lower than 3.5 and only 15% of scores were higher than 4.6. However, this "range of typicality" varied based on course size. When course size was taken into consideration the range of 'typicality' included scores as low as 3.4 and as high as 4.8. The table below describes the range of "typicality" based on each course size category. For any given course size, scores within this range of typicality should be interpreted as reflecting a typical collective student experience.

Table 20

Mean ICM Scores, Standard Deviation, And Range of 'Typicality' By Course Size

Course size	*M*	Typical (middle 70%)	Lower than typical (bottom 15%)	Higher than typical (top 15%)
1-25	4.3	3.7 and 4.8	≤ 3.6	≥ 4.9
26-50	4.0	3.6 and 4.5	≤ 3.5	≥ 4.6
51-100	3.9	3.4 and 4.4	≤ 3.3	≥ 4.5
101-200	3.9	3.4 and 4.3	≤ 3.3	≥ 4.4
201+	3.8	3.4 and 4.2	≤ 3.3	≥ 4.3

Importantly, scores outside of this range of typicality do not, necessarily, indicate poor or exemplary teaching. ICM scores can be influenced by a number of factors, many of these outside of the control of the instructor. With that said, an atypical ICM score may warrant further investigation, especially if the score seems unusually low or high for a particular course or department. Possible sources of evidence for better understanding atypical ICM scores may include (but are not limited to): the instructor's narrative explanation of the course; course context variables; students' written comments; classroom observation; course materials, and/or other supporting documents.

9. Generalizability

A. The ICM exhibits identical reliability and validity patterns across academic divisions.
B. The ICM is generalizable to graduate-level courses.
C. The ICM is generalizable to dual-instructor courses, but the evaluation context differs.

A. The ICM exhibits identical reliability and validity patterns across academic divisions

The current validation study examined single-instructor undergraduate courses across the four largest undergraduate divisions at the University of Toronto. Across the four divisions, the ICM exhibited strong consistency in terms of item completion rates (Table 21), interrater reliability (Table 22), internal consistency (Table 23), test-retest reliability (Table 24), and convergent validity patterns (Table 25).

Table 21

Completion Rates for The 6 Institutional Items by Division

	FASE	ARTSC	UTM	UTSC
UofT1. Intellectually stimulating	99.8%	99.8%	99.9%	99.9%
UofT2. Deeper understanding	99.8%	99.3%	99.8%	99.8%
UofT3. Instructor created atmosphere	99.7%	99.4%	99.8%	99.8%
UofT4. Improve understanding	99.8%	99.5%	99.8%	99.6%
UofT5. Demonstrate understanding	99.8%	99.4%	99.9%	99.9%
UofT6. Overall learning experience	99.7%	91.0%	99.8%	99.8%
Correlation: length and complete rate	*r* = -.03	*r* = .001	*r* = -.004	*r* = -.02

Table 22

Interrater Agreement and Interrater Reliability, by Division

	FASE		ARTSC		UTM		UTSC	
	r_{wg}	ICC(*k*)	r_{wg}	ICC(*k*)	r_{wg}	ICC(*k*)	r_{wg}	ICC(*k*)
Item 1	0.74	0.84	0.72	0.81	0.72	0.82	0.73	0.80
Item 2	0.76	0.85	0.72	0.80	0.73	0.81	0.74	0.79
Item 3	0.72	0.92	0.69	0.87	0.70	0.87	0.70	0.87
Item 4	0.73	0.84	0.72	0.80	0.71	0.80	0.73	0.79
Item 5	0.73	0.83	0.72	0.80	0.71	0.80	0.73	0.78
Item 6	0.73	0.89	0.74	0.84	0.72	0.85	0.73	0.84
ICM	**0.93**	**0.89**	**0.92**	**0.85**	**0.92**	**0.85**	**0.92**	**0.85**

Table 23

Factor Loadings And Internal Consistency, By Division

	FASE	ARTSC	UTM	UTSC
Variance Explained	**83%**	**85%**	**87%**	**87%**
Item 1 factor loading	.87	.89	.92	.91
Item 2 factor loading	.94	.93	.94	.93
Item 3 factor loading	.82	.86	.89	.89
Item 4 factor loading	.92	.93	.93	.92
Item 5 factor loading	.91	.91	.91	.92
Cronbach's alpha (α)	**α = .94**	**α =.95**	**α =.96**	**α =.96**

Table 24

Reliability Across Students, Courses, Instructors, and Course-Instructors by Division

	FASE	ARTSC	UTM	UTSC
Same student (across different courses)	0.64	0.62	0.66	0.63
Same course topic (regardless of instructor)	0.71	0.72	0.73	0.67
Same instructor (regardless of course topic)	0.70	0.73	0.72	0.75
Same course with the same instructor	0.81	0.80	0.83	0.79

Table 25

Patterns of Convergent and Discriminant Validity with The ICM by Division

	FASE	ARTSC	UTM	UTSC
Intellectual engagement	0.77	0.85	0.87	0.86
Higher order learning	0.87	0.83	0.87	0.89
Clarity of instruction	0.88	0.86	0.80	0.82
Learn from components	0.91	0.75	0.82	0.78
Fairness of assessment	0.80	0.70	0.74	0.77
Attendance	0.16	0.47	-0.04	0.28
Workload	0.31	0.03	0.01	0.04
Expected Grade	0.43	0.59	0.40	0.33

B. The ICM is generalizable to graduate-level courses

ICM values are similar across graduate and undergraduate courses of the same size

When comparing ICM scores of graduate course sections with undergraduate course sections, it appears as if the ICM scores are higher for the graduate-level courses (*M* = 4.1) than for undergraduate-level courses (*M* = 3.9). However, those differences go away when taking into consideration course size. Indeed, 85% of all graduate courses have fewer than 25 students in them. In contrast, only 36% of undergraduate courses have fewer than 25 students. Once these differences are taken into consideration, graduate courses (SGS) have ICM scores comparable to similarly size undergraduate courses. Interestingly, however, graduate courses do have consistently higher response rates, no matter the course size category.

Table 26

Response Rates and ICM Scores for Graduate and Undergraduate Courses by Course Size

Course size	% UG	% SGS	RR SGS	RR UG	ICM SGS	ICM UG
1-25	36%	85%	58%	50%	4.3	4.2
26-50	26%	12%	53%	44%	3.9	4.0
51-100	18%	3%	51%	38%	3.9	3.9
101-200	13%	< 1%	42%	34%	3.8	3.8
200+	7%	0%	----	32%	----	3.8

Note. UG = undergraduate; SGS = graduate; % = percent of courses falling into each course size category; RR = response rates; ICM = Institutional Composite Mean.

Graduate courses exhibited similar internal consistency patterns as undergraduate courses

Graduate courses also exhibit similar internal consistency patterns to undergraduate courses. Indeed, the Cronbach's σ and factors analysis patterns are nearly identical to those found with the undergraduate courses. Indeed, in graduate-level courses the five items of the ICM loaded on to a single factor, with the single factor explaining 86% of the variance. All of the factor loadings were greater than 0.70 (the typical cut-off point to assess factor loadings is 0.40 or higher). Furthermore, the internal consistency was very high, with Cronbach's α = 0.96. The ICM appears to be a reliable composite variable, even in dual-instructor courses.

Table 27

Factor Loadings and Internal Consistency in Single- Versus Dual-Instructor Courses

Factor analysis	%	I1	I2	I3	I4	I5	α
Undergraduate courses	**86%**	0.90	0.93	0.87	0.93	0.91	**0.96**
Graduate courses	**82%**	0.91	0.92	0.87	0.92	0.90	**0.94**

Note. % = percent of variance explained, I1 – 16 = institutional items; α = Cronbach's alpha.

C. The ICM is valid for use in dual-instructor courses, but the evaluation context differs

The main validation study focused on single-instructor courses. To examine dual-instructor courses, a follow-up analysis examining 509 dual-instructor undergraduate course sections evaluated within the same divisions and time-period as the main sample was performed.

Course evaluations in a dual-instructor course context

The evaluation context differs slightly between dual-instructor and single-instructor courses. In a dual-instructor course, students rate 7 institutional items, instead of 6 institutional items. This is because institutional item 3 is an instructor-specific question. When students rate a multi-instructor course, they rate item 3 for each instructor under consideration (e.g., in a dual-instructor course item 3 is asked twice, once for each instructor). The other five institutional items focus on the course as a whole and are rated only once per item.

Students differentiate between instructors

In dual-instructor courses, the average difference in ratings between the two instructors on item 3 was 0.56 (S = 0.58), with differences ranging between a low of 0.00 (no difference) and a high of 3.78. In addition, the aggregated ratings of each instructor were only weakly correlated with one another, r = 0.24 (small effect). These results suggest that the ratings of one instructor did not heavily influence ratings of the other instructor. Importantly, there were no differences in ratings based on the order in which each instructor was listed on the course evaluation form, η^2 = .001 (no effect), suggesting that the differences in ratings were not an artifact of the order in which faculty were listed on the survey.

Students differentiate between the instructors and the course as a whole

Item 3 was differentially correlated with the other institutional items when comparing single-instructor course sections with dual-instructor course sections. In single-instructor course sections, item 3 was strongly correlated with the other institutional items. In contrast, in the dual-instructor sections, the item 3 rating of any given instructor was more moderately correlated with the other institutional items.

Table 28

Correlations Between Item 3 and the Other Items in Single- Versus Dual-Instructor Courses

Item 3	I1	I2	I4	I5	I6	ICM
Single-instructor section	0.80	0.81	0.78	0.77	0.89	0.91
Dual-instructor section	0.58	0.59	0.57	0.56	0.66	0.77

Note. I1 – 16 = institutional items; ICM = Institutional Composite Mean

Evaluations of the course are more strongly associated with the higher scoring instructor

An examination of correlations between the highest rated instructor versus the lowest rated instructor suggest that, on the whole, student ratings may be slightly more associated with their perceptions of the more higher scoring instructor, rather than the lower scoring instructor, although the differences in strength of correlations are small.

Table 29

ICM Scores Are More Strongly Associated with the Favoured Instructor

Item 3	I1	I2	I4	I5	I6	ICM
Highest rated instructor	0.67	0.69	0.66	0.68	0.76	0.81
Lowest rated instructor	0.62	0.61	0.59	0.59	0.72	0.74
Average of ratings	0.71	0.72	0.69	0.68	0.82	0.84

Note. I1 – 16 = institutional items; ICM = Institutional Composite Mean

Dual-instructor courses receive slightly lower ICM scores than single-instructor courses

When comparing dual-instructor course sections with single-instructor course sections, the dual-instructor course sections were rated slightly lower than the single-instructor course sections, however these differences were relatively small (d = 0.28, small effect).

Table 30

Institutional Item Means for Single- Versus Dual-Instructor Courses

Item 3	I1	I2	I3	I4	I5	I6	ICM
Single-instructor section	4.0	4.1	4.1	4.0	4.0	3.8	4.0
Dual-instructor section	3.9	4.1	4.0	3.8	3.8	3.6	3.9

Note. I1 – 16 = institutional items; ICM = Institutional Composite Mean

The ICM had similar factor analysis patterns in dual-instructor versus single-instructor courses

Given the differential correlation pattern between the instructor item (item 3) and the other institutional items in dual-instructor courses, this raises questions about the generalizability of the internal consistency of the institutional composite mean (ICM) when evaluating dual-instructor courses versus single-instructor courses. Fortunately, the items of the ICM seem to exhibit similarly high internal consistently in dual-instructor courses versus single-instructor courses.

Indeed, in dual-instructor courses the five items of the ICM loaded on to a single factor, with the single factor explaining 86% of the variance. All of the factor loadings were greater than 0.70 (the typical cut-off point to assess factor loadings is 0.40 or higher). Furthermore, the internal consistency was very high, with Cronbach's α = 0.96. The ICM appears to be a reliable composite variable, even in dual-instructor courses.

Table 31

Factor Loadings and Internal Consistency In Single- Versus Dual-Instructor Courses

Item 3	%	I1	I2	I3	I4	I5	α
Single-instructor	**86%**	0.90	0.93	0.87	0.93	0.91	**0.96**
Dual-instructor	**76%**	0.91	0.92	0.70	0.92	0.90	**0.91**

Note. % = percent of variance explained, I1 – 16 = institutional items; α = Cronbach's alpha.

Single-instructor and dual-instructor courses are not equivalent

Single-instructor and dual-instructor course sections do not result in equivalent evaluation contexts. In single-instructor course sections, students appear to be more likely to conflate their perceptions of the instructor with the course. This conflation does not occur to as strong of a degree in dual-instructor course sections. In dual-instructor course sections, students seem to make greater differentiation between their perceptions of any given instructor and their perceptions of the course as a whole. Because of these differences, direct comparisons should not be made between an instructor teaching a single-instructor course with an instructor teaching a dual-instructor course, no matter the similarity of the topic or the course structure. However, the current results also suggest that the key psychometric properties of the institutional items remain stable even in the context of dual-instructor course sections. These results suggest that it may be appropriate to use the Cascaded Course Evaluation Framework in dual-instructor course sections similar to those evaluated in the validation study, as long as the results are not treated as equivalent to those found in single-instructor course sections. At present, there is not enough data to examine the generalizability of the course evaluation framework in contexts involving more than two instructors.

IMPLICATIONS FOR INTERPRETATION

Adequate response rates

The table below outlines the response rates required to achieve "very" precise to "somewhat precise" ICM estimates for courses of varying sizes. ICM scores will be most meaningful when response rates are 50% or higher for small courses (< 50 students) and 20% or higher for larger courses (> 100 students). Certainly, the ICM can still be used for formative and summative purposes when response rates are lower than this, but in these cases the ICM score should be thought of as a general estimate of students' collective experiences, rather than as precise estimate of these experiences. If the goal of assessment is to make very precise estimates of students' collective experiences for the purpose of making fine-tuned comparisons across time points, course sections, course topics, or instructors, higher response rates are best.

Table 32

Response Rate Needed to Make Meaningful Inference

		Course Size				
Interval around the mean	**Recommended interpretation of the quality of the mean estimate**	1-25	26-50	51-100	101-200	200+
< ±0.1	Very precise estimate	>90%	>80%	>80%	>60%	>50%
< ±0.2	Precise estimate	>80%	>70%	>70%	>50%	>40%
< ±0.5	Somewhat precise estimate	>70%	>50%	>40%	>20%	>10%
< ±1.0	General estimate	>60%	>20%	>10%	>10%	>10%
> 1.0+	Very general estimate	< 30%	<10%	<5%	<3%	<1%

Note. Guidelines are based on a 95% confidence interval around the mean with margin of errors ranging from ±0.1 to ±1.0, a standard deviation of 1.0, and correction for the use of a finite population.

In addition to the actual response rate, it is also important to consider the nature of the respondents themselves. Even within the same course section, students may have drastically different learning experiences from one another. Our data suggest that, for the most part, students within the same course section tend to be in relatively high agreement in their institutional item ratings of the same instructor/course section. However, there is always the possibility that a student with an atypical experience can sway the overall results, particularly in smaller classes. **In general, encouraging high response rates is one of the best ways to ensure that the ICM score is a meaningful reflection of students' collective experiences within a course**.

ICM Interpretation

The ICM has been found to be a reliable and valid indicator of students' collective experiences with the University of Toronto's institutional teaching and learning priorities. As such, the ICM provides valuable information on the extent to which a particular course instructor/course section created:

- An engaging atmosphere for students.

- Opportunities for students to gain knowledge.
- An atmosphere conducive to learning.
- Opportunities for students to learn from assessment.
- Opportunities for students to demonstrate their understanding.

The ICM score is **not** intended to be a direct measure of student learning. Nor is it a measure of the appropriateness of the scope and depth of the content covered in the course. Rather the ICM is **one** of **many** pieces of evidence that can be used to better understand teaching and learning environments at the University of Toronto. The ICM, as a measure of students' collective experiences with the teaching and learning priorities at the University of Toronto should always be interpreted within the larger teaching and learning context.

Typical versus atypical ICM scores

The table below describes a "range of typicality" (i.e., the middle 70%) for any given course size. Scores within this range reflect a 'typical' collective student experience. Scores outside of this range are 'atypical' in that they reflect the bottom 15% of ICM scores and the top 15% of ICM scores. Importantly, however, atypically low scores do not, necessarily, indicate poor teaching, nor do atypically high scores, necessarily, indicate exemplary teaching. ICM scores can be influenced by a number of factors, some of which are outside of the control of the instructor. With that said, an atypical ICM score may warrant further investigation.

Table 33

Range of Typical ICM Scores for Each Course Size Category

Course size	*M*	Typical (middle 70%)	Lower than typical (bottom 15%)	Higher than typical (top 15%)
1-25	4.3	3.7 and 4.8	≤ 3.6	≥ 4.9
26-50	4.0	3.6 and 4.5	≤ 3.5	≥ 4.6
51-100	3.9	3.4 and 4.4	≤ 3.3	≥ 4.5
101-200	3.9	3.4 and 4.3	≤ 3.3	≥ 4.4
201+	3.8	3.4 and 4.2	≤ 3.3	≥ 4.3

ICM scores in a larger context

Course evaluation scores should always be interpreted within a larger teaching and learning context. Possible sources of evidence that can be used to contextualize ICM scores include (but are not limited to): an instructor's narrative explanation of their teaching contexts; course context variables; students' written comments; classroom observation; course materials, and/or other supporting documents. The University of Toronto provides a table outlining possible sources of evidence for contextualizing teaching competence (see: https://teaching.utoronto.ca/teaching-support/documenting-teaching/teaching-dossier/).

When interpreting ICM scores the results of the validation study suggest that the following contextual factors may be of particularly high importance for interpreting ICM scores:

Specific division/department

Although differences were small, and mostly explained by differing course sizes, ICM scores varied from division to division and from department to department. As such, ICM scores should be interpreted within the context of specific divisions and departments, rather than being compared directly across units.

The size of the course

Although the correlation between ICM scores and course size was "moderate", the average difference between a very small course (1-25 students) and a very large course (200+ students) can be as high as 0.5 points on a 5-point scale. Course size should always be taken into consideration when interpreting course evaluation scores.

Single instructor versus dual/multi-instructor courses

Similar ICM values emerged between single-instructor and dual-instructor courses, and the items were psychometrically similar (especially when it came to the factor structure). However, the analyses also suggested that students use different criteria to rate single-instructor versus dual-instructor courses, especially when it comes to the core institutional item 3 which focuses on specific course instructors.

In single-instructor course sections, students appear to be more likely to conflate their perceptions of the instructor with that of the course. This conflation does not occur as strongly in dual-instructor course sections. In dual-instructor course sections, students seem to make greater differentiation between their perceptions of any given instructor and their perceptions of the course as a whole. Because of these differences, direct comparisons should not be made between an instructor teaching a single-instructor course with an instructor teaching a dual-instructor course, no matter the similarity of the topic or the course structure.

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, and draws on input from diverse institutional stakeholders (e.g. the Course Evaluation Advisory Group) and experts who provided guidance around the key questions to ask to examine the framework's effectiveness. The study reflects the University's commitment to ongoing analyses and education related to course evaluation data quality and interpretation.

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.

Centre for Teaching Support & Innovation. (2017*). Developing & Assessing Teaching Dossiers: A guide for University of Toronto faculty, administrators and graduate students*. Toronto: Centre for Teaching Support & Innovation, University of Toronto.

Cook, C., Heath, F., & Thompson, R. L. (2000). A meta-analysis of response rates in web-or internet-based surveys. *Educational and Psychological Measurement*, *60*(6), 821-836.

Eagan Jr, M. K., & Garvey, J. C. (2015). Stressing out: Connecting race, gender, and stress with faculty productivity. *The Journal of Higher Education*, *86*(6), 923-954.

Goos, M., & Salomons, A. (2017). Measuring teaching quality in higher education: assessing selection bias in course evaluations. *Research in Higher Education*, *58*(4), 341-364. DOI 10.1007/s11162-016-9429-8

Le Breton, J. M., & Sentor, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815–852.

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319-383). Springer, Dordrecht.

Marsh, H. W., & Bailey, M. (1993). Multidimensional students' evaluations of teaching effectiveness: A profile analysis. *The Journal of Higher Education*, *64*(1), 1-18.

Morley, D. D. (2009). SPSS macros for assessing the reliability and agreement of student evaluations of teaching. *Assessment & Evaluation in Higher Education, 34*(6), 659-671, DOI: 10.1080/02602930802474151

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: what can be done? *Assessment & Evaluation in Higher Education*, *33*(3), 301-314.

Shih, T. H., & Fan, X. (2009). Comparing response rates in e-mail and paper surveys: A meta-analysis. *Educational Research Review*, *4*(1), 26-40.

Shih, T. H., & Fan, X. (2008). Comparing response rates from web and mail surveys: A meta-analysis. *Field Methods*, *20*(3), 249-271.

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.

Streiner, D. L., & Norman, G. R. (2003). *Health Measurement Scales: A Practical Guide to their Development and Use*. Oxford Medical Publications.

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