University of Toronto Course Evaluation Interpretation Guidelines for Academic Administrators

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 and Innovation (2018). *University of Toronto Course Evaluation Interpretation Guidelines for Academic Administrators.* 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:

Gregory Hum, Assistant Director, Teaching Assessment, CTSI, University of Toronto
Brandon Tam, Research Assistant, CTSI, University of Toronto
Kosha Bramesfeld, Data Analyst, Course Evaluations, CTSI, University of Toronto

Acknowledgements:

Diane Horton, Acting Director (January-June, 2018), CTSI, University of Toronto
Megan Burnett, Associate Director, CTSI, University of Toronto

Introduction

This document highlights key elements for effective practice in interpreting course evaluation reports for the purposes of assessing teaching. These recommendations are informed by statistical analyses conducted by the Course Evaluations team at the Centre for Teaching Support & Innovation (2018) and the wider literature on this subject (see references at the end of this document).

This guide has been prepared in reference to course evaluations conducted through the Centralized Course Evaluation Framework which is being progressively implemented across the University of Toronto. As well, for further guidance you may wish to consult:

- The primary course evaluation PDF reports generated since Fall 2017, which provide further information on terminology used in the reports and here.
- Relevant University of Toronto policies and guidelines (see references at the end of this guide).
- Relevant Divisional Teaching Evaluation Guidelines (see references at the end of this guide).
- Books and articles referenced in this document (see references at the end of this guide).
- The Centre for Teaching Support & Innovation (see 'contact information' below).

Three Core Principles of Effective Interpretation:

1. Course evaluations are a key component of the evaluation of teaching, but they only provide a partial portrait. The assessment of teaching should make use of multiple sources of evidence (triangulation of data).
2. Course evaluation scores, particularly the University of Toronto Institutional Composite Mean, are reasonably valid and reliable indicators of student experiences in a course. However, due to the margin of error inherent in all human-measurement, as well as contextual factors that come to bear, the scores are not sufficiently precise to support fine-grained comparisons and rankings on their own.
3. Conclusions from course evaluation data for the assessment of teaching should be drawn only from clear trends and patterns (e.g., not from isolated comments or scores), after considering all available data, and in consideration of context(s) (e.g., course type, size).

What do course evaluation scores tell us?

- Course evaluation scores represent students' perceptions of their learning environment and are an important mechanism for hearing student voices. As participants in a course throughout a term, theirs is an important perspective to consider. Research supports the notion that students can accurately report some (but not all) elements of teaching, such as: teaching actions and strategies employed, the ease/difficulty of their learning experience, and workload. Course evaluations at the University of Toronto strive to minimize questioning that is known to produce inaccurate or unreliable results (e.g., questions on instructors' knowledge or students' self-reported learning) (Theall & Franklin, 2001).

Examining course evaluation scores:

- The most reliable and valid score in the University of Toronto Course Evaluation Framework is the Institutional Composite Mean (ICM). It is a composite score formed by calculating the mean (i.e., average) of the scores on the five questionnaire items that form the scale, and as such, it incorporates the most information and is more reliable than any one of the individual items and their scores.

  - In general, it is best to consider the ICM rather than the any one of the individual items that comprise it, in isolation.
  - The overall mean of the ICM across all course evaluations at the survey-level[1] (i.e., not corrected for course size) for all divisions (undergraduate and graduate) is **3.9** and the standard deviation is **0.95**.
    - Please note Tables 2 and 3 described below regarding the importance of class size relative to the ICM

- Best practice includes considering the distributions of scores (Linse, 2017). These are provided in section 2 of the PDF course evaluation reports.
  - For all the institutional questions, and the vast majority of other course evaluation questions used (e.g., divisional), there is strong agreement between students on their responses.
    - Most commonly, the distribution is "skewed," with the majority of responses ranging from 3-5 on the five-point scale. This is a common pattern that has been observed elsewhere (e.g., Nulty, 2008; Theall & Franklin, 1991).
    - This majority opinion is what should be mainly considered. It is typical, even for exceptional instructors, to receive a small number of low scores.
  - Deviations from the typical distribution(s) described above deserve to be examined and considered more closely:
    - For example, distributions with highly divided scores, having multiple peaks, typically two (i.e., bi-modal), may suggest important contextual variables to consider (e.g., students with different levels of background knowledge or interest).
    - Students' qualitative comments are often especially helpful for understanding these atypical response patterns.

- Comparators provided on course evaluation reports (section 3) are not intended to be a definitive benchmark. They should be considered cautiously as they are calculated at the survey-level (see footnote 1) and as such, do not account for context (including course size).

Interpreting comments:

- In interpreting the student comments, one should be looking primarily for comments that are common and where students show significant consensus, since:
  - Not all comments are true, or meaningful.
  - Positive comments tend to be shorter in length, and less "salient" than negative comments. This can cause negative comments to appear more common or meaningful than they are. Avoid giving these undue focus.

[1] We refer to one student's response to the course evaluation questions for a particular course-section and term as a "survey". The survey-level mean calculations are made by averaging across all individual surveys (including multiple surveys done by a given student for each course they are enrolled in). Calculations done this way incorporate the most possible information, but *do not* correct or account for a number of factors including context or course size.

Context(s) to Consider:

The following recommendations were informed by analyses conducted by the Centre for Teaching Support & Innovation using data from the University of Toronto[2].

Multi-instructor courses:

- The vast majority of questions do not differentiate between instructors. Thus, in a multi-instructor course, responses are not necessarily indicative of any specific instructor's role in the course, but rather, the aggregate experience(s) of that course.
  - It may be informative to examine an instructor's other scores, for similar courses, to determine whether the scores in a multi-instructor course are representative.

Response rates:

- Response rates affect the relative precision of the observed/collected scores in their ability to estimate the score that would have been attained had 100% of students responded. Simply put, the higher the response rate, the more precise the estimate is.

- The response rate relative to the course size affects the quality, and therefore, interpretability of the course evaluation results.

- Based on our analyses[3] which used local data to estimate the size of the margin of error interval for a given margin of error and course size, we have provided interpretive labels of the relative quality of the estimate (Table 1).
  - The relative quality of the estimate represents a range. It is strongly advised that 'general' and especially 'very general' estimates be used only for formative purposes as these estimates are less likely to be accurate. Summative evaluations should only consider relatively precise estimates and more heavily consider more precise estimates.
  - For example:
    - For a course with 28 students, a response rate of 21% provides an estimate between 'somewhat precise' and a 'general' estimate.
    - For a course with 211 students, a response rate of 21% provides an estimate between 'precise' and 'somewhat precise'.
    - For a course with 211 students, a response rate of 51% provides a 'very precise' estimate.
  - A more detailed explanation of the development of Table 1 is provided in the appendix under "detailed explanation of Table 1".

[2] Conducted with undergraduate Fall/Winter course evaluations from 2015-2017 in: Faculty of Applied Science and Engineering; Faculty of Arts and Science; University of Toronto Mississauga; University of Toronto Scarborough. For these analyses, data were aggregated to course-section levels prior to analyses so that additional factors (e.g., course size) would have their influence removed or reduced.

[3] For greater detail into the underlying statistics and analyses used to make these recommendations see "response rate and interpretation" in the appendix of this document.

Table 1. Interpretation of precision based upon margin of error interval sizes for response rates at given course size ranges

		Course Size				
Margin of error interval	**Interpretation**	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.

- Note on overall institutional response rate quality:
    - We have found that 96% of the courses at U of T allow for at least a "general estimate" of students' collective experience, and 66% allow for a "very precise" to "somewhat precise" estimate.

Course Size:

- Course size has been identified as a main contextual factor associated with differences in both response rates and ICM. Larger courses are associated with lower response rates and lower ICM scores than smaller courses.

Table 2. Mean response rate and ICM for given course size ranges

	Course Size				
	1-25	**26-50**	**51-100**	**101-200**	**201+**
Response Rate	50%	44%	38%	34%	32%
Mean ICM	4.3	4.0	3.9	3.9	3.8

Overview of effective practices in the assessment of teaching:

Our key recommendations when making judgements across large groups of instructors:

- **Avoid sorting only on ICM scores across large groups**: Scores are rarely sufficiently precise to make meaningful distinctions between instructors, particularly on decimal point differences. At the decimal level, differences observed are more likely to be due to contextual factors or chance than actual differences in teaching (Boysen et. al. 2014; Linse, 2017)

- **Scores should, at most, be used to preliminarily sort into broad, predetermined ranges**: If using the ICM scores to make comparisons, it is recommended that they be used to sort into a number of predetermined score ranges (e.g., by percentiles).
    - This/these criterion scores may be informed by the qualitative response labels/anchors used in the course evaluation questions (e.g., 3=moderately…) and context (e.g., course size).
    - A purely "norm-referenced" approach (i.e., simply sorting all scores from highest to lowest) risks arbitrarily categorising effective instructors as low in performance since by mathematical definition, half will be "below average".

It is best to perform such preliminary sorting within groups of similar courses. As course size is the currently identified main factor in score differences, to help inform such an approach, Table 3 provides institutional trends on the distribution of scores within given percentiles.

- **Contextualize and triangulate**: Judgements, and particularly fine distinctions for comparative purposes, should not be based solely on course evaluation scores. Finer comparisons should consider the context (see previously for some important contextual factors) and make use of available relevant information from multiple sources (e.g., teaching statements, teaching awards, course syllabi).
    - It should be noted that, in and of themselves, atypically low scores do not, necessarily, indicate poor teaching; nor do atypically high scores, necessarily, indicate exemplary teaching. They are, however, highly helpful in identifying cases where further examination is warranted.

Table 3. ICM Means and distribution by percentiles for given course size ranges

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

Contact information:

- For additional CTSI consultation regarding course evaluations, please contact CTSI at ctsi.teaching@utoronto.ca
- For questions or comments on this document please e-mail Gregory Hum, Assistant Director, Teaching Assessment/CTSI: gregory.hum@utoronto.ca

Appendix:

References/Additional Resources:

Boysen, G. A., Kelly, T. J., Raesly, H. N., & Casner, R. W. (2014). The (mis) interpretation of teaching evaluations by college faculty and administrators. *Assessment & Evaluation in Higher Education, 39*(6), 641-656.

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. Retrieved from: https://teaching.utoronto.ca/wp-content/uploads/2018/09/Validation-Study_CTSI-September-2018.pdf

Hativa, N. (2014). *Student ratings of instruction: recognizing effective teaching.* Oron Publications: USA.

Linse, A. R. (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in Educational Evaluation, 54*, 94-106.

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.

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction?. *New directions for Institutional Research, 2001*(109), 45-56.

Theall, M., & Franklin, J. (1991). Using student ratings for teaching improvement. *New Directions for Teaching and Learning*, *1991*(48), 83-96.

Relevant Policies:

University of Toronto Policy on the Student Evaluation of Teaching in Courses (2011):

http://www.governingcouncil.utoronto.ca/Assets/Governing+Council+Digital+Assets/Policies/PDF/studenteval.pdf

University of Toronto Provostial Guidelines on the Student Evaluation of Teaching in Courses (2017):

http://www.provost.utoronto.ca/Assets/Provost+Digital+Assets/Provostial+Guideline+on+the+Student+Evaluation+of+Teaching+in+Courses.pdf

Divisional Teaching Evaluation Guidelines:

https://www.aapm.utoronto.ca/academic-administrative-procedures-manual/teaching-guidelines/

Detailed explanation of Table 1:

Using University of Toronto data, we calculated, for a range of course sizes, the response rate required for particular levels of margin of error, and have set thresholds of interpretability based on this (Table 4). For example, in a course with 75 students, where the response rate was greater than 80%, and the mean ICM value was 4.0, we are confident that this observed score is accurate to within 0.1 of the "true score", and thus we describe the score as a "very precise estimate". So one could be very confident that the students "mostly agreed" that the course reflected the key institutional teaching and learning priorities of U of T. On the other hand, if the response rate was lower than 10%, one would have much less confidence in the precision of this score. In this case, we are confident only that the observed score is accurate to within 1.0 of the "true score", so an observed mean ICM value of 4.0 is estimating that the actual mean response could lie anywhere between (3) "moderate" and a (5) "great deal" on the scale. In this case, we describe the score as a "very general estimate."