UNIVERSITY OF TORONTO | Centre for Teaching Support & Innovation

A Step-by-Step Guide to Reviewing Course Evaluations

Course evaluations, along with other feedback sources, offer valuable insights into students' experiences in and perceptions of your courses. This guide is designed to help you effectively analyze course evaluation reports, focusing on both quantitative and qualitative feedback to identify strengths and areas for improvement. A **Process Map** is provided to guide you through five suggested steps sequentially to navigate course evaluation results (Figure 1). Within this guide, and where relevant, information that is specific to a particular audience is denoted with separate "**For Instructors**" and "**For Administrators**" headings.

Course Evaluations in the Broader Evaluation Framework

Course evaluations are a crucial component of a comprehensive teaching effectiveness evaluation framework, which can also include peer reviews, self-assessment and other assessment methods. While course evaluations provide insights into students' reported learning experiences, they do not offer a complete picture of an instructor's teaching effectiveness or overall teaching quality. Course evaluations are nonetheless one useful source of information for formative and summative assessment of teaching when interpreted effectively.

Interpreting Course Evaluation Results

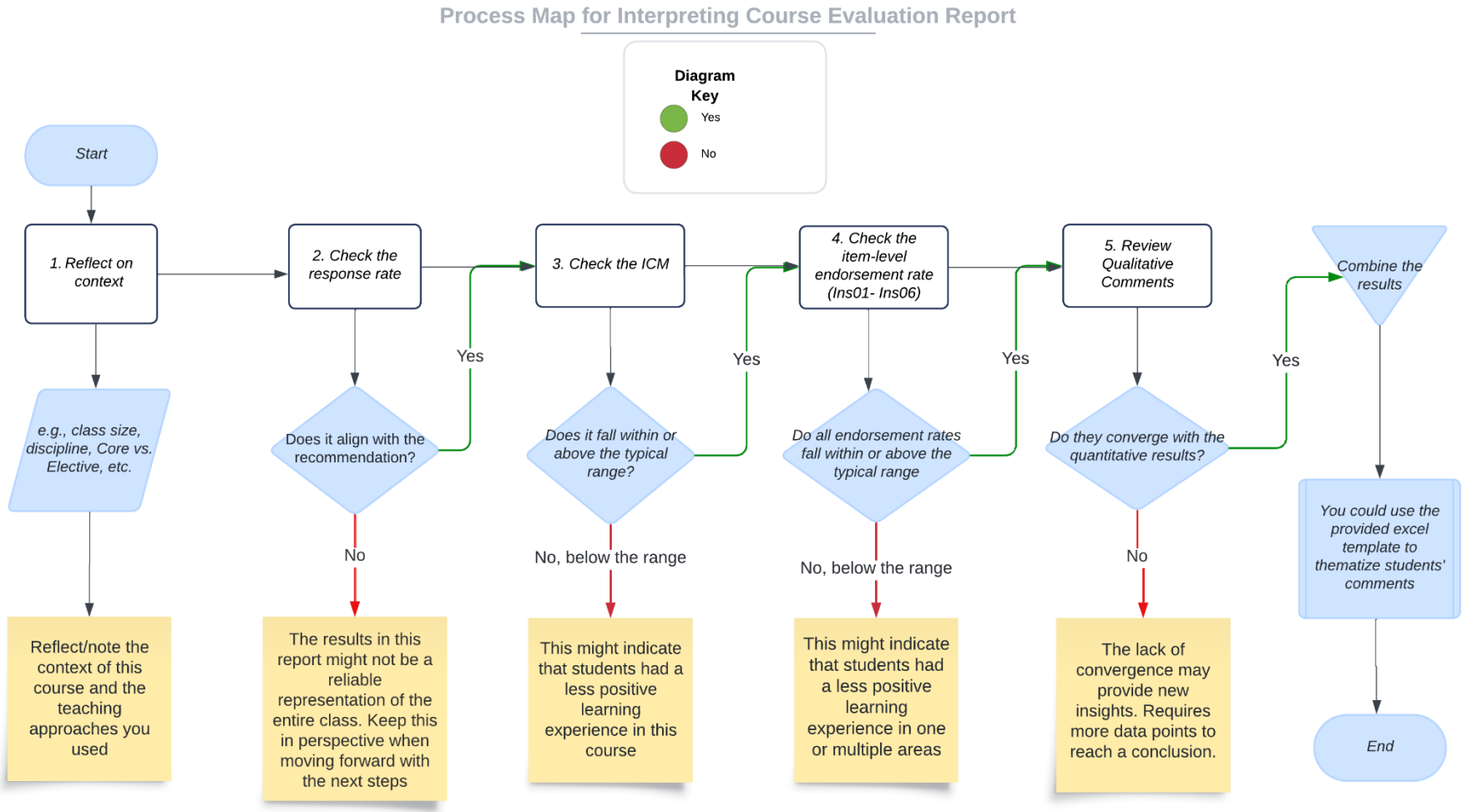
Interpretation of course evaluation results requires an understanding of the response rate, the Institutional Composite Mean (ICM), and item-level endorsement rates. Findings from the Renewed Validation Study (CTSI, forthcoming) highlight the importance of considering recommended response rates and the minimum number of responses necessary for reliable evaluations. The ICM is the average score of five institutional items and offers a broad view of students' overall learning experiences aligned with institutional priorities. It is also important to consider item-level endorsement rates, which are the percentage of respondents that selected the two most positive response options to a question (e.g., "A Great Deal" and "Mostly" combined in Ins01 to Ins05). These endorsement rates can provide more detailed insights that might be obscured by aggregate measures such as the ICM.

Important Considerations for Both Instructors and Administrators

- **Context Matters:** Class size, class level, and other contextual factors can lead to variations in ratings. These factors should be considered when interpreting results.

- **Focus on Patterns:** Look for consistent trends across multiple courses and semesters to make informed decisions about teaching effectiveness. Course evaluation scores are often less precise than expected, limiting the value of fine-grained comparisons or rankings.
- **Multiple Sources of Evidence:** Course evaluations capture student perspectives on their learning experiences. However, teaching evaluation experts advise that no single method can fully assess an instructor's effectiveness. A multifaceted approach promotes a comprehensive understanding of teaching effectiveness by incorporating multiple perspectives—those of students, peers, and instructors—through various types of evidence and teaching artifacts. This approach allows for the best practices in interpreting course evaluation data, ensuring a more accurate and fair assessment within the university context.

Figure 1. Process Map for Interpreting Course Evaluation Report



Step 1: Reflect on Context

For Instructors

Before diving into the report, take a reflective practitioner approach by considering the specific context of the course and your teaching methods:

- What were your course learning objectives?
- Were there unique aspects of this course?
- What went well, and what could be improved?
- Did you experiment with any new teaching strategies?

For Administrators

Prior to interpreting course evaluation results, consider the course context. Contextual factors such as course size can impact course evaluation results (CTSI, forthcoming).

Course Size

Course size has been identified as one contextual factor to consider, with larger course sizes demonstrating lower scores. An Institutional Composite Mean (ICM) of 3.7 might be typical for large classes (200+ students) but would be considered below the typical range for smaller classes (25 or fewer students). While undergraduate-level courses and multi-instructor courses also tend to be rated lower, these differences are negligible after accounting for course size.

Single-instructor vs. multi-instructor

The results of multi-instructor courses require particularly careful interpretation. Many questions—including four out of five questions that make up the ICM—do not differentiate between instructors. Thus, course evaluation responses predominantly capture students' aggregate experience of that course, rather than a single instructor's role.

Comparing Courses to Department Averages (Section 3: Comparative Data)

Comparators provided on course evaluation reports (Section 3: Comparative Data) are not intended to be a definitive benchmark, as they do not account for context (e.g., course size). Critical consideration into how a given course's context differs from the 'average' course context is required. If most courses in a department have small course sizes, and an instructor happens to teach several very large courses, it is likely their course results for those courses will reveal a pattern of performance below the department mean. In this scenario, an apparent trend of underperformance is at least partly due to factors outside of the instructor's control (i.e., course size). Consideration of context and utilization of multiple sources of information are important to avoid such misinterpretation of course evaluation results.

Step 2: Check the Response Rate

Included in the Report Intro (Page 2)

The absolute number of responses and the response rate are **both** important.

The reliability and accuracy of course evaluation results depend on how many students responded. While course evaluation results provide a snapshot of student perceptions, the precision of this snapshot can vary significantly based on the response rate. In some cases, the results are highly reliable, providing a good reflection of student learning experience. However, when fewer students respond, the results may only serve as rough estimates.

The Renewed Validation Study (CTSI, forthcoming) provides guidelines on the minimum response rates and number of responses required for reliable and generalizable results, accounting for both sampling and measurement error, which vary by course size. The study found that most courses at our institution have smaller enrolments: 44% of classes had between 1 and 25 students (referred to as "Very Small"), 25% had between 26 and 50 (referred to as "Small"), 16% had between 51 and 100 (referred to as "Medium"), 10% had between 101 and 200 (referred to as "Large"), and only 5% of classes had more than 200 students (referred to as "Very Large").

However, a significant portion of small classes do not meet the recommended response rate thresholds as displayed in Table 1:

- Only **4%** of Very Small, **22%** of Small, and **21%** of Medium-Size courses meet the threshold.
- In contrast, larger courses (101+ students) have much higher percentages of meeting these guidelines, **73%** of Large and **99%** of Very Large-Size courses.

If a course's response rate falls below the recommended threshold, interpret the results as an imprecise, general estimate rather than a definitive result. In such circumstances, seemingly large differences in the ICM (e.g., 3.7 vs 4.1) may be due to 'chance' or 'noise' rather than reflecting genuine changes in students' reported learning experiences.

For Instructors

Given the trend of smaller classes not meeting suggested response rates, we recommend instructors teaching smaller classes actively monitor their response rates during the evaluation period. For strategies to encourage student engagement, consult CTSI's resources page.

For Administrators

It is strongly advised that only courses meeting the recommended threshold be used for summative evaluations. This ensures a suitable degree of precision in results. It is advisable to treat the course evaluation results of courses not meeting the threshold as broad indicators rather than precise measurements.

Response rates should not be used as a metric to evaluate teaching. Many factors outside of an instructor's control can influence response rates.

Table 1. Response Thresholds for Reliable Results

Course Size	Response Thresholds
Very Small (1-25 students)	Receive a response rate of 70% or higher **and** at least 14 completed surveys.
Small (26-50 students)	Receive a response rate of 50% or higher **and** at least 14 completed surveys.
Medium (51-100 students)	Receive a response rate of 40% or higher **and** at least 15 completed surveys.
Large (101-200 students)	Receive a response rate of 20% or higher **and** at least 15 completed surveys.
Very Large (201+ students)	Receive a response rate of 10% or higher **and** at least 18 completed surveys.

Note: Recommended response rate (%) was operationalized as an interval around the mean of <= .05 (i.e., margin of error of .025) based on 95% confidence interval around the mean, a *t* distribution, a standard deviation of 1, and a finite population correction (CTSI, forthcoming; James et al., 2015, p. 1129). Recommended number of respondents was operationalized as a predicted phi (ϕ) reliability coefficient of >= .80 for a given course ICM, based on empirical results from a generalizability theory d-study (CTSI, forthcoming).

Step 3: Compare Institutional Composite Mean (ICM) to the Typical Range

Included in Report Section 3: Comparative Data

Institutional Composite Mean (ICM): A mathematical average of the first five institutional rating scale items (Ins01-05), which represent institution-wide teaching and learning priorities. The 2018 Validation Study established the reliability and validity of using the ICM as a metric to understand students' collective experiences.

The Institutional Composite Mean (ICM) provides a general sense of how students perceived their learning experience in the course. The ICM is not intended to measure student learning or teaching effectiveness. Table 2 demonstrates typical ranges of the ICM at the University of Toronto based on the middle 70% of the distribution. However, atypically low (bottom 15%) or high (top 15%) scores don't automatically indicate ineffective or effective teaching. Various factors outside an instructor's control—like class size or course difficulty—can influence these scores.

An unusually low ICM score is an indication to explore what might have affected students' perceived learning experience, while an unusually high score may not always reflect a positive learning experience but could result from other factors. Triangulation with other sources of evidence (e.g., instructor's narrative explanation of the course, course contextual variables, student open-ended comments, course materials) may be helpful in investigating atypical responses. When comparing ICM scores across courses or years, avoid drawing strong conclusions from small differences. Such variations often result from natural fluctuations rather than real changes. The range of ICMs (Table 2) should be used as a contextual tool to guide further investigation and improvement efforts, rather than as definitive judgments of teaching quality.

Table 2. Typical ICM Values by Course Size.

Course Size	ICM Mean	Typical ICM (the middle 70%)
Very Small (1-25 students)	4.3	3.8 - 4.9
Small (26-50 students)	4.1	3.6 - 4.6
Medium (51-100 students)	4.0	3.5 - 4.5
Large (101-200 students)	3.9	3.4 - 4.4
Very Large (201+ students)	3.8	3.3 - 4.3

Note: ICM refers to the institutional composite mean, an average of the first five institutional items for a course-section. The mean and typical values in this table correspond to the average of ICMs over all course-sections taught in 2018/19 – 2022/23 (undergraduate & graduate) from all 22 divisions within the central course evaluation framework. Typical values correspond to the 15th and 85th percentiles of the distribution of course-section ICMs.

Step 4: Explore Item-level Metrics for Deeper Insights

Report Section 1: Quantitative Data & Report Section 3: Comparative Data

Item Endorsement Rate: The percentage of respondents that selected the two most positive response options ("A Great Deal" and "Mostly" combined in Ins01 to Ins05; "Excellent" and "Very good" combined in Ins06).

Item-level endorsement rates in Report Section 1 offer a more detailed view of student feedback. The typical ranges for these rates (Table 3) offer context and help guide interpretation. Item endorsement rates of very small courses (1-25 students) tend to be positive.

In addition to interpreting item-level endorsement, considering the distribution of item responses can be helpful. In such cases, the "majority" opinion should be the primary consideration. It is typical, even for exceptional instructors, to receive a small number of low scores. Distributions with multiple peaks may suggest that contextual variables should be considered (e.g., students with different levels of background knowledge providing different scores).

For Instructors

By examining item-level endorsement rates, you can gain insights into strengths and areas for improvement that might be obscured by overall scores. For instance, high endorsement rates on "Ins02: The course provided me with a deeper understanding of the subject matter" may highlight some effective pedagogical practices. Similarly, lower endorsement rates on "Ins04: Course projects, assignments, tests and/or exams improved my understanding of the course material" may indicate areas needing attention. Note that some items tend to have higher endorsement rates on average.

For Administrators

It is beneficial to interpret item-level data in addition to the ICM when interpreting course evaluation results for summative purposes. Identifying trends and patterns is even more important when interpreting item-level results, as any one question cannot provide a complete snapshot of students' reported learning experience. In fact, "general" questions such as Ins06 ("Overall, the quality of my learning experience in this course was:") may be less precise than items measuring specific behaviours or experiences, as ambiguity encourages students to rely on heuristics to answer them. While the comparative data table in section 3 provides item-level means and medians as aggregated scores, interpreting the frequency distribution of items remains important.

Table 3. Typical Item Endorsement Percentages by Course Size.

Course Size	**Institutional Items (Ins) Typical Item Endorsement**					
	Ins01	Ins02	Ins03	Ins04	Ins05	Ins06
Very Small (1-25 students)	62% - 100%	67% - 100%	67% - 100%	60% - 100%	62% - 100%	50% - 100%
Small (26-50 students)	56% - 97%	60% - 100%	57% - 100%	56% - 95%	57% - 100%	42% - 92%
Medium (51-100 students)	50% - 91%	57% - 93%	52% - 95%	50% - 90%	52% - 90%	36% - 85%
Large (101-200 students)	50% - 87%	56% - 91%	48% - 92%	48% - 85%	50% - 86%	33% - 79%
Very Large (201+ students)	48% - 85%	56% - 89%	47% - 90%	46% - 82%	48% - 84%	32% - 77%

Note: Values correspond to typical item endorsement percentages (i.e., 15th & 85th percentiles) of institutional items (Ins01 to Ins06) for all course-sections within the central CE framework from 2018/19 to 2022/23 academic years (CTSI, forthcoming). Item endorsement is the percentage of students who selected the two most positive Likert-scale response options (i.e., 'Mostly' and 'A Great Deal' for Ins01 – Ins05, and 'Very Good' and 'Excellent' for Ins06).

Step 5: Review Qualitative Comments

Included in Report Section 4: Qualitative Comments

Student comments add context and depth, helping depict students' perceived learning experiences in a manner that numbers alone can't capture.

Key Considerations:

1. **Identify Recurring Themes**: Pay attention to patterns in the comments. Repeated concerns or suggestions point to areas that may benefit from adjustments. Don't dismiss single comments too quickly; they could still highlight useful insights.

2. **Balance Comments with Quantitative Data:** Contextualize student comments by comparing them to quantitative ratings.
    - If the endorsement rates are consistently high, occasional negative comments may not be representative. However, repeated comments pointing to factors that may inhibit student learning likely indicate areas needing attention.
    - If student comments are aligned with quantitative ratings, the feedback from students is consistent.
    - If the two data sources diverge, consider gathering more input through mid-course evaluations or peer observations for further insight.

There is a multitude of strategies to identify recurring themes that emerge from student qualitative comments. Some people may prefer to use an exploratory approach to allow themes to emerge naturally from the comments (i.e., inductive coding). Other people may prefer using a coding template with pre-determined categories (i.e., deductive coding). An exemplar coding template is provided in Table 4 below.

For Administrators

When considering student open-ended comments for summative purposes, prioritize areas where students demonstrate significant consensus. It is possible that not all comments are meaningful, particularly if they are outliers. Additionally, resist giving negative comments undue focus. Positive comments may appear less "salient" than negative comments because they are usually shorter and less attention-grabbing.

For Instructors

Analyzing these comments systematically can reveal themes and inform targeted improvements.

Consider these guiding questions:

- Which aspects of your course do multiple students find helpful/positive? These are the successes you should try to build into the next iteration of the course.
- Which aspects of your course do multiple students find challenging? These are areas that require further reflection and that you may need to consider modifying.

- What recurring strengths could you highlight in your teaching dossier, or other narratives about your teaching (e.g., for teaching awards)?
- Are there other sources of feedback (e.g., peer observations) that could provide more insights?
- If you intend to make changes to this course in future, what feedback will you seek next time?

Once you have reviewed and categorized the qualitative comments, consider whether any comments could be incorporated into your teaching dossier. If the themes identified from these comments align with your course learning outcomes, and teaching philosophy, or fit within the broader narrative of your dossier, they may be worth including.

Table 4. Example Deductive Coding Scheme for Open-Ended Course Evaluation Comments

CATEGORY	Tally		Illustrative Quotations
	Aspects Students Perceive as Facilitating Their Learning	**Aspects Students Perceive as Impeding Their Learning**	
DELIVERY			
MATERIAL			
COURSE DESIGN			
ASSESSMENTS/ LEARNING ACTIVITIES			
TECHNOLOGY / TOOLS			

Conclusion

When integrating course evaluations into your teaching dossier, it's important to present them as one piece of evidence among several indicators of teaching effectiveness, such as:

- Mid-course feedback surveys
- Peer observation of teaching (formative, not summative)
- Unsolicited emails/letters from students or colleagues
- Examples of student work and outcomes
- Teaching awards
- Instructional grants

Teaching excellence is an ongoing, iterative process. Course evaluations can provide valuable insights into students' experiences and perceptions of teaching if proper consideration for contextual factors, trends/patterns, and multiple data sources is utilized. Following effective practice in interpreting course evaluations, including leveraging multiple sources of information, can help derive relevant insights.

References

Centre for Teaching Support & Innovation. (Forthcoming in 2025). *A Renewed Validation Study of University of Toronto's Cascaded Course Evaluation Framework*. Toronto, ON: Centre for Teaching Support & Innovation, University of Toronto.

James, D. E., Schraw, G., & Kuch, F. (2015). Using the sampling margin of error to assess the interpretative validity of student evaluations of teaching: Assessment & Evaluation in Higher Education. *Assessment & Evaluation in Higher Education*, *40*(8), 1123–1141. https://doi.org/10.1080/02602938.2014.972338