Waterloo Student Course Perception Survey

Guide for Academic Administrators

Last updated: January 2022

This user guide is a work in progress that will remain open for input, updates, and revisions. Please contact Sonya Buffone, Director of Teaching Assessment Processes (sonya.buffone@uwaterloo.ca) with any questions or concerns.

Contents

1 Introduction

Teaching assessment, and its longstanding reliance on student course perceptions (also referred to as "student evaluations of teaching" or "course evaluations") as a primary mechanism for data collection is riddled with conflicting perspectives, but most scholars in this field agree that student voices are essential when it comes to understanding how students perceive their learning experience, particularly at institutions of higher learning.

The challenge lies in finding a way to balance students' insights against their implicit and explicit biases, which can find their way into the feedback they provide to instructors and, indirectly, merit-based decisions made by academic administrators. While all evaluation metrics are subject to bias, the effects of bias may be compounded on members of equity-deserving communities, who often face discrimination, inequity, and injustice within academia, with serious implications for tenure, promotion, and job retention.

The research literature on student course perceptions is complex, with limited consensus on the extent to which different factors influence scores. Despite a substantial body of academic research to support the use of student course perceptions, concerns about bias have received significant attention in both the popular press and academic literature. What is clear is that context matters. An understanding of data gathered at the University of Waterloo is crucial to understanding the ways in which these interactions play out within this specific teaching and learning context.[1]

In the context of completing a student course perceptions (SCP) survey, bias may stem from:

- **student impression of instructor's perceived race, gender, sexuality, ability, etc.**
- **class size**
- **whether a class is online or face-to-face**
- **whether a student is taking the class out of interest or as a program requirement**
- **the time of day (or day of the week) that the class is scheduled**
- **a student's expectations of their course grade**
- **student perception of workload**

The potential for these kinds of variables to negatively impact SCP scores to varying degrees has been well-documented, including the results of a pilot study conducted at Waterloo.

1.1 The role of academic administrators

It is imperative that academic administrators take potential bias into account when reviewing and interpreting the results of SCP surveys, to ensure that instructors undergoing review are not unfairly impacted.

1.2 About this user guide

This document seeks to provide guidance with respect to interpreting scores collected through the Student Course Perceptions (SCP) survey. The purpose of this guide is to help academic administrators

[1] Compared to research literature conducted, for example, at a university in France (Boring, Ottoboni & Stark, 2016), a five-week online course in one department (Mitchell & Martin, 2018), the Netherlands (Mengel, 2018), in a male-dominated field (Burnell, Cojuharenco, & Murad 2018), or at a male-dominated American military college (Carrell & West, 2010).

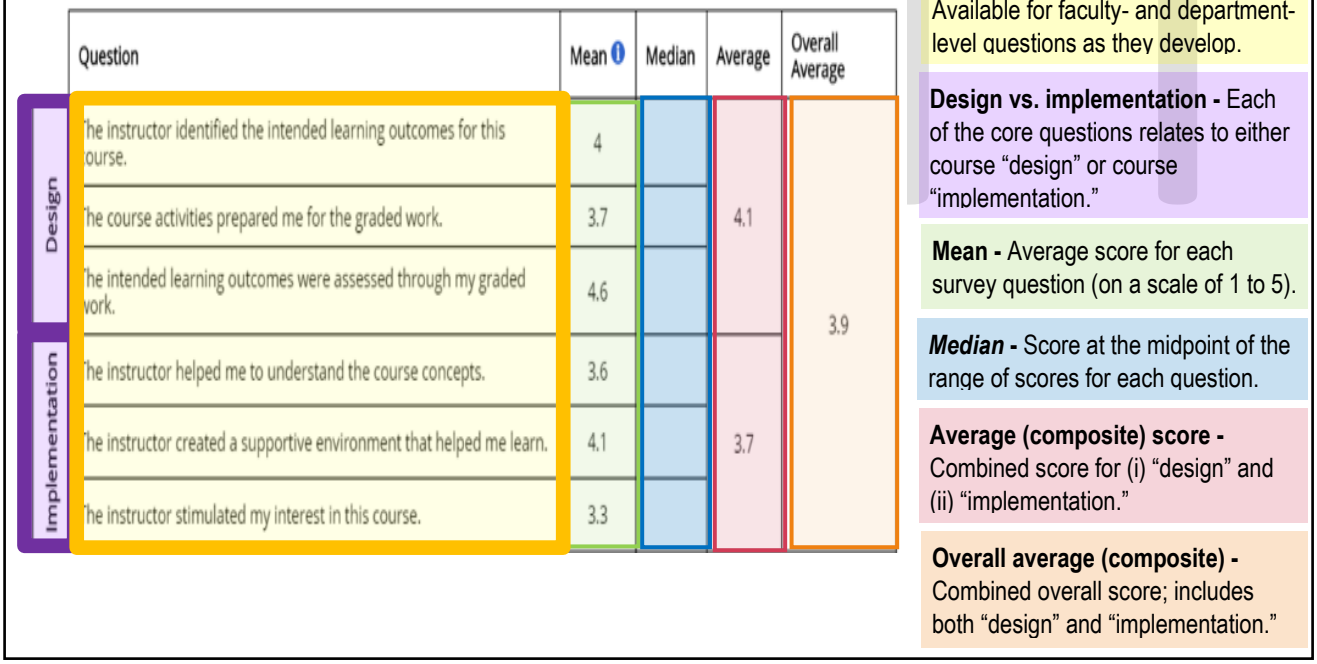
at Waterloo interpret SCP scores and to provide guidance for understanding the contextual factors that may impact those scores. The guide will be updated and informed by ongoing testing and monitoring of the SCP process, discussions with advocacy groups at Waterloo, input from department Chairs and others acting in an administrative role, experiences and reports shared by other Canadian institutions, and continued review and analysis of the literature.

2 Overview of the SCP data report

The SCP data report summarizes the results of the SCP survey for each instructor/course. It currently contains the results of core measures collected for institution-level analysis and will eventually contain the results of faculty- and department-specific measures. The report provides an overview of all seven core survey items and their associated scores (Figure A) and statistics and a detailed view of the scores and statistics associated with each survey item (see Figure B). The report will also include the results of three open-ended questions.

The following figures illustrate the overview (Figure A) and detailed view (Figure B) of the core survey questions. Some faculties will develop faculty-level and/or department level questions in addition to the core items; the report will include similar figures for each level. The data in Figure A provide an overview of the scores for each question for a particular instructor for a specific course:

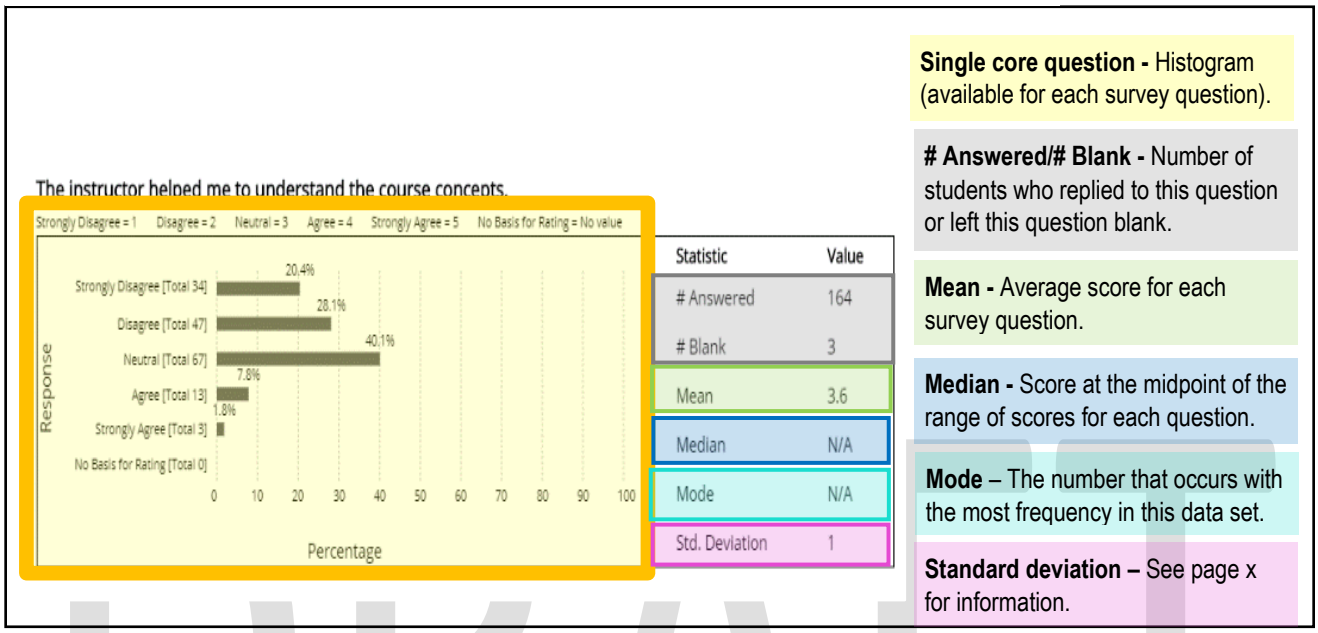
Figure A: Overview of survey questions and associated scores



The average (composite) scores allow academic administrators to focus their review on design, implementation, or both, depending on the instructor's involvement with the course. **Scores relating to course design should not impact an instructor who was only responsible for delivering lectures.** Being mindful of large discrepancies in these scores (e.g., a pattern of weak scores more focused on

design than implementation) can facilitate mentoring and advising with respect to targeting areas for instructor improvement.

Figure B: Detailed view (histogram) of the scores associated with each survey question.



The data in Figure 2 tell us that almost half of the students in this class selected "neutral" in response to the item, "The instructor helped me to understand the course concepts," and that 40% of respondents selected "strongly disagree" or "disagree." To the right of the histogram, #Answered indicates that 164 students answered this survey item; #Blank indicates that three respondents did not answer this item. The mean shows that this item has an average score of 3.6 with a standard deviation of 1.

3 Reviewing the SCP report

> Too often, SET (Student Evaluation of Teaching) systems have been […] uncontextualised, unsupported, simplistic and interpreted in isolated ways, features which render SETs punitive bureaucratic tools rather than supportive mechanisms through which enhanced learning environments can be created and sustained
>
> Moore & Kuol, 2005

The results of the SCP survey revolve around numbers: a five-point Likert scale rating system, means and standard deviations, charts, and tables. However, effective use of this tool requires a less tangible sense of how to interpret those numbers—requiring attention to context, an understanding of instructors' concerns, and the awareness to recognize when bias may be at play. This section of the Guide for Academic Administrators seeks to provide guidance with respect to interpreting SCP scores with these factors in mind. The following steps are critical to reviewing SCP results in a way that optimizes efficacy and recognizes the importance of equity in high-stakes decision-making: **Contextualize**, **critique**, **compare**, and **communicate**. In practical terms, these steps will intersect and overlap as academic administrators work through SCP results.

3.1 Consider context

Empirical research conducted at Waterloo and beyond indicates that scores can be affected by contextual variables. The first step is to recognize that some of these variables may be at play with respect to the scores under review. Given a lack of consistency in the research literature, the Teaching Assessment Processes team explored potential associations with data drawn from a cross-sectional University of Waterloo-designed pilot test of the new SCP survey.[2] The results showed that (1) **workload, expected grade, and class type (online vs. face-to-face)** were most strongly associated with SCP scores, and (2) there were smaller (but significant) associations between SCP scores and **gender** as well as **class size.** This first pilot test was unable to investigate instructor racialized identity: At the time of testing, these data were not being collected by the University. Plans are underway to look at this variable.

These quantitative data reveal important nuances specific to the Waterloo context, but other sources of data are equally important to consider, namely, the lived experience of Waterloo instructors. Cross-campus consultations confirm that instructors from underrepresented, marginalized, and racialized communities face discrimination, inequity, and injustice within academia. Evidence derived from lived experience should also be taken into account when reviewing SCP scores. Currently, these include instructor gender, racialized identity, and precarious employment.

Before reviewing Student Course Perceptions (SCP) scores, take a step back to think about context. Empirical research and consultations conducted at Waterloo and beyond indicates that scores can be affected by variables that do not reflect teaching effectiveness. The following variables (and their potential associations with SCP scores) reflect both empirical research conducted at the University of Waterloo and the results of consultations with Waterloo stakeholders.

Variable	Strength of association with pilot test results	Identified via consultation	Associated with lower score	Associated with higher score	Notes
Class size	**HIGH**	Yes	Large class	Small class	Larger classes (100-200 students) tend to receive lower scores than smaller classes. Pilot test results found this difference to be marginal but other U15 institutions found class size was a significant factor (0.05-0.13).
Expected grade	**HIGH**	Yes	Expecting lower grade	Expecting higher grade	Students who expected higher grades tended to give higher SCP scores (0.22 - 0.81 points higher for a grade >90).
Perception of workload	**HIGH**	Yes	Lower workload	Higher workload	Courses for which students rated the workload as "average" or "high" received higher SCP scores than courses perceived as having a "very low," "low," or "very high" workload (0.4 - 0.58).
Course format	**HIGH**	Yes	Online class	Face-to-face class	Online courses received lower scores compared to face-to-face classes. In-class courses received scores 0.1 - 0.28 points higher across SCP items. In practical terms this would be a difference between 4.3 and 4.6 on a survey item (0.1 - 0.28).
Instructor gender	**LOW**	Yes	Men	Women	Instructors at the University of Waterloo report that gender bias impacts SCP scores. The 2018 SCP pilot test suggested only marginal difference between male and female instructors overall and

[2] Statistical analyses were conducted by Waterloo's Statistical Consulting and Collaborative Research Unit.

					significant differences between male and female *probationary* instructors teaching large courses* (0.04 to 0.12). Male instructors scored lower in small classes (1-25 students) (0.1). This does not account for the possibility of interaction effects (e.g., gender/racialized status). *Interpret with caution: extremely small sample <5.
Instructor racialized identity	**Not measured**	Yes	Instructor racialized identity	Non-racialized identity	No data available
Workload	**HIGH**	Yes	Small workload	Big workload	Courses for which students rated the workload as "average" or "high" received higher SCP scores than courses perceived as having a "very low," "low," or "very high" workload (0.4 - 0.58).
Requirement status	**NEGLIGIBLE**	Yes	Required course	Elective	Elective courses received marginally higher scores than required courses. An increase of 1% in the proportion of students taking the course as "required" resulted in a slight decrease in average score for some SCP items.

The discrepancies between SCP pilot test findings and consultations with campus stakeholders reveal the importance of ongoing examination of these variables of interest. The Teaching Assessment Processes team will continue to investigate these variables to better understand the relationship between the SCP survey and instructor racialized identity, precarious employment, and gender, as well as the possible interaction effects that might influence these relationships.

The Teaching Assessment Processes team is committed to ongoing longitudinal analysis of SCP survey results as well as ensuring that SCP scores are triangulated with other methods of teaching assessment when these items (peer review and dossiers) become available.

3.2 Review scores with a critical eye

Review SCP scores with a critical eye. Ensure that you understand what the mean score does and does not represent, how the standard deviation of a score should be interpreted, and how the score relates to the typical distribution of ratings for this tool.

Don't over-rely on the mean score without incorporating other evidence and instruments

The mean score provides information about the "typical score" for students' perceptions of the quality of instruction for a specific course but does not provide a complete picture. The five-point Likert scale used for the SCP survey is **ordinal**, not continuous: it uses a scale that arbitrarily numbers an ordered series of labels ranging from "strongly disagree" to "strongly agree." With an ordinal scale, the difference between a mean score of 3.9 and 4.2 is not overly meaningful.

A continuous scale measures numerical data. We can measure numerical differences in dollar amount. If Amy has $5 and Ping has $4, we can say that Amy has precisely 1 dollar more than Ping.

An ordinal scale orders nominal data (e.g., categories) to make it possible to measure it in a numerical way. For example, if Amy received an overall mean SCP score of 5 while Ping received an overall mean score of 4, we can say that Amy obtained a higher score than Ping. To say that Amy is a "more effective teacher" by 1 point would be over-interpreting the numbers applied to the categories.

In fact, unless Amy and Ping taught the same course to the same cohort of students, any comparison of their scores is meaningless. All we can say is that a set of students rated Amy at a score of 5 and a different set of students rated Ping at a score of 4.

Consider the following fictional scenario:

Amir receives an overall score of 3.2. A closer look at the distribution in scores on the histograms for each item shows there are 4 extreme outliers (students who selected 2 on the scale for every survey item) but the rest of the scores are clustered between 4 and 5.

In this case, the mean score does not reflect most students' perception of this course.

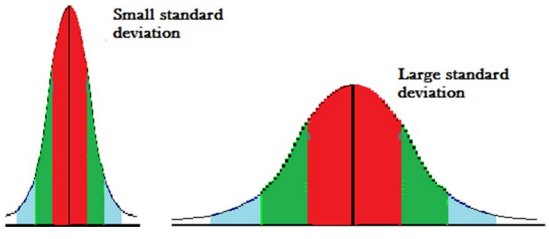
Be very cautious in assigning significant weight to Amir's low average. This score is the result of only a handful of students' experiences in this course. It would be advisable to examine Amir's scores in other courses to get a clearer picture of student perceptions of his teaching performance.

Consider the standard deviation

Standard deviation indicates the variability of data—the degree to which SCP scores vary around the mean. A higher standard deviation means that there is high variability in the data. A lower standard deviation means that there is less variability in the data. A low standard deviation inspires more confidence that the mean represents the 'typical case.' Note: the standard deviation can also be affected by extreme outliers.

A higher standard deviation (SD) means that there is high variability (less agreement) in the data. Dan received a mean score of 4.5 on the SCP survey, with an SD of 2**.** An SD of 2 is quite large in the context of a five-point scale (translating to a two-point differential in scores), which means that there was a lot of variation in students' responses. Anyone reviewing Dan's score should be cautious about interpreting it as reflective of the collective experience of all students in this class. *With a higher SD, administrators should take a closer look at the scores to see if they can identify discrepancies in students' experiences.*

A lower SD tells us that scores are close to the mean, meaning that there is less variability (more agreement) in the data. Mitra received a mean score of 4.5, with an SD of 0.5, which is quite small (see above comments). With a lower SD, we can be more confident that the mean measures the typical case.



An unusually high or low score may require investigation but should not be considered in isolation. The important thing to consider is that the mean is not the best measure to focus on because it is extremely sensitive to outliers. That is, a few low ratings can pull the mean downwards, which is particularly problematic for courses with a small number of students.

Response rate

Don't assume that average score represents the collective experience of the entire class. Exercise caution, especially when there is a low response rate. Scores associated with higher response rates are a better reflection of the collective experience.[3] For summative evaluations, high response rates are crucial.

The size of a course determines what should be considered an adequate response rate (Table 4). Generally speaking, smaller classes require a higher response rate to achieve the same confidence that scores reflect the collective experience. Larger courses result in more data than smaller courses, so even with similar response rates, a larger class is more likely to achieve a "reflective" estimate of the overall experience.

Academic administrators should be extremely cautious when interpreting course data from smaller numbers of responses. There is a high level of inaccuracy in mean scores calculated in samples with fewer than 20 respondents. In a smaller class, one unhappy student can seriously undermine the overall average. The outlier cases can have a significant impact on the overall score. With smaller classes, it is extremely important for administrators to review the distribution of scores.

Consult Table 4 to determine how much confidence you can have in the scores based on the level of "precision" for the response rate in a particular course.

Table 4: Response rate necessary to have a 'very reflective,' 'reflective,' 'somewhat reflective,' or 'generally unreflective estimate' of the collective experience of students based on class size. Adapted from University of Toronto data.

Interval around mean	Quality of mean estimate	Course size				
		1-25	26-50	51-100	101-200	200+
<±0.1	Very reflective of the collective experience	>90%	>80%	>80%	>60%	>50%
<±0.2	Reflective of the collective experience	>80%	>70%	>70%	>50%	>40%
<±0.5	Somewhat reflective of the collective experience	>70%	>50%	>40%	>20%	>10%
<±1.0	Generally unreflective of the collective experience	>60%	>20%	>10%	>10%	>10%
>1.0+	Not at all reflective of the collective experience	<30%	<10%	<5%	<3%	<1%

[3] If every student in the class completed the SCP, the response rate would be 100% and indicative of the "true" collective experience of the entire class. In reality, however, a response rate of 100% is extremely rare so we have to rely on an estimate of the true experience of the collective. This is important because estimates always include some degree of measurement error.

3.3 Compare carefully

Take a closer look at any scores that stand out as unusual. Look for patterns. Pay attention to scores that are higher or lower than expected or that reflect unusual patterns over time or between courses. Investigate the reason by revisiting context.

Focus on an instructor's scores across different types of courses. An occasional low rating is to be expected, but a long-term pattern of low scores might signify something to be concerned about and whether, for example, there is bias or contextual circumstances at play (i.e., the instructor is consistently tasked with teaching difficult or required courses) and/or if the instructor may require guidance from one of the academic support units (i.e., the Centre for Teaching Excellence).

Focus comparisons on sections of the same course over time. Comparing scores from a first-year course to those from a fourth-year class is not useful. Similarly, scores from an undergraduate course should not be compared to scores from graduate-level courses.

Focus on patterns of results within courses over time, rather than single assessments, but exercise caution in making comparisons. A single assessment might indicate areas of potential strength or weakness, but patterns of results help to identify trends.

Avoid comparing online and in-class courses. Scores for online courses tend to be lower than scores for in-class courses.

Exercise caution comparing large and small classes, or large and small sections of the same class. Larger classes tend to receive lower scores than smaller classes. These details will be different for different departments.

Avoid focusing on small decimal differences (e.g., 4.2 vs. 4.3). Differences of a decimal point or two are not meaningful.

When using SCP scores to inform decisions about an instructor, refer to several of their SCP scores over time.

Why do larger classes tend to receive lower scores than smaller classes?

- It can be harder for instructors to establish rapport with students in large classes.
- It can be more difficult to facilitate group work and discussions in larger classes.
- Smaller classes often have more opportunities for students to interact with classmates, speak up in class, ask questions, and establish a relationship with the instructor.
- In smaller classes instructors can more reasonably assign lengthier writing assignments, more graded homework, and essay exams (vs. multiple choice), which could lead to greater student learning (Hativa 2013b; Marsh 1987). If students feel they have learned more, they are more likely to assign higher ratings (Benton & Cashin, 2014).

Consider the following fictional scenario:

Alli receives a score of 3.5 on a course that she taught recently, which seems a bit low for a course that you know is traditionally a popular elective.

To dig deeper, scan Alli's scores from previous courses. You notice most of her scores are between 4.5 and 4.8, but with small class sizes. When you dig even deeper you realize that this was Alli's first time teaching a course with more than 70 students.

This lower score is not necessarily a reflection of Alli's poor teaching (from the student perspective); it could very well be a result of her inexperience teaching large classes. Alternatively, Alli may have used an innovative teaching method for the first time and students may have had difficulty adjusting to this new method. The important thing to recognize is that **context is crucial**. Do not assume lower evaluations equate to poor teaching.