Guidance for interpretation of course evaluation results

The purpose of this document is to provide guidance on the use of course evaluation results. We focus on the instructor assessment questions within the course evaluations, and discuss how to interpret their summary statistics as they relate to instructor effectiveness.

Key takeaways:

- Students' evaluations of courses are influenced by factors unrelated to teaching effectiveness.
- Only trends/patterns observed across multiple courses and multiple semesters, interpreted with consideration of their broader context, should be used as one source of inferences about an instructor's effectiveness.
- The margin of error can be used to interpret the precision of the mean score received on an evaluation question. For example, for a course with 25 respondents and a respondent mean score of 4.0, the true mean is somewhere between 3.7 and 4.3 with a 95% confidence level. The fewer the number of responses, the larger the margin of error.
- The median score is less susceptible to extreme values and often provides a better summary of student's responses than the mean score.
- Apparent differences between two means scores are often due to chance and thus cannot be used to infer a difference in instructor effectiveness. The difference between two mean scores can only be taken as evidence of a difference in instructor effectiveness if the distribution of scores and class size are taken into account, ideally by testing the difference against chance.

Course evaluation results are useful for identifying trends in students' reports of their experiences in courses. While students' evaluations provide valuable information and are one metric for evaluating instructor's effectiveness, they are limited in that they represent students' reports of their experiences, which are not an objective measure of an instructor's teaching ability or performance. A student's perception of the instructor is influenced by a number of variables unrelated to teaching effectiveness. For example, analysis of course evaluation results at Lehigh has shown that the mean score varies between departments and colleges, and tends to decrease with class size and increase with class level. Research has also shown that other variables such as time of day of the course, instructor grading reputation[1][2], and instructor gender[3][4] also influence the mean score.

It is also important to carefully consider the types of summary statistics available when assessing course evaluation results, as discussed in greater detail in the sections below. While the mean score is commonly used, if the number of responses received is small or if the standard deviation of the responses is large, the mean score should not be trusted as a good summary of students' reported experiences. The margin

[1] https://pdfs.semanticscholar.org/df16/2c7413b2085ba184df1927171adc0966f338.pdf

[2] https://www.washington.edu/assessment/course-evaluations/reports/course-reports/adjusted-medians/

[3] MacNell, L., Driscoll, A. & Hunt, A.N. What's in a Name: Exposing Gender Bias in Student Ratings of Teaching. *Innov High Educ* **40,** 291–303 (2015). https://doi.org/10.1007/s10755-014-9313-4

[4] Mitchell, K., & Martin, J. (2018). Gender Bias in Student Evaluations. *PS: Political Science& Politics, 51*(3), 648-652. doi:10.1017/S104909651800001X

of error is useful in determining how good of an approximation the mean score is as a measure of students' actual reported experience in a course.

It is important to keep these limitations in mind when considering course evaluation ratings, acknowledging that they should only be used alongside other indicators when assessing instructor effectiveness. In particular, one should be very cautious when comparing mean scores across instructors or courses, or when using mean scores to identify effective or ineffective instructors. Only trends/patterns observed across multiple courses and multiple semesters, interpreted with consideration of their broader context, should be used as *one* source of inferences about an instructor's effectiveness.

Evaluation process

End-of-Term Course Evaluations at Lehigh are administered online via EvaluationKit, using the evaluation form available here. For most courses, evaluation forms are open for students to complete during the last week of classes. After evaluation forms have closed and final grades have been submitted, the responses for each course are made available in an aggregate format to instructors, department chairs and coordinators through EvaluationKit. The responses are further summarized as needed for other reporting purposes (Faculty PAR, Department review, etc.).

Six Likert-Scale questions on a scale of 1 (Disagree Strongly) to 5 (Agree Strongly) are used to assess the instructor of the course. For each of these six questions, the students' responses are summarized in the EvaluationKit report in the form of a mean score, median score, standard deviation of responses, and number of respondents. Within each course summary report, the mean score, median score, standard deviation and number of respondents are also provided for the Department and College that the course belongs to in order to provide some context for that specific course.

Summary Statistics

Number of respondents

The number of respondents for a specific question is the number of students in a course who responded to this question on the evaluation form, the sample size for the statistics that are captured for that course. The closer this number is to the number of students enrolled in the course, the more likely the results are to reflect the overall perception of students in the course. Likewise, the fewer number of responses in relation to the number of students enrolled in the course, the less meaningful the evaluation results.

Mean score

For each question, the mean score is the arithmetic mean of the responses, or the sum of the individual student ratings divided by the number of students who answered that question. The mean is affected by extreme scores or outliers, especially when there are not very many responses.

Median score

The median score is the score at the midpoint of all the student responses, where half of the responses are above it and half are below it. The median is a more robust statistic than the mean because it is less susceptible to outliers. When available, the median often provides a better summary of students' responses (especially, for example, when there are extreme scores).

Standard deviation

The standard deviation is a measure of the variation of student responses around the mean score. Since the instructor questions are on a five-point scale, the standard deviation will be between 0 and 2. A larger standard deviation indicates that the responses are spread out from the mean, whereas a smaller standard variation means that the responses are closely clustered around the mean. As such, the standard deviation is a measure of consistency among the respondents.

Interpreting the mean score

As was mentioned above, if the number of responses received is small or if the standard deviation of the responses is large, the mean score should not be trusted as a good indicator of students' reported experiences. The margin of error helps us determine how good of an approximation the mean score is as a measure of students' actual reported experience in a course.

The table below provides an estimate of the 95% margin of error for the mean based on the number of respondents. This margin represents the distance from the mean within which the true population mean will be 95% of the time.

	Small Sample			**Large Sample**		
Number of respondents	**5-10**	**10-19**	**20-29**	**30-49**	**50-99**	**>=100**
95% Margin of error	0.81	0.46	0.30	0.23	0.28	0.13

Table 1: 95% Margin of Error estimates by count of respondents. The margins in this table are calculated based on the minimum number of respondents in each bin, and using a standard deviation of 0.65 (average standard deviation across questions I1-I6 across all courses in Fall 2018, Spring 2019 and Fall 2019). See section "Additional resources" for more detail on the calculations.

For example, for a course with 25 respondents and a respondent mean score of 4.0 on a given question, we can conclude that the true mean is somewhere between 3.7 and 4.3 with a 95% confidence level.

Comparing mean scores between courses

Apparent differences between two mean scores are often due to chance and thus cannot be used to infer a difference in instructor effectiveness. As such, a simplistic comparison between two courses is a misuse of the course evaluation results. The difference between two mean scores can only be taken as evidence of a difference in instructor effectiveness if the distribution of scores and class size are taken into account, ideally by testing the difference against chance.

A two-sample independent t-test can be used to determine if the difference observed between two mean scores is statistically significant and not likely to be due to chance. A t-test can be performed using any statistical software given two sets of responses. It can also be performed using the number of respondents in each course together with the mean and standard deviations of the responses. For example, if considering two courses with 10 respondents (median count of respondents across Fall 2018, Spring 2019 and Fall 2019) and equal standard deviation 0.6 (median standard deviation across the same three semesters), a difference smaller than about 0.57 between their respective means is not statistically significant at a level of p=0.05. Due to the large variability in class size and spread of scores, one cannot generalize this example to all Lehigh courses.

Calculating the margin of error

This section explains how to calculate the 95% confidence interval for any given course and question using the mean score standard deviation and number of respondents. Table 1 above can be used as an approximation when it is not possible or appropriate to calculate course specific confidence intervals.

The standard error of the mean is the standard deviation of the sampling distribution of the mean. It indicates how much variability there is across samples from the same population. Large values indicate that the mean from a given sample may not be an accurate reflection of the population from which the sample came.

The standard error is calculated using the following equation:

$$Standard\ Error = \frac{Standard\ Deviation}{\sqrt{Number\ of\ respondents}}$$

The standard error is used to calculate a margin of error, a margin around the sample mean within which we expect to find the true population mean with a specific confidence level. Most often, a confidence level of 95% is used. The resulting range around the mean is called the 95% confidence interval (95% CI).

The margin of error is calculated using the standard error and the size of the sample, or the number of respondents. For large samples (30 responses or greater), we can assume that the sample means follow a normal distribution and, therefore, the 95% CI of the sample mean can be calculated as:

Sample Mean ± 1.96 * standard error

For small samples (less than 30 respondents), we cannot assume that the sample means are normally distributed. Instead, the sample means follow a *t*-distribution, with degrees of freedom obtained by subtracting one from the sample size. The 95% CI of the sample mean can be calculated as:

Sample Mean ± t_{n-1}* standard error

Where t_{n-1} refers to the value of *t* for a two-tailed test with probability of 0.05 for n-1 degrees of freedom.