LMU|LA
Center for Teaching Excellence

Guidelines for Interpreting Student Teaching Evaluations

Student teaching evaluations are the most commonly used measure for evaluating teaching in higher education. There are at least two purposes for evaluating teaching: to improve the teaching and to make personnel decisions (merit, retention, promotion). When using student teaching evaluations for either of these purposes, it is essential to follow certain guidelines to ensure valid interpretation of the data. The following guidelines are adapted from Theall and Franklin (1991) and Pallett (2006).[1]

#1. Sufficient Response Ratio
There must be an appropriately high response ratio.[2] For classes with 5 to 20 students enrolled, 80% is recommended for validity; for classes with between 21 and 50 students, 75% is recommended. For still larger classes, 50% is acceptable. Data should not be considered in personnel decisions if the response rate falls below these levels.

#2. Appropriate Comparisons
Because students tend to give higher ratings to courses in their majors or electives than they do to courses required for graduation, the most appropriate comparisons are made between courses of a similar nature. For example, the Bellarmine College of Liberal Arts average would *not* be a valid comparison for a lower division American Cultures course.

#3. When Good Teaching is the Average
When interpreting an instructor's rating, it is more appropriate to look at the actual value of the rating instead of comparing it to the average rating. In other words, a good rating is still good, even when it falls below the average.

#4. Written Comments
Narrative comments are often given great consideration by administrators, but this practice is problematic. Only about 10% of students write comments (unless there is an extreme situation), and the first guideline recommends a minimum 50% response threshold. Thus decisions should not rest on a 10% sample just because the comments were written rather than given in numerical form! Student comments can be valuable for the insights they provide into classroom practice and they can guide further investigation or be used along with other data, but they should not be used by themselves to make decisions.

#5. Other Considerations

- Class size can affect ratings. Students tend to rank instructors teaching small classes (fewer than 10 or 15 students) most highly, followed by those with 16 to 35 and then those with over 100 students. Thus the least favorably rated are classes with 35 to 100 students.
- There are disciplinary differences in ratings. Humanities courses tend to be rated more highly than those in the physical sciences.

#6. One Final Point
Teaching is a complex and multi-faceted task. Therefore the evaluation of teaching requires the use of multiple measures. In addition to teaching evaluations, the use of at least one other measure, such as peer observation, peer review of teaching materials (syllabus, exams, assignments, etc.), course portfolios, student interviews (group or individual), and alumni surveys is recommended.

Contact the Center for Teaching Excellence (310-338-5866) if you need assistance in adopting one of these alternate measures or have any questions about these guidelines.

[1] Pallett, W. "Uses and abuses of student ratings." In *Evaluating faculty performance: A practical guide to assessing teaching, research, and service.* Peter Seldin (ed.). Bolton, MA: Anker Publishing, 2006.
Theall, M. and Franklin, J. (eds.) *Effective practices for improving teaching*. New Directions in Teaching and Learning, no. 48, San Francisco: Jossey-Bass, 1991.

[2] The following describes how to compute the response ratio for a given set of forms from one section of one course . First, note the number (n) of forms returned and the number (N) of students in the class, compute the ratio n/N, and then multiply by 100% to convert the ratio to a percent. Then, for each question under consideration, from this percent subtract the percent of blank and "Not Applicable" responses. The result is the response ratio for that particular question. If the result does not meet the threshold recommended in Guideline #1 above, the data from that question should not be considered.