Understanding Student Evaluation Data: The Numbers

Promoting professional excellence for faculty and graduate students

Michael Palmer, Teaching Resource Center, University of Virginia, 2006

When a course is over and the end-of-semester student evaluations arrive, it's easy—and tempting—to judge the success of the course from the numerical data alone. The numbers don't lie—right? While the data can be meaningful and ultimately useful in improving your instruction and future iterations of the course, care must be taken to determine the significance and validity of the data. Here are some general principles and guidelines to help you get the most out of the numerical data reported on evaluations.

Response Rates[1]

The following class size and response ratio standards help ensure valid interpretation of student rating data:

Class Size (students)	Minimum Acceptable Response Ratio
5-20	80%
21-30	75%
30-50	66%[a]
50-100	50%[b]
> 100	50%

[a]75% recommended
[b]66% recommended

The Mean

The mean reported on U.Va.'s online evaluation system is an arithmetic mean, which is the sum of all the individual student ratings for a particular question divided by the number of students answering that question. By itself, this number has very little meaning.

Important Point: Research has clearly shown that numerical student evaluation data is positively biased, meaning that responses are skewed toward the positive end. Thus, the average rating on a five-point scale tends to fall around 3.4 rather than the expected 3.0.

The Standard Deviation

The standard deviation (std dev) is simply the "average" or "expected" variation around the mean. A large standard deviation indicates that the data points are far from the mean and a small standard deviation indicates that they are clustered closely around the mean. The standard deviation is a useful measure when interpreting the mean for data which are well (and normally) distributed. When large clusters of data exist on both ends of the ratings spectrum, the standard deviation is virtually meaningless.

Important Point: On a five-point scale (e.g. Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree), a standard deviation less than 1.2 indicates relatively good agreement.

Confidence Interval

Like other numerical data, student evaluation data have margin of errors and confidence intervals within which the true values *probably* lie. Knowing something about the "true" value of a response is sometimes more meaningful than a general statistic like standard deviation. While a confidence interval is not reported for items on U.Va.'s online evaluations, it can easily be calculated from the mean, standard deviation, and number of students responding to an item.

The first step is to calculate the margin of error:

$$\text{margin of error (95\%)} = \frac{\text{standard deviation}}{\sqrt{\text{number of responding students}}} * 1.96$$

Note: The value 1.96 is a statistically-based constant for a 95 percent confidence interval.

Then, simply add the margin of error to the mean for the item under examination to obtain the highest probable limit to the true score and subtract the margin of error from the mean to obtain the lowest probable limit. The range from the highest probable limit to the lowest probable limit is the 95 percent confidence interval.

[1] Theall M. and Jennifer Franklin. "Using Student Ratings for Teaching Improvement." New Directions for Teaching and Learning 48 (1991): 83-96.

To illustrate, if the response to a particular question has a mean of 4.35, a standard deviation of 0.45 and a sample size of 25 students, then the margin of error in the mean is calculated as follows:

$$\frac{0.45}{\sqrt{25}} * 1.96 = 0.18$$

Thus, the confidence interval is 4.35 + 0.18 = 4.37 and 4.35 – 0.18 = 4.33. In other words, the true value of the item—with 95 percent confidence—lies between 4.37 and 4.33.

NORMS

In order to accurately interpret and compare evaluation data, context is necessary. A norm, or reference group (e.g. course, department, type of class, etc.), provides this context.

Important Point: At U.Va., the norm is often too large (e.g. all department courses) to be meaningful for instruction and course improvement purposes.