

Why Most Teaching Evaluations Do Not Evaluate Teaching

Kurt C. Schaefer

Summary: It is not appropriate to evaluate a teacher—or anything else, for that matter—by comparing her course-evaluation average to the average of some peer group, even if the peer group is standardized for discipline.

Say that Jergen and Kurt form a doughnut taste-test panel. They are asked by a baker to rate Bismarck and French doughnuts on a five-point scale:

- 1 - Very Good
- 2 - Good
- 3 - Mediocre
- 4 - Bad
- 5 - Very Bad

After a taste, the results are:

	<u>Bismarck</u>	<u>French</u>
Jergen	1	4
Kurt	2	1
<i>Average</i>	<i>1.5</i>	<i>2.5</i>

What can we say with confidence from this information? Kurt prefers French doughnuts, and Jergen prefers Bismarcks.

Can we say that, on average, Bismarcks are preferred to French doughnuts (since 1.5 is smaller than 2.5)? Economists know from Intermediate Micro Theory that the answer is "no." In fact, Bismarcks and French doughnuts are preferred exactly the same number of times: once. Can we say that Kurt likes French doughnuts twice as much as Bismarcks (since 2 is twice as big as 1), or that Jergen likes Bismarcks just as much as Kurt likes French doughnuts (since one equals one), or that Jergen would be indifferent between one Bismarck and four French doughnuts (since 4 is four-times one)? In each case, no. But let "Jergen" and "Kurt" represent groups of students, and doughnuts represent faculty members, and you see that we are making just such statements when evaluating teaching by comparing evaluation averages. (This whole discussion sets aside the issue of constructing a meaningful instrument to measure preferences in the first place.)

Why aren't evaluation averages meaningful? Computation of an average requires both addition and division. In order to do these operations to the measure of a characteristic, one must first establish that the characteristic behaves the same way that numbers behave. (This would involve empirical study of the characteristic until a set of its properties could be set out as axioms. Then a "representation theorem" would be proven, establishing a correspondence between the characteristic and the real numbers.)

In the case of weight, for example, we can identify a true zero and an arbitrary unit of measure (or "interval"), and by observation we have established that the weight of two objects is equal to the sum of the individual weights. So there is a direct correspondence between the act of combining weights and the mathematical operation of addition. Since we also have a true zero, it makes sense to multiply and divide weights. So the idea of "average" weights makes sense. The same could be said of time and length. Such scales are called "ratio" scales.

For some things (like Fahrenheit and Celsius temperature) we can establish an interval, but not a true zero. These are "interval" scales. Addition and subtraction are then meaningful, but not multiplication and division. It makes sense to say "my temperature has gone up two degrees," no matter what its initial level. But it doesn't make sense to say "98 degrees is twice as hot as 49 degrees," since zero degrees doesn't indicate the absence of heat.

For utility (including preferences about teachers) we cannot even establish a meaningful interval. (Try explaining what a "util" is to a normal person.) We are forced to use an "ordinal" scale, a rank-ordering. Since rank-orderings don't behave like numbers, arithmetic is not meaningful here. No sensible interpretation can be attached to the numbers that result. Faculty teaching evaluations are conducted on an ordinal scale; therefore averages don't mean anything.

It would make more sense (though not perfect sense) to evaluate faculty by reporting the proportion of students that ranked the faculty member "good" or "very good" on our five-point scale above. This would at least leave us manipulating actual numbers--the number of students that agreed with the proposition "this teacher is good or very good," divided by the sum of this number with the number who did not agree.

This approach is not only more theoretically sound; it also reports the information in a more vivid, useful form. Try it the next time you need to evaluate a speaker series or conference session. In the doughnut example, if average rankings were reported the baker would go away thinking that most people found both varieties pretty tasty, since 1.5 and 2.5 are both in the "good" range of the scale. But if the proportion ranking the product "good or very good" is reported for both products, the baker finds that only half of the tasters think French doughnuts fall into this category.