What Were They Thinking: Can Student Responses to Teaching Evaluation Instruments Reveal the Qualities of an Effective Instructor?

William P. O'Dea*

OBJECTIVE

The object is to develop a set of simple metrics that department chairs can use to evaluate the data generated by the student evaluation of instruction (SEI) instruments used by their departments. I am operating on the premise that department chairs typically inherit an SEI instrument they had no hand in designing. The instrument may in fact be poorly designed. Even if the instrument is not statistically reliable, the faculty may still resist replacing it with a better alternative. At my institution, the Committee on Instruction spent several years and engaged in two pilot studies to produce a more reliable SEI instrument. Union opposition prevented the adoption of the new instrument. Department chairs are thus in the unenviable position of having to extract the maximum possible meaning from data generated by what could be viewed as a haphazardly designed research project.

RESULTS FROM THE FIRST PHASE OF THE PROJECT

The results of the student evaluations are not noisy. The ranking of instructors from most to least effective are very stable from semester to semester. In any given semester, the difference in the student responses to the "overall evaluation of teaching effectiveness" question [the summative question on my department's SEI instrument] between the most highly rated and lowest rated instructors is large and ranges from two to three standard deviations. For the most part, the results do not appear to be influenced by factors beyond the control of the faculty such as their gender, their ethnicity, enrollment in the course section, and the time of day at which the course was taught. What does seem to make a difference is the level of the course. The average response to the "overall evaluation" question is about .5 point higher, which with our scoring scale would be worse, in 100 level courses than it is the higher level course sections. The results are not influenced by factors under the instructor's control such grading policy and the response rate to the SEI instrument. [Since instructors control when the SEI is administered, the possibility exists that instructors can pick a day in which they expect the students in attendance to rate them highly.] In my department, assigning high grades is not a winning strategy to receiving favorable student evaluations. The results appear to be driven by what is going in the classroom. The questions remains: what are the qualities that differentiate more effective from less effective instructors?

*Department of Economics, Finance and Accounting, School of Economics & Business, State University of New York College at Oneonta, Ravine Parkway, Oneonta, NY 13820

A POSSIBLE APPROACH

Our SEI instrument has 13 items. The first 12 items deal with the qualities that my department thinks (or thought at the time it designed the instrument) that a good instructor should possess. The thirteenth item is the summative "overall evaluation of teaching effectiveness" item. Our scoring scale ranges from 1 (excellent or the equivalent) to 5 (poor or the equivalent). Our hope is that the students in completing the questionnaire will carefully answer the first twelve questions and then make a bottom line assessment. I have access to SEI data from 277 course sections taught in the Department of Economics Finance and Accounting (EFA) between the fall of 2008 and the spring of 2012. I used OLS to determine which of the 12 qualities have the greatest impact on the "overall effectiveness." The results are shown in table 1. I tested for the presence of heteroskedasticity using the Koenker-Bassett test and found none. As a variation on this theme, I eliminated very highly ranked course sections (overall rating of 1.5 or less) and very poorly rated course sections (overall rating of 4 or higher). The change in the estimated coefficient was negligible.

Table 1: OLS Regression Results

Variable	Estimated Coefficient	Standard Error	t-Statistic
Constant	-.021	.062	-.337
Rigor of the Course	-.013	.032	-.433
Organization	.132	.047	2.783**
Teaching Skill	.361	.056	6.475**
Coordination of Materials	.142	.052	2.744**
Poise	-.01388	.036	-.384
Planning and Clarity of Exam Questions	.084	.034	2.469*
Ability to Answer Questions	.22	.039	5.537**
Tolerance	.043	.033	1.294
Maintain Control	-.069	.026	-2.606**
Availability for Consultation	.032	.0333	.956
Demanding grading	-.041	.043	-.955
Fairness of Grading	.158	.033	4.799**

Number of Observations=277; R^2=.9821; adjusted R^2=.9629; F=599.48
**significant at the 99 percent confidence level *significant at the 95 percent confidence level

Overall, the twelve items account for 96 percent of the variation in the "overall effectiveness" ratings. This suggests that the student ratings are not much influenced by factors that our instrument does not address such as sense of humor, wardrobe choices, or physical attractiveness. The most highly valued qualities are teaching skill by which we mean the instructor's ability to deliver a coherent presentation and the ability to answer questions. These findings are consistent with the SEI literature and match up well with our intuition. Also important are the organization of the course, the coordination of instructional materials with the lecture, and fairness of grading practices. Again, there are no surprises here. What might be surprising is that delivering a rigorous and demanding course and maintaining demanding grading practices have a negligible impact on the overall SEI ratings. It could be that our students expect economics, finance and accounting courses to be hard and are not overly surprised when this expectation is met.

BUT

Given the results shown in table 1, it would seem that I as a chair can declare victory and tell my faculty that as long as they are well-organized, can deliver an effective lecture and make sure that they answer questions to the students' satisfaction they should be fine. [I would also encourage them to use good examples to illustrate their theoretical points and involve the students in their presentations.] The problem can be found in table 2 which shows a portion of the correlation matrix between the items in the SEI questionnaire. The first column shows that "overall evaluation of teaching effectiveness" is highly correlated with most of our qualities of good teaching. This is not a problem. The problem can be seen in examining the "organization" and "teaching skills" columns. It is readily apparent our qualities of effective instruction items are not only highly correlated with the 'overall evaluation question", which is a good thing, but they are also highly correlated with each. The correlation between some of the items, for example between organization and teaching skills, are well above the levels which would indicate the presence of multicollinearity. In fact, a regression of "overall effectiveness" on teaching skill alone produces an adjusted R^2 of .92.

The high correlations suggest that highly rated instructors perform well on most of the constituent items in the SEI and conversely for poorly rated instructors. It could be that highly rated instructors being aware of the criteria on which their teaching will be evaluated adjust their teaching methods accordingly and poorly rated instructors are uninterested in conforming to the SEI. There is another angle from which to approach the high correlations. We hope that the students are evaluating each of the attributes of effective instruction carefully before they make their bottom line evaluation. Is this true?

Table 2: Correlation between SEI Questions

	Overall	**Rigor**	**Organization**	**Teaching Skill**	**Coordination of Materials**	**Poise**
Overall	1					
Rigor	.05	1				
Organization	.94	.14	1			
Skill	.96	.12	.94	1		
Coordination of Materials	.95	.11	.94	.96	1	
Poise	.86	.17	.84	.88	.83	1
Exam Questions	.91	.06	.89	.9	.91	.83
Ability to Answer Questions	.94	.07	.9	.94	.89	.89
Tolerance	.79	-.12	.74	.74	.74	.66
Control	.79	.03	.82	.79	.81	.71
Availability for Consultation	.79	.07	.72	.75	.77	.68
Demanding Grading	.16	.85	.21	.26	.23	.31
Fairness	.76	-.04	.7	.66	.7	.6

WHAT ARE THEY THINKING

One hypothesis that has been advanced in the SEI literature is that student responses can be manipulated through “priming.” The story is that students are asked a series of questions about behaviors that even ineffective instructors can exhibit. For example, did the class meet every day or were examinations given on the day specified in the syllabus. In answering a series of softball questions, the student falls into a fugue state and when they get to the summative question they have been “primed” to bubble in the excellent answer. This hypothesis has not been tested empirically. I don’t think that our instrument biases our students to give students an excellent rating to the “overall evaluation” question, because our last two questions deal with grading policy, which is a an issue likely to capture the attention of our students.

A second possibility is that the students don't consider any of the individual items particularly carefully. Rather, they start by making an overall assessment ("Professor A is wonderful" or "Professor Z" is a waste of oxygen"). Having made this assessment, the student then bubbles in the rest of the responses accordingly. That is, if the student's gut level assessment is that the instructor is excellent (1 on our scale) the student simply fills in 1 for the other 12 items. Encouraging this type of behavior is that if the evaluation is given at the end of the period, filling out the SEI form is all that stands between the student and freedom. Bubbling in a series of 1s takes about 15 seconds which gains the student about 4 minutes to check text messages or play "Angry Birds." If either of these stories is correct (or approximately correct), we would expect to find student SEI sheets with runs of the same response. Of course if every student filled out the SEI forms in this manner then in an analysis that uses course section level data the result would be perfect multicollinearity. In order to have a statistical estimation problem at least some of the students have to vary their answers.

WHAT I DO

To get a sense of how the students fill in the questionnaires, as a one-off exercise I decided to examine the individual SEI forms from all course sections taught in EFA during the fall of 2012. The students filled out a total of 957 questionnaires. The only piece of information that we ask the students to provide is their cumulative GPA for all courses taken at SUNY-Oneonta. The students are offered six ranges (e.g. 3.5-4) and asked to select one. First semester students who don't have a SUNY-Oneonta GPA would bubble in the "no GPA" option. The students are also offered the opportunity to make a written assessment on the back of the SEI form. I have no information about the identity of the students who completed the SEI sheets. Indeed, even to attempt to match the sheets to the students who wrote them would be ethically dubious since we tell the students that their responses will be anonymous.

From the individual sheets, I then pulled out the following pieces of information: the respondent's GPA; whether the student made a written response; whether the response was positive; whether the response offered a substantive comment on the instructor's performance; the response to the "overall effectiveness "; the number of answers that were identical to the bottom line score; the spread between the high and the low response; and the spread between the high and low responses to the organization, teaching skills and coordination of instruction materials questions. I'm interested in whether the student made a written response because if a student is willing to take the time to make a written response it suggests that the student might have given some thought to filling out the questionnaire. Following this train of thought, I also identify those responses that made at least some attempt to identify the qualities of instruction that the student liked or disliked. I collect data on the number of answers that are the same as the bottom line assessment to establish whether the students are filling out the questionnaires in a rote manner. The spread between the high and the low responses measures the extent to which the students attempts to differentiate between the instructor's performance in each of the areas addressed by the 12 questions. The reason for looking at the spread between the high and low response to the subset of questions is that these are the questions where the correlations between the responses are especially high as shown in table two. For future work, I also identified the level of the course in which the SEI form was completed. I distinguished between 100 level courses, 200 level required courses, 200 level courses taken by majors, and 300 level courses. My thought is that students in these courses have different levels of interest in the subject and different expectations. For example, students in the 100 level courses night be taking the course to satisfy a related work requirement for their major and thus might want to take the easiest possible course. In future work, I will record the number of item responses that were higher than the "overall evaluation" and the number that were lower. Table 3 provides the summary statistics.

Table 3: Summary Statistics of Individual Student Responses to the SEI Questionnaire

Variable	**Mean (proportion)**	**Standard Deviation**
GPA*	2.93	.39
Written	.495(474/957)	
Written Favorable	.323(310/957)	
Written Substantive	.789(370/474)	
"Overall Effectiveness"	2.39	1.2
Number of Responses the Same as "Overall Effectiveness Response"	5.93	3.12
Spread between High and Low Response	2.21	.98
Spread between High and Low Response to Three Items	.69	.76
Responses from 100 level sections	467	
Responses from 200 level required courses	309	
Responses from 200 level major courses	84	
Responses from 300 level courses	97	

*Based on the 863 questionnaires that reported a GPA.

The response to the "overall effectiveness" questions indicates that on our rating scale our students consider the quality of instruction in EFA to be a little less than very good. When the students take the trouble to make a written response, they generally offer some concrete reason(s) why they like or dislike the course. On the typical SEI sheet, six answers are the same as the bottom line evaluation. This response suggests that neither the “priming” or “student as robot” hypotheses are correct. The spread between the high and low responses indicates that the students means that the students make at least some attempt to differentiate an instructor’s performance in the areas addressed by the individual questions. The spread between the high and low responses to the three highly correlated questions is much lower, which suggests that in any statistical analysis two of these questions are redundant (but which two?).

WHAT THE DATA TELL US

A good place to begin is by considering whether a relation exists between the reported GPA and the response to the “overall effectiveness” question. Do self-reported high performing students give higher ratings than lower performing students? The answer is they do not. The correlation between GPA and the “overall evaluation” is a miniscule -.068.

To determine, the relation between the willingness of students to discriminate on the SEI questions (i.e. to not bubble in the same answer all the way down the form), I split the overall population into three groups: students who gave the same response to nine or more questions that they did to the overall evaluation (more than one standard deviation above the mean), students who gave the same response to between three and nine of the questions as they did to the overall evaluation (the mean plus or minus one standard deviation) and students to who gave the same answer to two or fewer questions as they did to the overall evaluation question (more than one standard deviation below the mean). I then calculated the mean response to the overall evaluation question for the students in each group. The results are shown in table 4.

Table 4: Relation between Answer to “Overall Evaluation” Questions and Students’ Discrimination

Number of Answers the same as “Overall Evaluation”	Number of Students in Group	Mean Answer to “Overall Evaluation” Question	Standard Deviation
9 or more	254	1.52	.8559
8 to 3	554	2.57	1.107
2 or less	149	3.32	1.1854

The striking thing to note in table 4 is that the students who discriminate the least in answering the SEI give their instructors the lowest (best) evaluations. Conversely, students who discriminate the most give their instructors the highest (worst) evaluations. The differences are large and statistically significant. The question is what to make of these differences. It is possible that students who are pleased with the quality of the instruction that they receive are not inclined to give much thought to the specific qualities that an effective instructor possesses. It could also be that effective instructors really do possess all (or most of the qualities) that we think that a good instructor possesses. It also appears that once students start thinking carefully about the individual questions the "overall effectiveness" scores start to do down. The results in table 4 do suggest that when students give an instructor a high (unfavorable) "overall effectiveness score" they seldom give the same high score to the constituent items; i.e. they follow my mother's advice and think of something nice to say (or to be more accurate bubble in). The results also suggest that to the extent that there is variation in the explanatory variables it comes from students giving the least favorable evaluations.

As a variation to this exercise, I also broke the overall population into three subgroups based on the spread between the lowest and the highest scores: students with a spread between the lowest and highest answer of one or less (more than one standard deviation below the mean), students with a spread between the lowest and highest answer of between two and three (the mean plus or minus one standard deviation) and students with the maximum spread between the lowest and highest answers of four (more than one standard deviation above the mean). The results are shown in table 5.

Table 5: Relation between Answer to "Overall Evaluation" Questions and Spread

Spread between the lowest and highest responses	Number of Students in Group	Mean Answer to "Overall Evaluation" Question	Standard Deviation
1 or less	212	1.76	.8534
2 or 3	651	2.38	1.1266
4	94	3.9	1.038

The data in the table tell the same story. The responses of students who give instructors the best ratings tend to be the most tightly bunched. The responses of students who give the least favorable ratings are the most dispersed

.