WHAT DO STUDENTS" EVALUATIONS OF TEACHING (**SETs**) TELL US ABOUT TEACHER EFFECTIVENESS?

An Interview With Prof. Larry Lilliston *

OJ: Dr. Lilliston, you have been recommended to us as one having an unusual degree of experience with student evaluations of teaching (SETs). What specifically is your experience in this area?

Lilliston: As department chair for many years, I was responsible for seeing that the teaching performance was assessed properly. Just as relevant, as a psychologist who studies his field—and I am diligent in following this literature and contributing to the field—I have been privy to the primary academic literature that has focused on the SETs. I also led a study here at Oakland in the 1970s on the validity and usefulness of the SETs; this is the study that became known as the "Lilliston Report."

OJ: Before discussing the student evaluations of teaching (the SETs), would you describe where the SETs fit into an ideal teacher effectiveness evaluation process.

Lilliston: What I see as ideal is that student evaluations should play a very minimal role in the teacher evaluation process because of problems with them. If they are included at all when evaluating teachers for promotion *et cetera*, give

*Conducted by Sherman T. Folland

them an extremely low weight and find various other ways of getting at whether someone is a competent teacher. To be sure, there are some positive things about student evaluations, but I think that the positive things are outweighed by the negatives, by the problems.

Many people overestimate how useful student evaluations of teaching are. But, let me explain that I come at this from a different angle from most faculty. There is a field in psychology called Psychology of Measurement, sometimes it is called Psychology of Assessment. This field's research and its development of research techniques started right after World War I and has generated a voluminous amount of research and several scholarly journals dedicated to this subject. So, psychologists look at this differently, for us it is not just a practical, administrative necessity, but it is also a question of 'what does valid research tell us?' Through this perspective, we see that many faculty are overconfident in their knowledge about this, because it seems like a very simple idea: "Just ask the customers whether they are satisfied with the product." It seems like a simple idea, but this is much more complicated than most people think. The assumptions underlying the SETs reveal a very complicated situation.

OJ: What is the history of student evaluations of teaching?

Lilliston: Yes, there is a history to these evaluations. They weren't around when I was in college, and I have no reason to think that teaching has become better since

then. My old teachers were pretty damned good. And, if someone was not good, students and chairs and deans knew about it. If they were really extremely bad, the administration would take steps to either remedy the situation or maybe even to get rid of the person. One reason for the strength of my feelings about this, and I know that this sounds strange and radical, but I think our Oakland University professors generally are pretty competent.

I don't accept the idea that we all have to be somehow evaluated as "excellent" or even "above average" to pass the test. Being competent is pretty good, it's enough. We have a right to expect competence in our teachers, and if they aren't competent, we take steps for that, but that's good enough. We've become so crazy about this stuff, you've got to be "excellent" to get tenure.

After the historical beginning of student evaluations, I don't think teaching has improved, I think it is probably declining. In fact, now if you take chances, if you are really willing to take a risk, the use of student evaluations by your administration will probably deter you. The SETs really came into existence in the late sixties and early seventies. They did not arise from teachers or departments or whatever, they weren't responding to an epidemic of teachers who were not teaching well enough; they came from administration initiatives that can be accurately correlated with the rise of the corporate model. SETs didn't come from some 'strong professors', and there were only a few exploratory, often half-baked

attempts by students to conduct evaluations---like we had the "Oakland Undiapered" thing here that never amounted to much—they came instead from an era in which students had started to question everything. They came because administrators wanted to deal with people they were not happy with. SETs arose and began to be widely used, but then in a couple of years people began doing evaluations of the evaluations, asking: 'Are they good ?'; and, 'are they reliable?"

To test reliability, you give one evaluation one day, another evaluation on another day and then test to see if the two are consistent. Validity is tested, for example, by testing whether they are correlated with things like the experience of the teacher or the standing of the teacher. Do they correlate with what teachers themselves say when they visit classrooms. Here's what I see. When these studies were finished, very few of them supported student evaluations. The majority of them showed low or no correlations between all these things.

OJ: Do we have a history, here at Oakland, with the SETs?

Lilliston: We did that study at Oakland back in the 1970's on student evaluations, which was known as the "Lilliston Report", but it actually had many people involved in it. We tested for reliability by comparing the consistency of two evaluations, two weeks apart, toward the end of the semester, not a long period of time. People generally expect these SETs to be measuring some stable

characteristics—"is he a good teacher"—so you would expect the elements of the evaluation—e.g., "does he explain things clearly?", "is he helpful?", and so forth—to be highly correlated between the two evaluations. They were not insignificant but the correlations were very low like 0.40 or 0.50; you know 0.50 implies about 25% agreement. The only item that had—and we had samples all over this campus, because Oakland was thinking about going to a standard evaluation that all of us would give—the only item that had an acceptable reliability coefficient, and even that was only marginally acceptable, was the item that said: "I am a (Freshman, Sophomore, Junior...)". That reliability was 0.80, so there was about 60% agreement on what the hell class you're in!

OJ: What is the so-called "Dr. Fox Effect?"

Lilliston: There were a couple of studies back in the 70s that reported correlations were significant regarding teacher assessments and the experience of the teacher, but, the overwhelming majority did not show correlations. Then some studies back then took a different tack. They investigated variables that might influence the student evaluations, such as the famous "Dr. Fox" studies. These showed that the ratings were related to things that were not basic to teaching, such as "outgoing personality", "entertainment value"; there is certainly nothing wrong with teachers being outgoing or entertaining but there is something more to the story. Dr. Fox

studies had a guy come in and deliver the exact same lecture in terms of words and the lecture is meaningless but it sounds very realistic. It was laden with jargon and all that. And, for one group he gives it in a very standard lecture format, not intentionally dull but, in the other one he gives it in a cheerful, outgoing and kind of charismatic way—this is the same guy, he is an actor, same guy. Of course it's not just that students like the second guy most, he probably was more lively, but they also claimed that they learned so much more from him—yet they couldn't have learned anything from either of them, because it was gibberish. It was full of stuff like Prof. Irwin Corey, the old comedian on the Ed Sullivan Show. So, they learned lots of stuff.

OJ: Does giving higher grades result in higher student evaluations?

Lilliston: Yes, they give high evaluations on other things, and one of them that came up in these studies was the relationship of student evaluations and grades. There is a relationship. If you want to pump your evaluations up you can give higher grades. There are studies that show this. What you do, you don't just give high grades, you tell them how hard this material is: "You guys must be working really hard on this"; "You are doing really well on this". That will pump them up. Anyway, the research demonstrated that giving higher grades improved your SETs, yet you hear people talk about it today as if the results had come out the other way.

You know it's kind of like the research in psychology on ESP by real psychologists that stopped in the early years of the Twentieth Century, because their studies had shown that there was no such thing as ESP. So people kept talking about ESP because they falsely assumed that the research had shown it to be effective, which it had not. So these studies in the psychology of student evaluations are being revisited today, because you have to show people all over again what was settled earlier.

OJ: What is the "portfolio method" of teacher evaluation, and what is your opinion of it?

Lilliston: This is a good way to evaluate teaching, understandably we want people to be competent teachers. The portfolio method would mean basically that when you come up for tenure or promotion, you make your best case that you are a competent teacher, actually you might have to make your best case that you are a good teacher or even an outstanding teacher like that. I think to be competent is good enough.

What you do is you establish a portfolio that has evidence in it that you are or at least can be a satisfactory teacher. The kind of evidence that you might put into that—you could include your student evaluations in that—but you could choose to highlight whatever you wanted. The best is to have colleagues who are

experienced in the area to visit the class. Some faculty are concerned that some colleagues might take unfair advantage of them in that situation, but I think that most colleagues are very fair and would do the best job they could. I think there should be a system within each department regarding how class visits are done. Some methods make for a far better assessment than others. These visits should include, after hearing the lecture, interviews with students, and with the class as a whole. The visitor would ask students: "Was this a particularly difficult class? What is pretty typical? What you like best about this teacher? What might you want to see changed? Do you have some suggestions, can you give some feedback?" This is not kept from teacher.

So, class visits should be important and should be included in the evaluations by the department, by the committee. The materials used to teach. Syllabi, assignments, tests. Class visits should go on even after you are tenured. The goal should be to help ourselves be good teachers. So, you got evaluations, course materials, class visits, statements of the person on what their goals are. You should also have statements from people not solicited by the professor. I think it's a good rule that any letter that comes in that is not solicited by the committee is suspect. Because there are so many people who might just call up their old students, presumably ones that like them. But these could be solicited in a quasi-random way. Not each year. This would be for promotion, not each year.

So, you would have a portfolio and you'd have these things in the portfolio. Just like an artist's portfolio. You are responsible for assembling the portfolio. Just as an artist—and I imagine this applies to even the best artist in the world—her portfolio doesn't include every damn thing she has done. She says: "Oh, I want this one in, this one is representative, this is good." Artists that have portfolios aren't obligated to have everything they have done in it. And, I don't think a professor should be obligated to do so either, she should get to emphasize what she wants to. Suppose she gets some negative teaching evaluations, students say she is not very good: The candidate gets to say: No I don't want those faculty evaluations in there. In balance, that's a good system. What if a candidate comes forward with an empty portfolio, well, that just tells you that much more.

As a rule, in my department, since we do not have that kind of system, we do have classroom visits and the like. Student evaluations in my department are rated very low, if you look at our criteria you will see that they are down at the bottom. There is a lot of sentiment in the department to get rid of them altogether. This just reflects the fact that psychologists, with their background in this psychological measurement literature, tend to see these issues differently from faculty in many other departments, who often strongly prefer to keep the SETs.

OJ: Is the portfolio method expensive in either time or money?

Lilliston: What would be expensive? The things needed for it are already being done in my department each semester. The class visits? Some people say: "You know, it takes up too much time." To them I say, "Bullsh*t". You know, being a professor isn't hard manual labor, and most of us have very nice teaching loads, trust me, we do, in comparison to other schools. Every semester each faculty—the way I would like to see it—would have to visit two classes. And, you have to get written feedback to the teacher whose class you visited, so how hard is that? A one page report. And, this is for tenured and non-tenured, everybody. Two classes. How hard is that? That also means incidentally that my classes will get visited two times—I visit two times, everybody's getting visited two times. So for the person who is tenure track but not tenured, when you come up for tenure you have 24 class visitations to measure your teaching, either supporting it "good" "good enough" as most people actually are. If they have a bad day, throw those out, say "I've got 20." This is a lot. It's not expensive. Don't tell me that this kind of feedback is not really valuable. Don't tell me that some 19-year-old with no experience can do better.

OJ: What is the correlation between SET scores and independent tests of student performance?

Lilliston: There are studies that have been done on this. Those studies have often been done in departments where there are several sections of the same course: like Intro Chemistry, ... My recollection is that the correlations are statistically significant but pretty low.

Keep in mind that when we instituted the SETS we turned over to the teachers the sole responsibility for what a student learns. There is another variable in that connection, and that's the student and his or her motivation. We know that teaching is really a two-way dialogue and by treating the teacher as solely responsible we are assuming passive students, this is really what we do. You might have an active learner in your class and you might be a terrific teacher for that person. For others, passive learners (and most of our students are passive), you might not be a good match for the student. You may not be a good teacher in that case because your teaching rests on assumptions that "I have an active learner."

There has been recent research on teaching techniques that hinge on whether the student is a positive or an active learner. Those students who like Power Point and the web page stuff prove to be those students who are very passive in their approach. They tend to give positive ratings to those who use Power Point. Most of the teachers who rely on this technique in holding a class don't really give lectures. Students who don't like Power Point, in contrast, would rather hear a story. I don't mean an entertainment story, it's a little more complex. Students

who like Power Point often assert that they have learned better. However, when they are given a departmental test on the subject matter, the results show that they have learned less. One of the problems with student evaluations is that there isn't much of a relationship between student learning and the scores they give to the teacher.

OJ: What personal characteristics of teachers tend to be favored by students when assigning scores on the SETs?

Lilliston: I have mentioned some of these, such as congeniality, but there are others. Confidence. Come into the classroom glowing with confidence and good humor, and you will do well on the SETs, regardless of what you know. Another thing that I didn't mention was discussed in a study last year, I believe it was in *Psychological Assessments*. This study showed that those faculty who are physically attractive get higher ratings than those less attractive. Physical attractiveness has a small effect, but it is significant.

OJ: What makes a popular teacher?

Lilliston: There is no sin to being a popular teacher. The problems are the kinds of things that some teachers do just to gain popularity. I am absolutely sure that many popular teachers are excellent teachers. If you are talking about passive

students, popularity might help the students somewhat. If you have active learners then it doesn't matter. The active learner sees a professor differently.

There are people who sacrifice teaching standards in order to be popular. The SETs can be destructive in this way, that is, if the professors opt to be popular rather than do an honest day's work in teaching. The most destructive thing about them is the way they manipulate teacher behavior. But, what happens when we look over our SETs knowing that we know the stuff, knowing that we presented a good course, and yet we get low student evaluations. Many of us overreact to this. We become sullen or depressed, angry, outraged And the only reason is because we bought into the validity of the SETs. Then, our behavior changes.

Most of these student evaluations pit teacher against teacher. These student data are at best weak data and they are merely ordinal in nature. Even if you assumed that it were legitimate to calculate means and standard deviations from ordinal data, you'll get many people from departments, on CAPs and FRPCs, who won't know what a mean or standard deviation is. For example, they have no understanding that if you have a mean of 2.5 and the standard deviation is 0.7 that a score of 2.3 is not below the mean in any meaningful sense.

The other thing is that when you calculate means from these data, what do they *mean*? Say you are a very good teacher, but you are in a department of supermen. So, you come out below average. That's crazy! These numbers don't

mean that. Whenever you calculate a mean in this sense you are comparing people. You could be in a department where everybody is crappy, it doesn't mean you're a good teacher. I hate that. People are using these data to make life decisions about people, and they don't understand what these data mean. I hate things like Teacher Excellence Awards. I hate it when you start pitting teacher against teachers. This is a noble calling, and you pit people against each other effectively on the basis of how noble they are versus the next guy: "I'm nobler than you."

OJ: Should as some say, the SETs only be used to compare multiple sections of the same course?

Lilliston: No, all the problems would still apply. You can take a test and break down all the elements in that score. People who believe strongly in teaching evaluations are those who believe that what goes into those scores is almost entirely a genuine measure of teaching effectiveness. That is, they believe that it is all systematic variance. And, that that systematic variance relates to something in teacher assessment in some absolute context, that it is a pure, absolute evaluation of how good that professor is.

There are many things that go into it, for one thing a substantial degree of nonsystematic variance. But, even among the systematic variance, most of it is not just 'how good is the teacher' but it's all this other stuff. Likeability, grades,

personality characteristics, how much work you have to do. Those things are not just eliminated by comparing only one section of a course with another. They are there just the same. One section may be harder than another section, so they get lower ratings.

OJ: On-line course evaluations are being considered by several departments. What is your opinion of these?

Lilliston: To be candid, I haven't given it much thought, though, I have a general dislike for such things. I don't know how various departments do this. I have looked into the student on-line evaluations. There is no gate-keeping there. I could go on-line and give myself great evaluations. A friend of mine could do it. "Hey everybody come!"

If you control that problem on-line, I still think that is not the best way. For in-class evaluations, when you have students rate you, that have to be there to do so; the best way to control the response rate is to have students do this in class. Probably not a good idea. The way it should be done is to have someone, not the prof, come into the classroom and present the evaluations: "You should take these seriously, they are used for promotion etc." When it's on-line you have no control. It's naive to think that the response rate doesn't matter.

OJ: Have the SETs harmed anyone?

Lilliston: Yes, they corrupt the whole process. And, you don't have to be a brain surgeon to understand this. The teacher will think about his students: "My livelihood, my self-esteem, and my little babies are dependent on what you think of me. Even though I know that I should assign that extra book for my students to read for class, I'm concerned about the student responses: 'Oh he made us read a lot for class'." It corrupts the process, rather than teaching the way you think it should be taught.

Here's how my education worked. I'll be watching Jeopardy, and they'll have a question on it and I'll think, jeez, I ought to know that. I had that in classical literature, and I remembered that I had it. Then, I'll go back and I'll read the book. It was pretty good. So, it's not just the facts walking around in a student's head. It's these other things, you know. The criterion variable "learning" is very difficult to evaluate, once you start talking about what you think real learning is.

OJ: The evidence you describe seems to pile up against the SETs. Can you tell us, what keeps these student evaluations going on? Who benefits and who loses?

Lilliston: I think the reason they have kept going on historically is that people decided back in the 70s, incorrectly, that the research had shown them to be effective. That's the reason they keep going on. Another factor is that the system

has bought into them, buying into the assumption that the SETs actually work. Some teachers who get high ratings say they benefit, but the student evaluations don't benefit education generally and I don't think that they benefit the students. What genuine benefit there is to students is to empower them, give them an open channel to administration. But, there are other ways to do this. If you wanted you could empower students without the SETs, there are other ways.

Occasionally you may find that students aren't strongly interested anyway. Just take an example. Several years ago I thought of putting a box at the back of the room, and I told students that you can send me messages about the course, anonymous messages. You can tell me 'good job today', you can tell me 'I did not understand this, could you go over that again'? Just anything to guide me. I did this for two sections of psychology two semesters. Two messages. One of them was a little obscene. The other was something about not understanding a point, a legitimate message.

We could hand out evaluation sheets to students and say "put your evaluation numbers on the sheet, and the section number, and turn it in to the department secretary" and say "you'll be heard". The students have been talking for the last four or five years doing their own student evaluations. Some of these have said that students aren't able to give feedback and they want to be an aid to those students--in terms of who is good, what professors to take and all this. So, I

did a little study. I got a couple hundred students. I asked them basically "what would you like to know about a class, a professor, about the course, so as to choose the right class. Just write down what things would you like to know". They didn't mention anything about professors such as whether he's a good teacher, instead they wanted to know how easy it is, whether there are any term papers, how many tests are there going to be. Students really, our students anyway, want to know those kinds of things.