Analysis and Interpretation of the University of Scranton Course Surveys

Course Evaluation Committee
February 2009

1. Introduction

The University of Scranton Course Survey was designed taking into account the large amount of research conducted on course evaluation forms over the past several decades. Among other things, this research considered the structure of the forms (what questions should be asked) and tested for reliability (the consistency of a measure over multiple applications) and validity (the degree to which the measure gets at its target).

Research shows that students do not simply view instructors or courses along a single dimension, say, from good to bad. Instead, there are several relatively independent dimensions along which students may vary in their rating of an instructor or course. The University of Scranton Course Survey was designed using commonly agreed upon dimensions of effective teaching. These dimensions transcend different teaching styles and environments. One dimension, student learning, is addressed in the Progress on Objectives section of the form. Only items rated by the instructor as "important" or "essential" from a list of objectives appear. Seven other dimensions related to enthusiasm, rapport, fairness of evaluation, organization, providing context, in-class interaction, and appropriateness of learning materials are the subject of the Instructional Methods questions. (The questions themselves and the parenthetical examples on the evaluations serve as elaborations of the dimensions.) Lastly, a question dealing with workload is also asked.

Students often form an overall judgment about the quality of an instructor and of a course as a separate factor, likely some unique amalgam of the other dimensions. Therefore students are asked to evaluate the instructor and the course as a whole. A question asking for the student's initial interest in taking the course is included to facilitate the statistical analysis done on the evaluations.

Research indicates that, with an adequate number of raters (say, 15 or more), student ratings of instructors for a single course are quite reliable. For some of the dimensions described above, ratings are also reliable, while for others, those more influenced by the nature of the course, ratings can be less consistent. Validity of the student ratings has been supported by research on the correlation between student ratings and other assessments of teaching quality, including ratings by colleagues, administrators, and alumni surveyed several years after completing the course. Additional studies, involving the comparison of student ratings of instructors to average learning demonstrated in multi-section courses with sections taught by different instructors, showed modest correlations.

2. Comparative Analysis

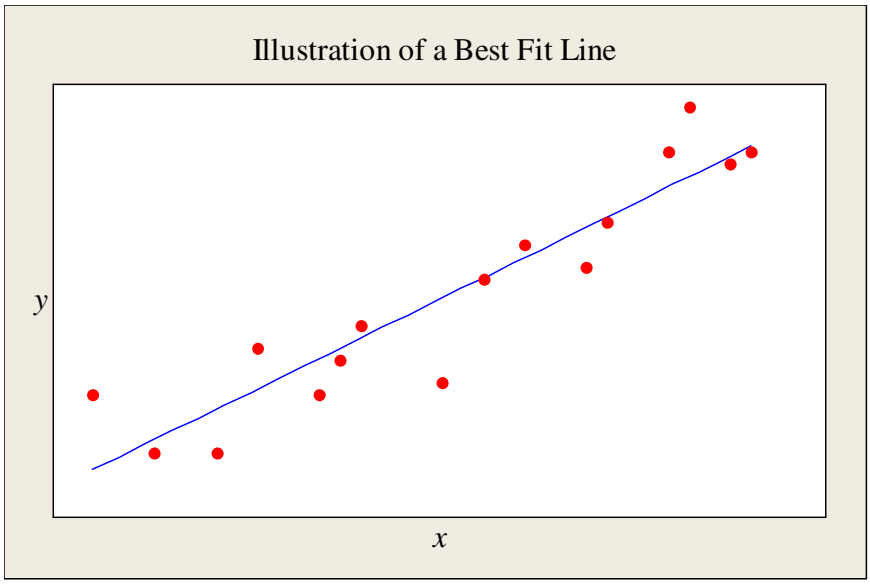
The comparative analysis done to course evaluations is a systematic process that determines how the student ratings for a particular course compare to the ratings on evaluations across the University of Scranton. This serves two purposes – it determines whether scores are significantly different from the campus average, and it allows correction for biasing factors that are out of the control of the instructor.

It is important to emphasize that the result of the analysis is a conclusion about how the ratings of a particular course compare to the mainstream ratings obtained by instructors *at this university*. In Spring and Fall semesters of odd-numbered years, it is mandatory that all courses at the University be evaluated by students, and all evaluations from that semester form the Comparison Group. For other semesters, the Comparison Group is the set of evaluations from the most recent mandatory semester of the same type (Fall or Spring). The comparative analysis compares an individual course to the Comparison Group.

An initial concern about course evaluation scores is their skew. In a perfect world, the average scores on a 1-5 scale would be near 3. However, course evaluation scores tend, in most cases, to have an overall average much higher than 3. (This is sometimes referred to as the Lake Wobegon effect – everybody is above average.) Because of this, one may rightly be concerned about the effect that a few low scores would have. (It might take three ratings of 5 to counter the effect of one rating of 1.) To remedy this, a transformation is done to the data to help eliminate the skew.

Another concern many have with course evaluation ratings is that those courses that are unpopular amongst students might receive lower ratings than those that are popular. Students' desire to take a course is measured by the question "Before enrolling, I really wanted to take this course REGARDLESS of who taught it." Indeed, a positive correlation exists between the rating on this question and the ratings on the other course evaluation questions here at the University of Scranton. The comparative analysis introduces an adjustment for this effect, which is out of the control of the instructor.

To explain how this is done, we use the Instructor Rating (IR) score as an example. (This refers to the response to the question "Overall, I rate this instructor an excellent teacher.") We abbreviate the Initial Student Interest score described above as ISI. After the skew-removing transformation is done, the average ISI and IR scores are calculated for each course in the Comparison Group. For each course, this forms a point with the ISI average as the x-coordinate and IR average as the y-coordinate. The best-fit line is then calculated for these points. This is the line that comes "closest" to passing through all of these points, and should be interpreted to represent the *expected* IR average for a given ISI average amongst courses at the University of Scranton. This line consistently has positive slope, meaning that higher IR averages are expected for higher ISI averages and vice-versa. The following figure is a simple illustration of a line that best fits a set of points.



The last step in the analysis process involves assessing each particular course. For a given course, the ISI average is calculated. The line described above then provides the expected IR average for that course. In virtually no cases will the actual IR average for the course match this expected value. It is important to determine whether this deviation is due to random error or indicates a significant difference. In this scenario, it is standard to perform a statistical test (called a *t*-test) to determine if the difference is significant. If the IR average of the course is found to lie statistically significantly above the expected IR average, then the IR ratings are labeled "Above Average." Likewise, if the IR average of the course is found to lie statistically significantly below the expected IR average, then the IR ratings are labeled "Below Average." Otherwise, no statistically significant difference exists, and the IR ratings are labeled "Average."

It should be noted that the Comparison Group for a Progress on Objectives item naturally consists only of those courses whose instructor rated that item as "important" or "essential." The Progress on Objectives categorization results from a weighted average of the *t*-test scores of the items, with the "essential" items counting twice the "important" items. This underscores the need for instructors to thoughtfully rate objectives for their courses.

The one item not characterized as above is the ISI question itself. These ratings are labeled "Very Low," "Low," "Moderate," "High," or "Very High" if the ISI average lies in the 0^{th}-20^{th} percentile, 20^{th}-40^{th} percentile, 40^{th}-60^{th} percentile, 60^{th}-80^{th} percentile, or 80^{th}-100^{th} percentile, respectively, of ISI averages in the Comparison Group. (An ISI average lying on the boundary between two groups is put into the higher group.) These labels are purely descriptive and do not represent the result of a statistical test.

3. Interpretation

Before drawing conclusions from any data, it is important to clearly acknowledge exactly what the data represent. In the case of course evaluations, the data represent students' perceptions of an instructor's accomplishment in the questioned items. They are no more or less than that. Assessing whether or not an instructor did a "good job" teaching a course is a complex endeavor, and the course evaluations provide only a piece to that puzzle. What students perceive as "good" and what faculty perceive as "good" do not always align. Nevertheless, course evaluations testing the dimensions we do have been thoroughly researched and found to be a reasonably valid, reliable way of measuring students' impressions and observations, and as those are valuable, so are the course evaluations.

It is important to note that, in order to ensure that a stable picture is obtained, researchers recommend that ratings be available for at least five or six courses, with an adequate number of raters (at least 15) from each course, before drawing conclusions. It is hazardous to draw firm conclusions from fewer sets of evaluations.

One must also exercise care in the interpretation of the "Above Average," "Average," and "Below Average" categorizations. These labels represent the relationship of one course to the population of courses offered at the University of Scranton. In particular, "Average" means within the mainstream of courses offered at the University of Scranton. If the level of instruction offered at this university is high, then "Average" means in line with that.

In particular, one may see "Average," or even "Below Average," assigned to ratings that seem to be high. This does not necessarily mean that the ratings are not high, but could indicate that the ratings across the University are also high. That ratings on an item are notably high (or low) should not require categorizations to be observed.

It may also be the case that, when one considers two different courses, one course may have a lower categorization despite having the same, or even higher, average rating. The most likely explanation is that the classes have different initial student interest averages. Even if that is not the case, differences can result in the *t*-test outcomes if the classes have different numbers of students or if their ratings have different standard deviations. This is not an error in the *t*-tests. To be categorized different from "Average" requires the surpassing of a threshold that depends on the standard deviation and class size. This is all the more reason to look at the entire package and not just the categorizations when considering course evaluation data.

4. Further Reading

The interested reader may further pursue matters related to the design and analysis of course evaluations by consulting the following sources. Cashin (1995), principal architect of the Kansas State *IDEA* form previously used at the University of Scranton, presents an excellent summary of the research. Marsh (1987), author of the *Students'*

Evaluations of Educational Quality (SEEQ) form, provides probably the most comprehensive review of the literature, including an important discussion of methodological issues. Cohen's (1981) meta-analysis is a crucial summary of one type of validity evidence: correlations with actual student learning. For other reviews of the research, see Abrami, d'Apollonia, and Rosenfeld (1997), Feldman (1997), Hogan (2003), Marsh and Dunkin (1997), and McKeachie and Svinicki (2006). All these sources show reasonable agreement on the major findings regarding course evaluation forms.

As mentioned above, the information provided from course evaluations by students should constitute a part, not the whole, of an assessment of an instructor's teaching effectiveness. For details of other useful methods of teaching assessment, the reader may refer to McKeachie and Svinicki (2006), especially Chapter 26, and Cashin (1990) or consult with members of the University of Scranton Center for Teaching and Learning Excellence.

References

Abrami, P. C., d'Aplollonia, S., & Rosenfeld, S. (1997). The dimensionality of student ratings of instruction: What we know and what we do not. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 321-367). New York: Agathon Press.

Cashin, W. E. (1990). Assessing teaching effectiveness. In P. Seldin and Associates (Eds.), How administrators can improve teaching: Moving from talk to action in higher education (pp. 89-103). SanFrancisco: Jossey-Bass.

Cashin, W. E. (1995). *Student ratings of teaching: The research revisited* (IDEA Paper No. 32). Manhattan, KS: Center for Faculty Evaluation and Development.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51,* 281-309.

Feldman, K. A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 369-396). New York: Agathon Press.

Hogan, T. P. (2003). *Psychological testing: A practical introduction*. New York: Wiley.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*, 253-388.

Marsh, H. W., & Dunkin, M. J. (1997). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 241-320). New York: Agathon Press.

McKeachie, W. J., & Svinicki, M. (2006). *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers* (12th ed.). Boston: Houghton Mifflin.