Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 10 Number 1, April 2005 ISSN 1531-7714

A longitudinal approach to understanding course evaluations

Larry H. Ludlow, Boston College, Lynch School of Education
Department of Educational Research, Measurement and Evaluation

Tenure, promotion, and salary decisions are influenced by student ratings of instructors (SRIs) (McKeachie, 1979; Theall & Franklin, 1990; Yao, Weissinger & Grady, 2003). The literature addressing their validity, as a consequence, is voluminous (Marsh, 1987; Seldin, 1997; Silverman, 2001). From one institution of higher education to the next, however, the evaluation instruments take on different formats and address different aspects of teaching and teacher quality. And, from one institution to the next, the degree to which administrators incorporate this information into an overall judgment about teaching quality is often vague at best (Millman & Darling-Hammond, 1990; Scriven, 1995; Theall, Franklin & Ludlow, 1990; Wachtel, 1998).

In general, little systematic or in-depth analysis is performed by instructors upon the SRI reports they receive. Many faculty, perhaps most, look first at their overall rating—what percent of their students marked them "excellent" or "poor"? After that they tend to ignore the rest of the report and proceed directly to the student narratives, if any are provided.

Part of the lack of respect for SRIs has been attributed to a belief that students rate highly only those faculty who are easy graders and are personable—a "halo" effect (Alemoni, 1999). Another reason is that faculty typically receive little guidance on how to systematically analyze, interpret, and act upon their evaluations. For example, a typical institutional report compares an individual's ratings to an aggregated result—such as the combined undergraduate or graduate school results. This type of comparison, however, is often rejected by faculty as too confusing, confounding, irrelevant, and inappropriate. Furthermore, faculty are skeptical about SRIs since analyses of university-wide SRIs are often contradictory and ambiguous in terms of the extent to which teaching evaluations by students serve any useful function (Aleamoni, 1999; Wachtel, 1998).

Another common problem with using SRIs to assess teaching quality is that it is common practice to review the ratings in a relatively narrow time frame. For example, annual reviews typically consider just the two semesters in a given academic year. While this approach has some merit in terms of evaluating how students felt about an instructor's class during that period, it ignores the trajectory and pattern of evaluations for that instructor over the course of contiguous years. This means that significant contextual information about an individual is missing when teaching reviews ignore past performance and circumstances.

Frustration with these various problems provoked a line of research that focuses on the individual instructor. This research has established that a single-subject longitudinal analysis of course evaluations can be effective in revealing how different types of course related, instructor specific, and administrative operational variables are related to an instructor's ratings (Ludlow, 1996; Ludlow & Alvarez-Salvat, 2001; Ludlow, 2002). The emphasis is on an instructor's teaching over

time and it is in direct contrast to administrative snap-shot comparisons that offer little evidence and understanding of factors that contribute to systematic growth, maintenance, or deterioration in teaching. A key characteristic of this approach is that the analysis of one's SRIs is intended to inform and guide individuals—regardless of what the administration's aggregated summaries may suggest.

This paper presents a strategy for systematically analyzing one's student ratings of instruction. This strategy includes the types of questions and statistical tools appropriate for a single-subject design. In addition, the paper illustrates practical insights that may be extracted from one's longitudinal record of teaching. The purpose of this paper and line of research is, therefore, to make the analysis and interpretation of course evaluation summaries more useful and valuable through an analytic approach that can be implemented by any instructor.

DATA

The data file consists of 99 separate records summarizing the course evaluations received by one instructor for all classes taught from fall 1984 through fall 2003. The data were extracted from the end-of-semester SRI summaries compiled by the university. The specific course evaluation questions are fairly typical of those used by many institutions. For example, they asked students about their perceived need to attend class, the extent to which principles and concepts were understood, instructor promptness returning assignments, instructor enthusiasm, and instructor subject matter knowledge. The evaluation questions remained the same across the 20 years covered by this dataset.

The file is updated each fall with the SRI ratings from the previous fall, spring, and summer classes. Although the specific courses are largely irrelevant for the purposes of this paper, they range from entry-level freshman "Child Development" to a capstone third-year doctoral "Seminar in Statistical Methods." Most of the evaluations are for graduate courses in applied statistics. The data file consists of four relatively distinct categories of variables for each class taught: (a) administrative characteristics (e.g. year taught, class size, course code, level of students), (b) student-level perceptions (e.g. percent of time spent on the course, extent to which they acquired factual information), (c) instructor-specific variables (e.g. tenure status and marital status at the time the class was taught), and (d) overall evaluation ratings (percent who marked excellent, very good, good, acceptable, or poor). There are a total of 27 variables associated with each class. Overall, the dataset summarizes the evaluations submitted by 2174 students.

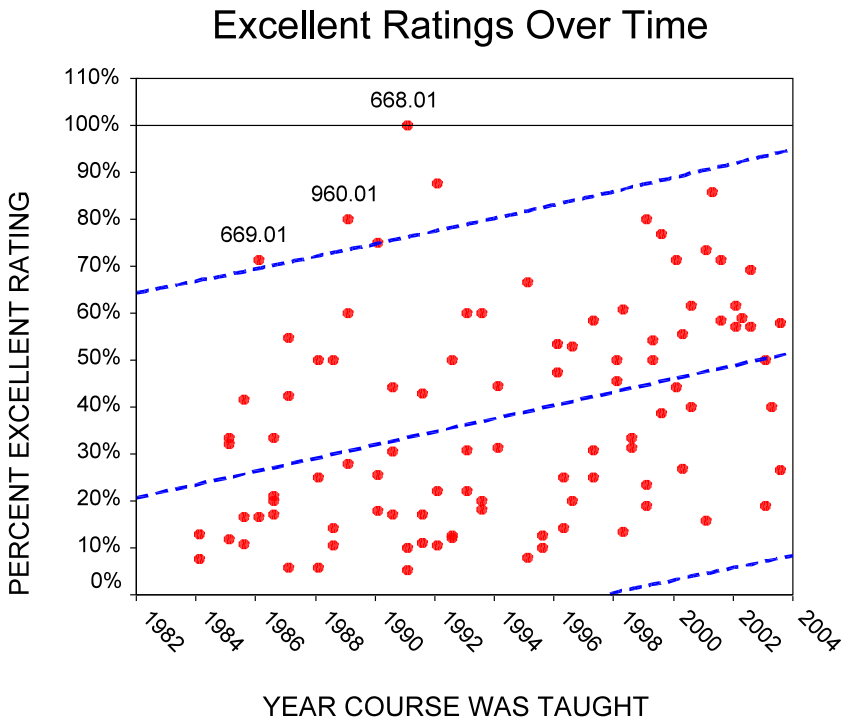
Ideally, one's institution provides the SRIs in an electronic format that is easily imported into a spreadsheet (EXCEL) or a statistical package (SPSS). In the event that SRI results are only returned in hardcopy, it is relatively simple to build a data file through hand-entering the results. In such a file each row of data corresponds to a separate class. Each column corresponds to a different aspect of the SRI summary results. An example of this type of file is illustrated in Ludlow (2002).

Once the file is created it is simple to add new information over time. This includes adding records for successive classes but it also includes adding new variables. For example, an indicator variable was added when the instructor was appointed department chair. This opportunity to add variables retrospectively makes it possible to test hypotheses about a wide variety of potentially influential variables, e.g. tenure, rank, or marital status at the time the class was taught.

ANALYSIS

One of the first things many faculty are interested in is: "What do my ratings look like over time?" Figure 1 shows the percent of students in each class who rated this instructor "excellent." Other outcomes such as the percent choosing "excellent + very good" could have been plotted but this particular instructor's interests are only on the variables that are useful in understanding the highest rating possible. The center dashed line is the regression line resulting from regressing "percent excellent ratings" on "year taught". It shows the predicted excellence rating for each class taught in any given year. Although the ratings show a positive upward trend over time across all classes, the graph does not differentiate the types of classes, e.g. undergraduate versus graduate.

Figure 1. How do the ratings look across all classes and 20 years?



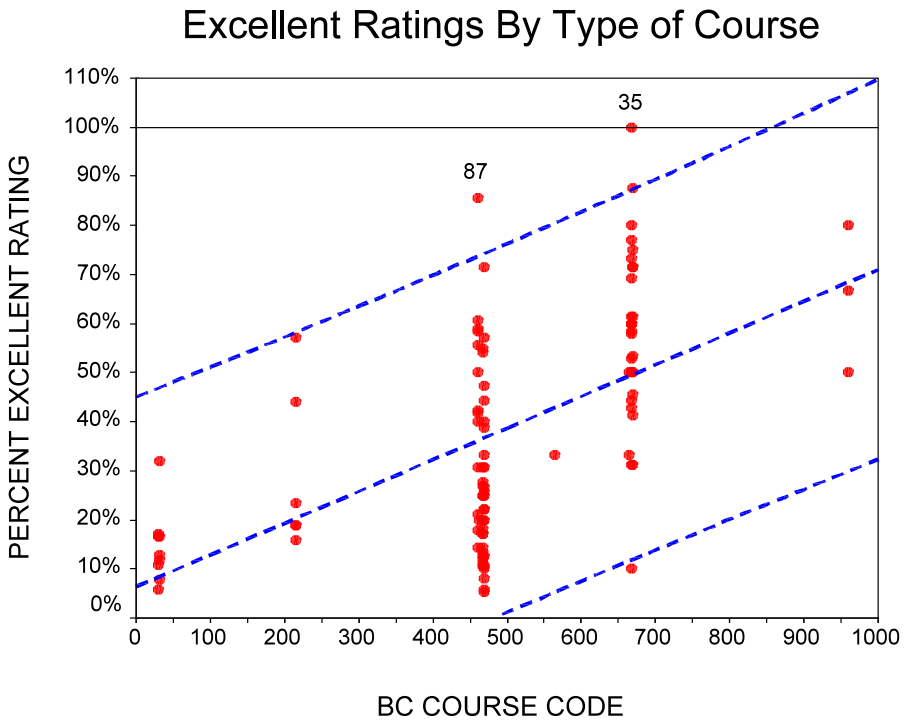
The other two dashed lines represent the 95% confidence interval around the regression line—the region within which most of the class ratings should lie (given this particular simple statistical model). There are four points corresponding to four classes that fall above the upper confidence interval. Those classes received excellence ratings that were much higher than expected. The class identifier reveals that three of these high ratings occurred the first time that particular course was taught by this instructor (indicated by the ".01" designation).

This is an interesting finding because faculty often believe that the first time they teach a course they work harder to get it "right" than for subsequent offerings. Hence, they believe ratings tend to be lower for first-time classes. Of course, if we wanted to know how the initial rating for a given course, say 669, compared to all other offerings of 669, we could easily generate a graph for just those classes. In fact, it can be very useful to plot the ratings over time for a single course that has been taught many times.

This is a useful graph for one's tenure and promotion portfolio. It is also useful when negotiating work area priorities and load considerations for the coming year and for documenting unexpectedly high ratings in the past year.

Continuing with this theme, what do the ratings look like for the different types of courses? Figure 2 plots the percent of excellence ratings for each class against the institutional course code assigned to that class. This analysis uses the course code as a proxy variable for the level of complexity of the material and sophistication of the student. The column of low ratings in the left region of the plot corresponds to courses 030 and 031—freshman child development courses taught early in the instructor's career. The next vertical column of slightly higher ratings corresponds to 216—undergraduate research methods. The next ratings correspond to 460, 468 and 469, required graduate research methods, introductory and intermediate statistics, respectively. The scale of this graph obscures the fact that the 460 ratings tend to be much higher than the 468 (which tend to be the lowest) and the 469 ratings. The differences in these three courses become evident when they are extracted and plotted in a separate analysis (not shown).

Figure 2. Do ratings differ by the type of course taught?

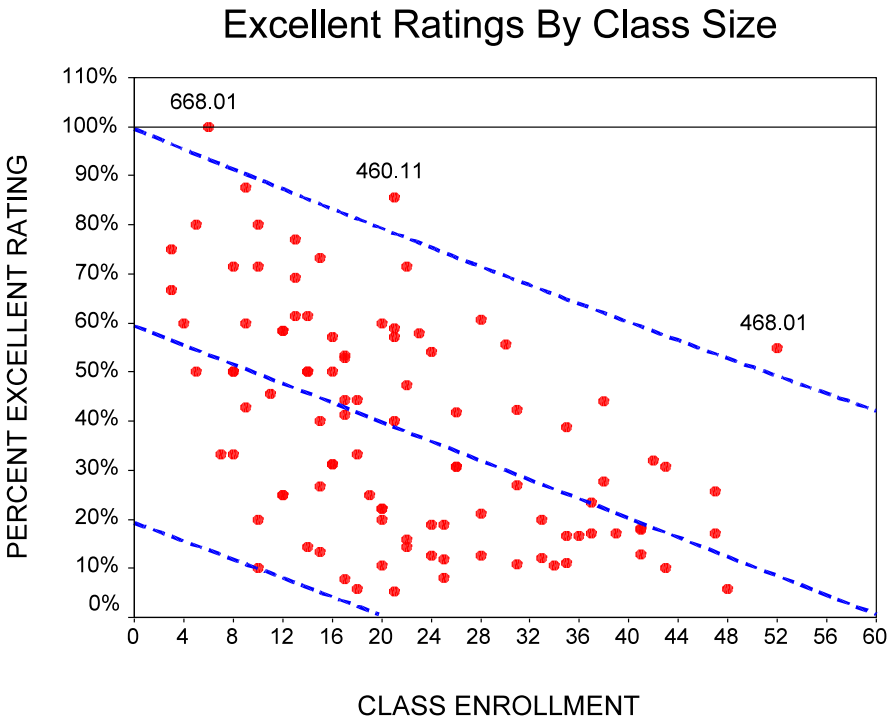


The vertical column containing the highest ratings corresponds to doctoral specialty courses (667, 668, 669—general linear models, multivariate statistics, psychometrics). The last column corresponds to a relatively non-technical seminar. Overall, there is a general upward trend in ratings as the course code increases—undergraduate courses have lower ratings than graduate courses, required statistics courses have lower ratings than the specialty courses.

This is an extremely useful type of analysis because it shows that ratings differ for the type and level of course taught. In particular, if an administrator must compare an individual's ratings against some aggregate, let the summary be constructed from similar relevant courses and students. Furthermore, two classes stand out as because of their unexpectedly high ratings: #87 is a research methods course (460.11), and #35 is a multivariate statistics course (668.01). It is a significant analytic point that these two classes will be identified as outliers in many of the following analyses.

One of the variables that faculty typically think has a negative effect on ratings is the size of the class. We generally think that we do better in small classes and, hence, receive better ratings than when we teach large classes. Figure 3 tests this hypothesis by plotting the excellence ratings by class size. There is a clear, unmistakable negative relationship between the ratings and class size—as enrollment increases, the ratings tend to drop. Statistically, for each additional student added to a class there is a decrease of about 1% in the excellence ratings.

Figure 3. Is there a relationship between class size and ratings?

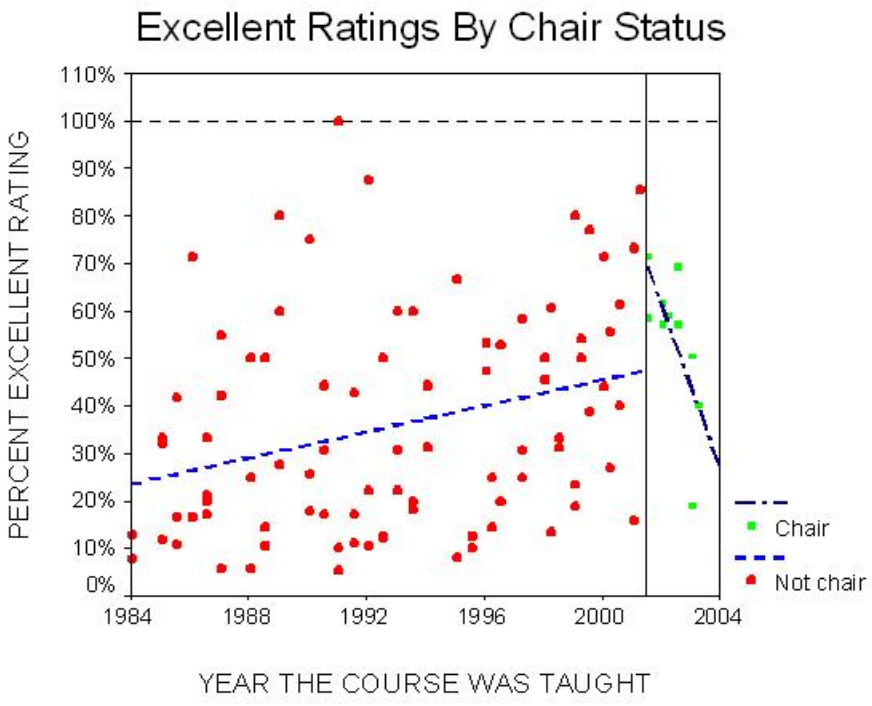


There are three classes with unexpectedly high ratings. The 668.01 class was described earlier. The 460.11 class was recently taught (for the 11th time) and its most striking characteristic was that it had an enrollment cap of 20 students in contrast to all previous classes which had had a cap of 30—all of which had lower ratings than this class. The particularly unusual class with the highest enrollment and relatively high rating is 468.01—this was the first time introductory statistics was taught by this instructor, it attracted a crowd of curious students, and its curriculum and format differently substantially from the way it had previously been taught (it changed from equation-based lectures to lectures followed by applications and statistical software instruction). This particular graph showing the relationship between class size and ratings has been used in numerous annual reviews to support arguments for reducing class sizes.

This instructor was appointed department chair in fall 2000. He quickly discovered that his new administrative duties were interfering with his class preparation, holding of office hours, and critiquing of assignments. He thought these problems might be reflected in his ratings.

Figure 4 shows the same ratings presented in Figure 1 but now the two periods of pre-chair and chair status are represented. It is apparent that the ratings in the two different periods reflect different trends. Although the overall trend seen in Figure 1 is positive, the ratings during the period as chair are dropping.

Figure 4. Is there a relationship between the ratings and administrative status?

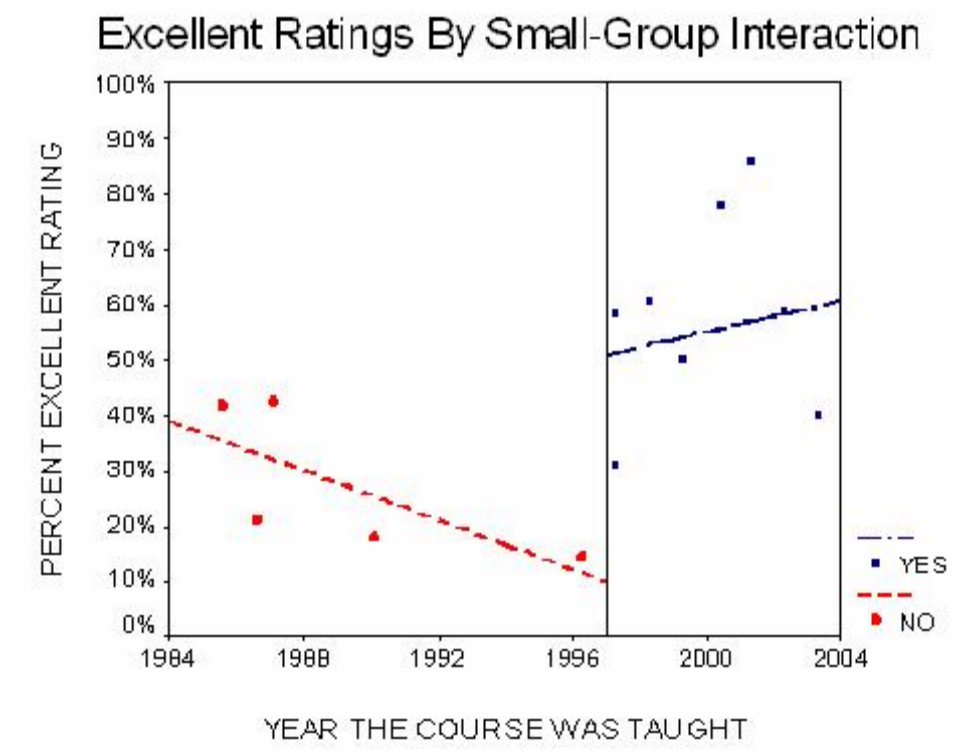


This type of event history graph is helpful when trying to understand the effects that critical experiences may have had, or are currently having, upon one's teaching effectiveness. Is there a drop in ratings from pre to post tenure? Is there a rise in ratings after a sabbatical or medical leave? Do ratings reflect personal changes such as marital status? By their very nature these types of self-reflective questions and analyses are unique for each individual.

These first four analyses looked at the relationship between the ratings and different types of administrative and structural variables. To a great extent these variables are not directly under the instructor's control. Although it is important and useful to understand how such variables are related to one's teaching practice as reflected in the ratings, faculty generally want to understand how their ratings reflect actual classroom practice and student experiences. For example, given that students often bring various forms of anxiety, even fear, into statistics courses, is there a relationship between how statistics is taught and the ratings or are the ratings simply overwhelmingly negative regardless of what the instructor does? The next three analyses illustrate how these questions may be investigated.

Positive educational effects that may be attributed to deliberate pedagogical change are often hard to detect and document. One way of detecting such an effect is illustrated in Figure 5. This instructor felt he was not adequately meeting the needs of students in his required graduate introduction to research methods course. The material was taught in a traditional lecture format that was boring him and apparently the students—as reflected by their low ratings. In 1997 (indicated by the vertical line) he changed the format to part lecture and part small-group interaction. The books, handouts, examples, and assignments essentially stayed the same. The primary change was a period of time during each class session that required students to interact with one another on practical exercises. The instructor wandered from group to group and served as a facilitator aiding and guiding their discussions.

Figure 5. What effect does a change in teaching practice produce?

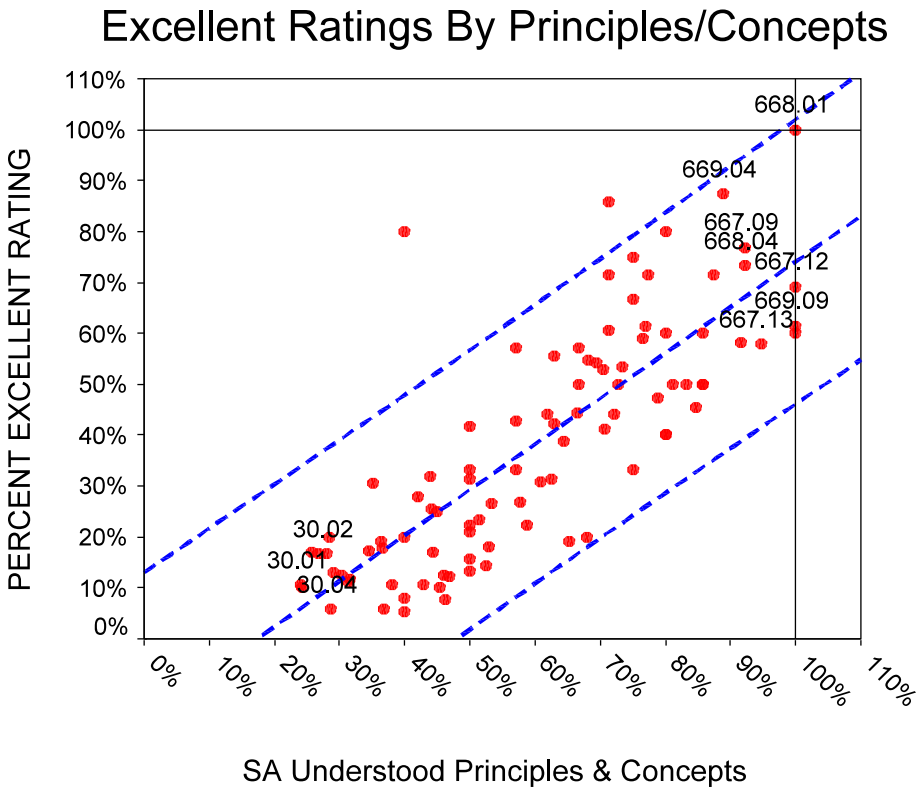


The effect of this relatively simple change is shown by the direction of the ratings before and after the change—the slope of the ratings quickly changed from negative to positive. In fact, this course is now one of the instructor's favorites. This particular analysis recently convinced the instructor to add small-group interactions and in-class exercises to his specialty classes. This type of analysis may be a useful tool as faculty debate the various pros and cons associated with moving from traditional chalk-and-talk formats to the various evolving point-and-click technology based formats.

Statistics faculty, like all faculty, must make choices about how a given topic is presented in class. For example, are equations presented as mathematical expressions to be memorized, is the emphasis placed on how to run and interpret statistical software, is the emphasis placed on linking the techniques in such a way that they are understood as logical ways to ask and answer increasingly sophisticated questions about one's data? The instructor's approach will influence which instructional questions on the evaluations are considered more relevant than others.

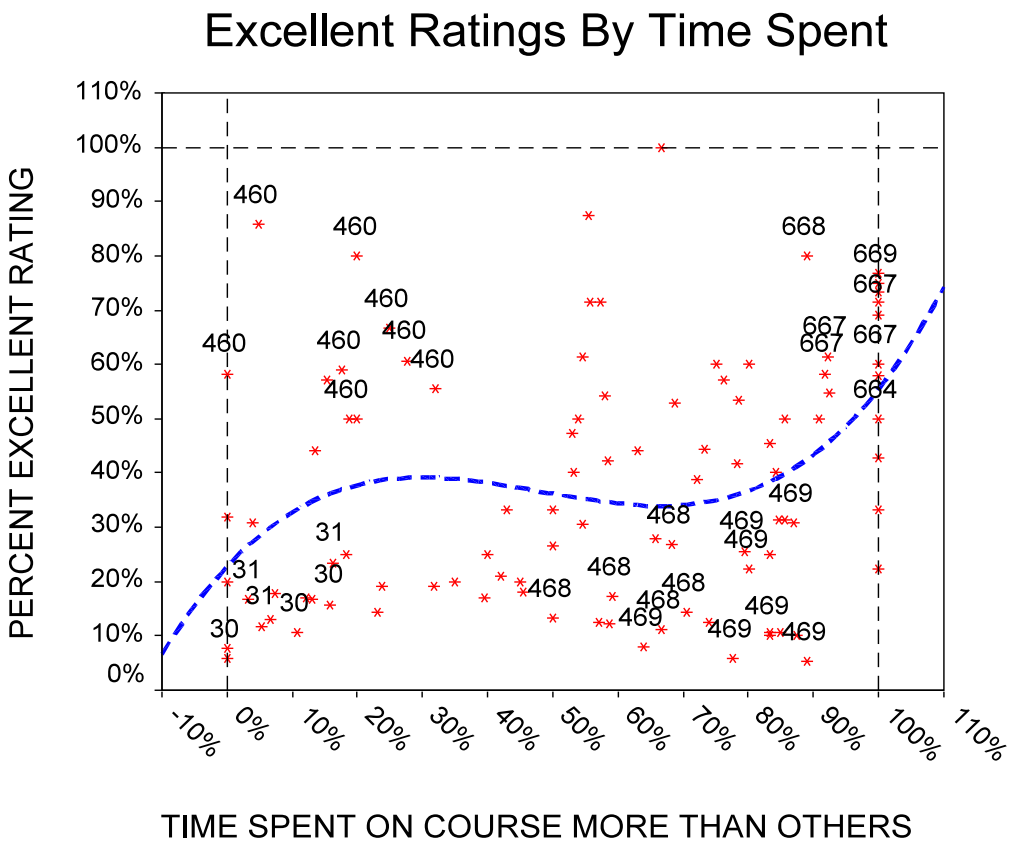
Figure 6 shows the percent excellent ratings plotted against the percent of students who strongly agreed that they understood principles and concepts. Note the extreme contrast between the 600-level courses (specialized statistics) with high principles and concepts ratings and the 30-level courses (child development) with low principles and concepts ratings. Given the analyses up to this point it is possible to state that excellence ratings increased over time, tended to occur in smaller classes in specialty courses, and were strongly related to the extent to which principles and concepts were emphasized and understood—not just whether factual information was presented (a different evaluation question).

Figure 6. What factors are controllable and how might they affect the ratings?



Faculty are often curious about the extent to which the workload required of students has an impact on the SRIs. While some faculty may believe that a heavy workload is desirable regardless of what the student thinks, others wonder if there is a negative relationship between the workload and ratings. Figure 7 show the ratings plotted against the percent of students who stated that they spent “much more” time on the present class than other classes that semester. Unlike the previous graphs, when this graph was first constructed it was immediately apparent that a simple linear trend was an inadequate representation of the relationship between these two variables. After a few exploratory attempts to find the best fitting line to these points, it was decided that this cubic relationship not only fit well but revealed a substantively interesting and useful pattern.

Figure 7. Is there a relationship between ratings and perceived workload?



There are four distinct clusters of courses in this graph. There are child development courses (30, 31) with trivial amounts of time commitments and the lowest excellence ratings. These were taught when the instructor first started teaching—and these were courses outside his training. There are the research methods classes (460) with slightly higher time requirements and some of the highest excellence ratings. These consist of frequent small-group interactions (as seen in Figure 5), are taught in the summer, and do not require extensive take-home assignments. There is a cluster of required statistics courses (468, 469) with a heavy time commitment and very low excellence ratings. These also define the mid-range cluster of points in Figure 6—students in these courses did not tend to strongly agree that they understood principles and concepts. Finally, there are the higher level specialty statistics courses (664, 667, 668, 669) with heavy time commitments and high excellence ratings. These courses tend to be the ones with small class sizes in Figure 3 and the ones in Figure 6 with the highest ratings on understanding principles and concepts. Apparently, heavy workloads and time commitments are valued by students if they understand why they are doing the work.

One of the key features of these graphs is that they show direction and magnitude of relationships, patterns over time, clusters of similar classes, and individual instances of surprising ratings. It is occasionally useful, however, to compute the simple correlations between pairs of variables. Table 1 contains a number of interesting relationships. This table suggests that higher ratings are related to (a) the extent to which students understood principles and concepts, (b) the extent to which students acquired factual information, (c) the extent to which the instructor was available outside of class, and (d) smaller classes. It is particularly interesting to see the relationship between class size and these other variables. The extent to which students understood principles and concepts, acquired factual information, and felt the instructor was available were all negatively correlated with class size. The larger classes not only had a negative relationship to the instructor's ratings, they had a negative relationship to the educational experience as perceived by the students.

Table 1: Pearson Correlations

Correlations

		PERCENT EXCELLENT RATING	Understood prin/concepts: SA	Acquired factual information: SA	Instrc. available outside class: SA	CLASS ENROLLMENT
PERCENT EXCELLENT RATING	Pearson Correlation	1	.807**	.719**	.562**	-.511**
	Sig. (1-tailed)	.	.000	.000	.000	.000
	N	97	92	92	92	97
Understood prin/concepts: SA	Pearson Correlation	.807**	1	.875**	.520**	-.599**
	Sig. (1-tailed)	.000	.	.000	.000	.000
	N	92	92	92	92	92
Acquired factual information: SA	Pearson Correlation	.719**	.875**	1	.457**	-.518**
	Sig. (1-tailed)	.000	.000	.	.000	.000
	N	92	92	92	92	92
Instrc. available outside class: SA	Pearson Correlation	.562**	.520**	.457**	1	-.527**
	Sig. (1-tailed)	.000	.000	.000	.	.000
	N	92	92	92	92	92
CLASS ENROLLMENT	Pearson Correlation	-.511**	-.599**	-.518**	-.527**	1
	Sig. (1-tailed)	.000	.000	.000	.000	.
	N	97	92	92	92	97

**. Correlation is significant at the 0.01 level (1-tailed).

It is possible to take the analysis of these particular variables one step further by considering how these variables might be used to predict specific course ratings. Table 2, for example, contains the results of a multiple regression using principles and concepts, factual information, availability outside the classroom, and class size as predictors of percent excellent ratings. The overall solution is statistically significant ($R^2_{adjusted}$ = .725, $p<.001$). Of greater interest, however, are the results for the individual predictors. Note that in this multiple regression (where the correlation between the predictor variables now is taken into account in estimating the effect that an individual variable has upon the outcome), the relationship between factual information and class size with the excellent ratings is no longer statistically significant. This is because factual information was highly correlated with principles and concepts and class size was highly correlated with all three other predictors—hence factual information and class size accounted for no statistically significant unique variance.

Table 2. A model for predicting excellent ratings.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.860[a]	.740	.725	12.22936

a. Predictors: (Constant), CLASS ENROLLMENT, Instrc. available outside class: SA, Acquired factual information: SA, Understood prin/concepts: SA

b. Dependent Variable: PERCENT EXCELLENT RATING

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-17.806	8.477		-2.101	.039
	Understood prin/concepts: SA	.731	.153	.650	4.775	.000
	Acquired factual information: SA	.068	.158	.054	.428	.670
	Instrc. available outside class: SA	.196	.077	.190	2.540	.013
	CLASS ENROLLMENT	-.033	.155	-.017	-.211	.833

a. Dependent Variable: PERCENT EXCELLENT RATING

Casewise Diagnostics[a]

Case Number	Std. Residual	PERCENT EXCELLENT RATING	Predicted Value	Residual
29	2.062	75.00	49.7787	25.2213
35	2.645	100.00	67.6562	32.3438
87	3.078	85.70	48.0602	37.6398

a. Dependent Variable: PERCENT EXCELLENT RATING

If we were to use this solution to predict the excellence rating for the next class taught the equation would look like:

> Predicted excellent rating = -17.8 + .73*(percent who strongly agreed they understood principles and concepts) + .196*(percent who strongly agreed the instructor was available outside class).

Although this particular regression solution is relatively arbitrary it does illustrate that it is possible to predict future class ratings based on past performance. More fully developed examples of this opportunity to predict ratings are presented in Ludlow (2002). Finally, the two classes that were not well fit by this solution were "Case Number" 35 (class 688.01 the multivariate class) and 87 (class 460.11 the research methods class).

The statistical point here is that a variety of relatively simple techniques have yielded consistent information that these two classes in particular were unusual. The pedagogical point of this exercise is that concentrating on making sure that principles and concepts are well understood by the students and making himself available outside the class are two variables that are directly under the instructor's control.

Of course, these statistical results are dependent on the specific variables chosen to construct the regression model where the model of choice would ideally be based on some theoretically grounded rationale—simply employing a stepwise regression is not a sound strategy for this type of investigation (Ludlow, 2002).

DISCUSSION

This paper presented a methodology for analyzing student ratings of instructors as a function of various student-level, instructor-specific, and administrative variables. The significance of this approach is that it enables an evaluation of an instructor in a longitudinal context, as opposed to a single year's snapshot of performance. This approach can yield valuable insight into the dynamics that may be operating within the professional career of an instructor.

Through the approach described in this paper, it has been possible to detect variables that influence one's teaching quality and effectiveness. For example, different statistics faculty have stated that they now: a) stress principles and concepts (instead of tedious calculations), b) form small-groups to facilitate interactions that reveal areas of confusion (instead of constant lecturing), c) incorporate real-world examples in all statistical applications (instead of artificial textbook examples), d) encourage email communications and hallway interactions outside the classroom (instead of sending everyone to the assistant), e) try to balance coursework between detailed thoroughness and unnecessary burden (instead of expecting every lecture point to be reflected in the assignments), and f) acknowledge to students when personal variables outside the classroom may affect day-to-day teaching effectiveness (instead of perpetuating the ivory-tower myth). These modifications were all prompted by patterns found through this approach to analyzing course evaluations.

In addition, this research has helped others prepare their teaching portfolio for promotion and tenure, and annual review considerations. Workshops, for example, have been conducted to show faculty that it is possible to extract useful, and sometimes unexpected, information from their course evaluations. These sessions have shown how to: a) extract meaningful data from their evaluation results, b) create an individualized database, and c) statistically analyze the ratings as a function of a variety of relevant professional variables.

The general workshop starts with a review of the course evaluation literature followed by examples of statistical analyses of course evaluations. The fundamental characteristics of the SPSS statistical package system are then explained. Next they are shown how to build a data file and how to run SPSS graphing and statistical analysis procedures. This step is extremely useful because it provides them with a simple set of tools for understanding and evaluating their SRIs in terms of their own unique situations.

The statistical methods employed in this paper are not restricted to the specifics of this particular study—they are generalizable in the sense that they may be appropriate for any other set of longitudinal SRI data. In particular this research is intended to more fully empower faculty in the analysis and interpretation of their own course evaluations. For example, one of the secondary consequences of this work has been that some faculty, who previously felt threatened by statistical questions and analyses, now feel sufficiently confident and competent to attempt creative analyses on their own. Ultimately, this type of longitudinal, single-subject research should contribute to a wider theory about how course evaluations may be effectively utilized to understand and improve teaching.

REFERENCES

Aleamoni, L.M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education. 13*(2), 153-166.

Ludlow, L.H. (2002). Rethinking practice: Using faculty evaluations to teach statistics *Journal of Statistics Education, 10(3).* www.amstat.org/publications/jse/v10n3/ludlow.html.

Ludlow, L.H. & Alvarez-Salvat, R. (2001). Spillover in the academy: Marriage stability and faculty evaluations. *Journal of Personnel Evaluation in Education, 15:2*, 111-119.

Ludlow, L.H. (1996). Instructor evaluation ratings: A longitudinal analysis. *Journal of Personnel Evaluation in Education, 10*, 83-92.

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

McKeachie, W.J. (1979). Student ratings of faculty; A reprise. *Academe: Bulletin of the AAUP, 65(6)*, 384-397.

Millman, J. & Darling-Hammond, L. (Editors) (1990). *The new handbook of teacher evaluation*. SAGE: Newbury Park.

Scriven, M. (1995). Student ratings offer useful input to teacher evaluations. *Practical Assessment, Research &*

Evaluation, 4(7). http://pareonline.net/getvn.asp?v=4&n=7.

Seldin, P. (1997). *The Teaching Portfolio*. Bolton, MA: Anker Publishing.

Silverman, F. H. (2001). *Teaching for Tenure and Beyond: Strategies for Maximizing Your Student Ratings*. Westport, CT: Bergin & Garvey.

Theall, M. & Franklin, J. (1990) (Editors). *Student Ratings of Instruction: Issues for Improving Practice*. San Francisco, CA: Jossey-Bass.

Theall, M., Franklin, J., & Ludlow, L.H. (1990). Attributions and retributions: Student ratings and the perceived causes of performance. *Instructional Evaluation, 11,* 12-17.

Wachtel, H.K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment and Evaluation in Higher Education, 23*, 191-210.

Yao, Y., Weissinger, E. & Grady, M. (2003). Faculty use of student evaluation feedback. *Practical Assessment, Research & Evaluation, 8(21).* http://pareonline.net/getvn.asp?v=8&n=21.

Acknowledgment:

This research was partially funded by a Boston College Teaching, Advising, and Mentoring grant.

Citation

Ludlow, L.H. (2005). A longitudinal approach to understanding course evaluations. *Practical Assessment Research & Evaluation*, 10(1). Available online: http://pareonline.net/getvn.asp/v=10&n=1