RIDIT ANALYSIS OF STUDENT COURSE EVALUATIONS

Dean Croushore
Professor of Economics and Rigsby Fellow
University of Richmond

Robert M. Schmidt
Professor of Economics
University of Richmond

June 2010

We thank participants at the 2009 Academy of Business Education for their comments.

Please send comments to the authors at Robins School of Business, 1 Gateway Road, University of Richmond, VA   23173, or e-mail: dcrousho@richmond.edu or rschmidt@richmond.edu.

RIDIT ANALYSIS OF STUDENT COURSE EVALUATIONS

ABSTRACT

Most analyses of student evaluations of courses and instructors are based on averaging the point values of the questions about course and instructor quality. But using the average assumes that the answers to the questions are uniform on an interval even though the categories may not be equally spaced in the minds of those filling out the forms. A better statistical method treats the responses as ordinal data, not equally spaced. For this purpose, we can use RIDIT analysis, which is analysis that compares the responses to a questionnaire relative to an identified distribution.

In this paper, we use RIDIT analysis on 20 years of data from the Robins School of Business at the University of Richmond. We show how the evaluation based on RIDIT analysis can lead to different conclusions than when using the arithmetic average. Statistical inference about significant differences in instructor quality can be made with a sounder theoretical basis.

Based on RIDIT analysis, we are able to compare performance of a variety of subgroups, as well as for individual courses and instructors. We first establish a baseline control RIDIT, then analyze how the control has changed over time and how it depends, for example, on the course level, the time of day of the course, and whether the course is a required course or an elective. We then compare these results to those that would be obtained with a more typical analysis based on the means of the evaluations assuming equally spaced intervals for the answers on the student-evaluation questionnaire.

RIDIT ANALYSIS OF STUDENT COURSE EVALUATIONS

I. Introduction

Questions on student evaluation forms of teaching effectiveness are qualitative or, at best, ordinal in nature. The questions ask students about the class and the instructor in terms of categories. Yet most analysis of student evaluations treats the answers as if they were interval, being placed into bins that are equally spaced apart. But if the differences between the intervals in the questions are larger or smaller in students' minds, then the quantitative analysis, such as evaluating the mean response to a question, may misrepresent students' views. This goal of this paper is to present the use of RIDIT analysis for evaluating student teaching, allowing for a more accurate representation of the results.

Consider the following example. A course evaluation form contains the statement: "I have learned a lot as a result of this course" and asks students to state whether they "strongly agree," "agree," "neither agree nor disagree," "disagree," or "strongly disagree." Typically, we assign numbers to each of the answers:

1 = strongly disagree

2 = disagree

3 = neither agree nor disagree

4 = agree

5 = strongly agree

Then we look at the number of responses in each of these categories and take the mean based on the assigned values of 1 to 5. Then we might compare a faculty member's result, say a 3.8, to the department's average, perhaps 3.6, to determine merit pay for teaching quality.

But what if, in students' minds, the true representation of teaching quality is very different. For example, suppose students think that a more appropriate scale is:

–5 = strongly disagree

0 = disagree

3 = neither agree nor disagree

10 = agree

20 = strongly agree

In this case, then mean could be very different. Such a scale would show much greater differences between strong teachers and below-average teachers.

Because eliciting a true scale from students would be difficult, it seems better to develop a metric that can be used without assigning values to each answer. The question is, how can we do so while still being able to test whether there are statistically significant differences between instructors? Or, if we want to test whether course evaluations for required classes differ significantly from those for elective classes, how can we do so?

Statistical theory provides one possible solution—a method known as RIDIT analysis. RIDIT stands for "Relative to an Identified Distribution." The concept is that there is a control group and a treatment group, and a statistical test investigates whether or not the treatment group has a distribution that is significantly different from the control group. This paper will show how RIDIT analysis can be used to analyze course evaluations, comparing the results to the more typical results from assigning an evenly spaced scale and comparing means from such a scale. In section II, we discuss the literature on course evaluations and how they have been analyzed historically. In section III, we describe RIDIT analysis and show why, in principle, it may provide a better framework for analysis than previously used methods. In section IV, we describe the data from the Robins School of Business at the University of Richmond, which covers 20 years. Section V presents our main results, testing a variety of hypotheses and showing how RIDIT analysis compares with the usual analysis that has been performed in the literature. Section VI provides conclusions and suggestions for additional research.

II. Literature on course evaluations

A substantial literature analyzes the factors that affect course evaluation results. Much of the research suggests that there are many problems with using student evaluations of classes to evaluate the quality of teaching. Typical of the literature is Dooris (1997), who argues that teaching evaluations represent a popularity contest. Emery (1995) showed that higher evaluations are achieved when teachers bring food to class. Abrami et al. (1982) find that charismatic and enthusiastic faculty get high ratings, even when they teach poor material; they also find that charisma and enthusiasm are unrelated to how much students learn. Feldman (1986) also finds a strong correlation between teacher personality and high ratings.

In the same vein, many studies have found little correlation between student achievement and evaluations. Cohen (1983) finds that student achievement accounts for only 14.4% of the rating variance across instructors. Damron (1996) finds that most factors contributing to high course evaluations are unrelated a teacher's ability to promote learning.

However, many studies find support for the use of student evaluations. Cashin (1995) examines many of the major studies of student evaluations of teaching and finds that "In general, student ratings tend to be statistically reliable, valid, and relatively free of bias . . ." (p. 6). Boex (2000) argues that "Despite anecdotal evidence to the contrary, much of the relevant literature concludes that student evaluations of teaching are generally consistent and valid." (p. 215) However, much research suggests that certain factors (including instructor popularity, expected grade, and difficulty of material) affect student evaluations, so these factors should be considered in using the student evaluation results.

Given that student evaluations of teaching are performed nearly everywhere, and administrators tend to use such evaluations for tenure, promotion, and merit pay increases, proper statistical analysis of them is of paramount importance.

In addition to the complaints that debate whether course evaluations are or are not useful in evaluating teaching quality, there is also a substantial literature on what affects course evaluations in terms of characteristics of the course itself, rather than the instructor. Cashin (1990) finds that course evaluations vary across disciplines, making cross-department comparisons difficult. Aleamoni (1989) finds that evaluations are lower for required courses than elective courses and the evaluations are lower for lower-level courses than for upper-level courses. Other papers in this literature look at the effects of class size, the sex of the student compared with the instructor, the time of day of the class, the grade-point average of the students in the class, and the students' reasons for taking the class.

All of this literature uses either factor analysis or is based on the assumption of equally spaced intervals for the answers to the course evaluation questions. We would like to investigate whether a better statistical method, RIDIT analysis, leads to different answers to these questions.

III. RIDIT Analysis

To perform RIDIT analysis, you first split the sample into two groups: a control group and a treatment group; see Fleiss (1976), pp. 102-108.[1] From the control group data, you calculate "control ridits," based on the frequency response to each question. From the treatment group data, you calculate the mean ridit and the standard error. Then a z-score determines the statistical significance of the difference between the control group and the treatment group.

Control ridits are calculated from the control group. Control ridits essentially describe the distribution of the control group's responses to each question. Follow these steps:

1. For each category (answer), calculate ½ the frequency (to determine the number of responses in that category up to the mid-point.

[1] RIDIT analysis was introduced by Bross (1958), with key contributions by Snell (1964) and Kantor, et al. (1968).

2. Calculate the cumulative frequency for all previous categories.
3. Add the results of steps 1 and 2 for each category.
4. Divide the results of step 3 for each category by the sample size of the control group.

A useful equation for calculating the control ridits can be developed as follows: If there are *c* categories, number them 1 to *C* from lowest quality to highest quality. Let the frequency in each category of the control group be F_c, where *c* = 1, 2, . . ., *C*. The total number of observations is: $N = \sum_{c=1}^{C} F_c$. The control ridit for each category is equal to:

$$R_c = (\sum_{i=1}^{c-1} F_i + \frac{F_c}{2})/N, \qquad (1)$$

for *c* = 1, 2, . . ., *C*.

These steps are not intuitive, so Table 1 presents an example, in which the frequencies are given and all the other elements of the table are constructed from the steps listed above.

Table 1: Example of Calculation of Control Ridits

Category (*c*):	(1) Strongly Disagree	(2) Disagree	(3) Neither	(4) Agree	(5) Strongly Agree	*N*
Frequency (F_c)	3	6	6	4	8	27
$F_c/2$	1.5	3	3	2	4	
$\sum_{i=1}^{c-1} F_i$	0	3	9	15	19	
$\sum_{i=1}^{c-1} F_i + F_c/2$	1.5	6	12	17	23	
$R_c = \frac{\sum_{i=1}^{c-1} F_i + \frac{F_c}{2}}{N}$	0.056	0.222	0.444	0.630	0.852	

The control ridits of 0.056 (strongly disagree), 0.222 (disagree), 0.444 (neither), 0.630 (agree), and 0.852 (strongly agree), show the proportion of all observations in the control group with an underlying value at or below the mid-point of the category. You can think of the categories of the answers as having the number of responses spread out evenly across a continuous scale. For example, the ridit for the "strongly agree" category is 0.852, so 85.2% of the observations in the control group have an underlying value at or below the mid-point of the category.

After the control ridits have been calculated, you then calculate the mean ridit of the treatment group, which is like a mean, but in comparison to the control group. To calculate the mean ridit, you multiply the frequency in each category by the control ridit for that category. Add up those products and divide by the number of observations in the treatment group to get the mean ridit.

Let the frequency in each category of the treatment group be F_t, where *t* = 1, 2, . . ., *C*. The total number of observations is: $n = \sum_{t=1}^{C} F_t$. The mean ridit is equal to:

$$\bar{r} = \frac{1}{n}\sum_{t=1}^{C}(F_t \times R_t). \qquad (2)$$

The standard error of the mean ridit is approximately equal to:

$$s.e.(\bar{r}) = \frac{1}{2\sqrt{3n}}$$

These calculations are illustrated for our example in Table 2.

Table 2: Example of Calculation of Mean Ridit

Category (*t*):	(1) Strongly Disagree	(2) Disagree	(3) Neither	(4) Agree	(5) Strongly Agree	*N*
Frequency $(\boldsymbol{F_t})$	5	8	6	2	6	27
Control ridit (R_t)	0.056	0.222	0.444	0.630	0.852	
$F_t \times R_t$	0.278	1.778	2.667	1.259	5.111	
$\boldsymbol{\bar{r} = \frac{1}{n}\sum_{t=1}^{C}(F_t \times R_t)}$						0.411
$\boldsymbol{s.e.(\bar{r}) = \frac{1}{2\sqrt{3n}}}$						0.056

The interpretation of the mean ridit is that it represents the probability that a randomly selected student from the treatment group will have a higher value for the question than a randomly selected student from the control group. So, if the question is about how much a student learned in a particular class (treatment), this is the probability that the student learned more in that class than in the control group (other classes).

In the example in Tables 1 and 2, the mean ridit of 0.411 means that a randomly selected student from this class has a 41.1% chance to have learned more in this class (the treatment class) than in the control class; or a 58.9% chance of having learned less.

If the treatment group has frequencies that are proportional to the frequencies of the control group, then the mean ridit will be exactly 0.50. Suppose $F_t = \alpha F_c$, so that $n = \alpha N$. From equation (2), $\bar{r} = \frac{1}{n}\sum_{t=1}^{C}(F_t \times R_t)$, and now substitute in the proportional relationships and equation (1) to get:

$\bar{r} = \frac{1}{n}\sum_{t=1}^{C}(F_t \times R_t) = \frac{1}{n}\sum_{c=1}^{C}(\alpha F_c \times R_c) \ = \frac{\alpha}{n}\sum_{c=1}^{C}[\sum_{i=1}^{c-1}(F_i + \frac{F_c}{2})/N]\, F_c$

$= \frac{1}{N^2}[\frac{1}{2}(F_1^2 + F_2^2 + F_3^2 + F_4^2 + F_5^2) + (F_1 F_2) + (F_1{+}F_2)F_3 + (F_1{+}F_2{+}F_3)F_4$

$+(F_1 + F_2 + F_3 + F_4)F_5]$

$= \frac{1}{2}\frac{1}{N^2}(F_1 + F_2 + F_3 + F_4 + F_5)^2 \ = \frac{1}{2}\frac{1}{N^2}N^2$

= ½.

The statistical significance of differences in mean ridits can be evaluated with a Z-test. The null hypothesis is that the mean ridit equals 0.5. The z-score is:

$$z = \frac{\bar{r} - 0.5}{s.e.(\bar{r})}$$

In our example, with $\bar{r} = 0.411$ and $s.e.(\bar{r})$ = 0.056, *z* = -1.60, which is not larger in absolute value than the critical value of 1.96, so this instructor for this question is not significantly worse than the control group.

Two treatment groups can be compared in one step, and the significance of differences between them can be evaluated directly. Suppose instructors A and B have mean ridits of 0.70 and 0.42 (respectively) against a relevant control group. The statistical significance of the difference in the mean ridits can be judged by the statistic:

$$z = \frac{\bar{r}(B) - \bar{r}(A)}{s.e.[\bar{r}(B) - \bar{r}(A)]}$$

where

$$s.e.[\bar{r}(B) - \bar{r}(A) = \frac{\sqrt{n_A + n_B}}{2\sqrt{3n_A n_B}}]$$

In our example, $z = \frac{\bar{r}(B)-\bar{r}(A)}{s.e.[\bar{r}(B)-\bar{r}(A)]} = \frac{0.42-0.70}{0.0619} = -4.53$, showing a significantly worse performance by professor B than professor A, relative to the control group.

IV. Data

The data that we use in this study come from the University of Richmond, Robins School of Business. Data cover the period 1990 to 2009, including course evaluations from every undergraduate course taught in the business school in fall and spring semesters. There is one record per course. Each data record includes information on the term, enrollment, number of responses, frequency of responses in each category, the course level (100, 200, 300 or 400), a code for each instructor, whether the instructor was tenure stream or not, whether the course was required or not, whether the class was early (before 9), or late (after 2), whether the class period was 50-minutes 3 times a week or not, the proportion of the class that is the same gender as the instructor, the class GPA, and the expected grade in the class.

In this paper, we use a subset of the data from economics courses. We focus on two questions: (1) Question 23: “I learned a lot as a result of this course”; (2) Question 24: “This instructor’s overall teaching ability is excellent”. The answers to both are in the categories: strongly disagree, disagree, neither agree nor disagree, agree, and strongly agree.

V. Results

1. Changes in Evaluations over Time

As time passes, a college’s faculty may become better (or worse) at teaching, which would lead to changes in the distribution of the course evaluations. To test if the evaluations now seem to come from a different distribution than before, we consider a simple experiment: use Fall 1990 as the control group, and look at Fall 1999 and Fall 2008 as treatment groups. Doing so gives the results shown in the first two rows of Table 3. With Fall 1999 as the treatment group, we find a mean ridit for question 23 (learned a lot) of 0.480, with a p-value from the z-score of 0.0857, so we do not reject the null hypothesis that the evaluation in Fall 1999 comes from the same distribution as that in Fall 1990.[2] However, for question 24 (overall teaching ability), the mean ridit is 0.543, with a p-value of 0.003, suggesting that the distribution changed between 1990 and 1999 towards better overall teaching ability. When we use Fall 2008 as the treatment group, we get a mean ridit of 0.637 for question 23 and 0.627 for question 24, both significant improvements over Fall 1990.

Table 3. Hypothesis Tests

		Q23: Learned a lot		Q24: Overall teaching ability	
Control	**Treatment**	**Mean Ridit**	**p-value**	**Mean Ridit**	**p-value**
Fall 1990	Fall 1999	0.480	0.086	0.543	0.003
Fall 1990	Fall 2008	0.637	0.000	0.627	0.000
1990s	2000s	0.600	0.000	0.566	0.000
Elective	Required	0.533	0.000	0.516	0.000
Tenure-Stream	Adjuncts & Visitors	0.445	0.000	0.420	0.000
Prime Time	Early (<9 am)	0.489	0.061	0.484	0.005
Prime Time	Late (> 2 pm)	0.557	0.000	0.559	0.000

[2] We use a 5% significance level throughout this paper.

The previous tests just picked a single semester to examine if teaching quality had improved. It might be more useful to compare longer time spans, in case the choice of a particular semester was an outlier. If we use the entire 1990s as the control and the 2000s as the treatment (results in the third row of Table 3), we get a mean ridit of 0.600 for question 23 and 0.566 for question 24, showing significant improvement in the 2000s compared with the 1990s.

2. Required versus Elective Courses

Faculty members often complain that when they are asked to teach required courses, they get lower teaching evaluations. We can test that hypothesis for our sample. Using all elective classes as the control group and required courses as the treatment group, we find that the mean ridit is above 0.5, suggesting that evaluations are actually higher in required courses. For question 23, the mean ridit is 0.533, and for question 24 it is 0.516, and both have small p-values that show a statistically significant difference between required and elective courses. So, our faculty should be happy when they are asked to teach required courses instead of electives.

3. Quality of Adjuncts and Visitors

In recent years, our department has worried that the teaching quality of our adjuncts and visiting faculty is significantly worse than the quality of our tenure-stream faculty. Using tenure-stream faculty as the control group, the mean ridit for non-tenure-stream faculty is 0.445 for question 23 and 0.420 for question 24, both statistically significant. So, our suspicion that the teaching quality of our adjuncts and visitors is well below that of our tenure-stream faculty is borne out in the data.

4. Teaching Early or Late

Some faculty members like to teach 8 a.m. classes, believing that they get better students that way. Others like teaching later in the day for the same reason. We can test those hypotheses about the time of day of classes by using prime-time classes (those taught between 9 a.m. and 2 p.m.) as the control group. For early classes, the mean ridit is 0.489 for question 23, not statistically significant; the mean ridit is 0.484 for question 24, and it is statistically significant.

For late-day classes, the mean ridits are above 0.5 (0.557 for question 23 and 0.559 for question 24) and are statistically significant. So, it appears that early classes do not lead to better evaluations, but late classes do.

Of course, all of these results could be subject to qualification because of sample selection. It could be that many good teachers prefer to teach in the afternoon, and worse teachers like teaching at 8 a.m. to minimize their class sizes, so self-selection could be the main cause of the difference. Or, it could be that if athletes and worse students and give worse teaching evaluations, and if they must take early classes because of afternoon athletic practice, then the evaluation results will be biased.

VI. Comparison with Means of Evenly Spaced Categories

One key question is: does RIDIT analysis tell us anything different than the more-typical analysis of means, assuming that the categories are evenly spaced. If we compare the mean ridit with the difference between the mean of the control group (under the typical assumption of assigning points to each category of 1 = strongly disagree, 2 = disagree, 3 = neither, 4 = agree, and 5 = strongly agree, then taking the mean) and the mean of the treatment group, we find that the difference in means suggests the same general results as the RIDIT analysis (Table 4). The question is whether or not we would also find a difference in statistical significance. Using a t-test for the difference in means and calculating the p-value based on that test, we can see that in our experiments, the p-values based on RIDIT analysis and the difference-in-means specification are very similar. If we run more experiments with RIDIT p-values near 0.05, we might find that there are marginal differences between the p-values based on RIDIT analysis and difference-in-means specifications with equally spaced categories. Or, it could be that the students in the sample think of the various categories as being evenly spaced, so that the two alternative approaches show no major differences.

Table 4. Comparison of RIDITs to Traditional Means

Q23: Learned a lot					
Control	**Treatment**	**Mean Ridit**	**p-value**	**Difference in Means**	**p-value**
Fall 1990	Fall 1999	0.480	0.086	-0.092	0.084
Fall 1990	Fall 2008	0.637	0.000	0.418	0.000
1990s	2000s	0.600	0.000	0.323	0.000
Elective	Required	0.533	0.000	0.102	0.000
Tenure-Stream	Adjuncts & Visitors	0.445	0.000	-0.177	0.000
Prime Time	Early (<9 am)	0.489	0.061	-0.036	0.080
Prime Time	Late (> 2 pm)	0.557	0.000	0.181	0.000

CONCLUSIONS

RIDIT analysis is a useful way to analyze and compare data on student evaluations of teaching, allowing tests of statistical significance. The RIDIT analysis showed improvement in course evaluations over time in our data set, found that required courses had somewhat higher ratings, that adjuncts and visitors have significantly worse ratings than tenure-stream faculty, that 8 a.m. classes do not have significantly worse ratings, and that classes after 2 pm have significantly better ratings.

What we don’t know is the extent to which self selection should cause us to doubt these results. It could be that we put our best teachers in required courses, resulting in higher ratings. Also plausibly, our best teachers like teaching in the late afternoon instead of the early morning. This suggests a need for a multivariate analysis to hold other factors constant, which will be the subject of our next paper on this topic.

REFERENCES

Abrami, P.C., L. Leventhal, and R.P. Perry, R.P. 1982. Educational seduction. *Review of Educational Research* 32: 446-64.

Aigner, D. J., and F. D. Thum. 1986. On student evaluation of teaching ability. *Journal of Economic Education* 17 (4): 243-66.

Aleamoni, L. 1989. Typical faculty concerns about evaluation of teaching. In Aleamoni, L., ed., *Techniques for Evaluating and Improving Instruction*, Jossey-Bass, San Francisco.

Becker, W. E., and P. Kennedy. 1992. A graphical exposition of the ordered probit. *Econometric Theory* 8 (1): 127-31.

Becker, W. E., and M. Watts. 1999. How departments of economics evaluate teaching. *American Economic Review* 89 (2): 344-49.

Boex, J. L. F. 2000. Attributes of effective economics instructors: An analysis of student evaluations. *Journal of Economic Education* 31 (3): 211-27.

Bosshardt, W., and M. Watts. 2001. Comparing student and instructor evaluations of teaching. *Journal of Economic Education* 32 (1): 3-17.

Bross, I.D.J. 1958. How to use ridit analysis. *Biometrics* 14: 18-38.

Cashin, William E. 1990. Students do rate different academic fields differently. In Theall, M., and Franklin J., eds., *Student Ratings Of Instruction: Issues For Improving Practice*, Jossey-Bass, San Francisco.

Cashin, William E. 1995. “Student Ratings of Teaching: The Research Revisited.” IDEA paper 32, Kansas State University, Center for Faculty Evaluation and Development.

Cohen, P.A. 1983. Comment on a selective review of the validity of student ratings of teaching. *Journal of Higher Education* 54: 448-58.

Damron, J.C. 1996. Instructor personality and the politics of the classroom. www.mankato.msus.edu/dept/psych/Damron_politics.html.

Dooris, M.J. 1997. An analysis of the Penn State student rating of teaching effectiveness: a report presented to the University Faculty Senate of the Pennsylvania State University. www.psu.edu/president/cqi/cqi/srte/analysis.html.

Emery, Charles R. 1995. Student evaluations of faculty performance. Clemson University, Clemson, SC., manuscript.

Emery, Charles R., Tracy R. Kramer, and Robert G. Tian. 2003. Return to academic standards: a critique of student evaluations of teaching effectiveness. *Quality Assurance in Education* 11 (1): 37-46.

Everett, M. D. 1977. Student evaluations of teaching and the cognitive level of economics courses. *Journal of Economic Education* 8 (2): 100-03.

Feldman, K.A. 1986. The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: a review and synthesis. *Research in Higher Education* 24: 139-213.

Fleiss, Joseph L. *Statistical Methods for Rates and Proportions* (New York: Wiley, 1973).

Kanor, Winkelstein, and Ibrahim. 1968. A note on the interpretation of the ridit as a quantile rank. *American Journal of Epidemiology* 87: 609-615.

Kelley, A. C. 1972. Uses and abuses of course evaluations as measures of educational output. *Journal of Economic Education* 4 (1): 13-18.

Seiver, D. A. 1983. Evaluations and grades: A simultaneous framework. *Journal of Economic Education* 14 (3): 32-38.

Snell, E.J. 1964. A scaling procedure for ordered categorical data. *Biometrics* 20: 592-607.

Stratton, R. W., S. C. Myers, and R. H. King. 1994. Faculty behavior, grades, and student evaluations. *Journal of Economic Education* 25 (1): 5-15.

Wilson, Robin. 1998. New research casts doubt on value of student evaluations of professors. *Chronicle of Higher Education* 16 (January 1998), p. A12.