Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.



Volume 16, Number 2, February, 2011 ISSN 1531-7714

# Do more online instructional ratings lead to better prediction of instructor quality?

Shane Sanders, *Western Illinois University*
Bhavneet Walia, *Western Illinois University*
Joel Potter, *North Georgia College and State University*
Kenneth W Linna, *Auburn University Montgomery*

Online instructional ratings are taken by many with a grain of salt. This study analyzes the ability of said ratings to estimate the official (university-administered) instructional ratings of the same respective university instructors. Given self-selection among raters, we further test whether more online ratings of instructors lead to better prediction of official ratings in terms of both R-squared value and root mean squared error. We lastly test and correct for heteroskedastic error terms in the regression analysis to allow for the first robust estimations on the topic. Despite having a starkly different distribution of values, online ratings explain much of the variation in official ratings. This conclusion strengthens, and root mean squared error typically falls, as one considers regression subsets over which instructors have a larger number of online ratings. Though (public) online ratings do not mimic the results of (semi-private) official ratings, they provide a reliable source of information for predicting official ratings. There is strong evidence that this reliability increases in online rating usage.

Teaching quality among university instructors is notoriously difficult to observe. Unlike primary and secondary schools, the academy does not generally utilize incremental standardized testing as a means to calculate student progress (teacher effect). In lieu of time intensive, external review, universities in the United States widely rely upon official student evaluations of teaching (SETs) to estimate an instructor's classroom performance. While their imperfections are well-established in the literature, SETs are an integral part of hiring and promotion decisions within the United States academy. Wolfer and Johnson (2003) write, "However little confidence instructors place in student evaluations, they continue to be widely used in higher education…evaluations of teaching have two primary purposes: administrative decision making and teaching improvement" (p.111). Cohen (1981), Feldman (1989), and others show that student learning has a moderately positive correlation to official SET scores. Bosshardt and Watts (2001) summarize, "…studies generally show that student evaluations and learning, as measured by objective tests, are positively correlated but generally not higher than simple r measures of 0.7" (p.4). Thus, official SET scores can be used to (imperfectly) predict a given instructor's marginal contribution to student learning.

In recent years, several independent websites have been established that allow university students to informally evaluate their instructors. Leading sites of this nature include ratemyprofessors.com, myedu.com, passcollege.com, professorperformance.com, reviewum.com, and ratingsonline.com. As students of a given university are typically unable to view an instructor's official SETs, said sites are valuable to students who wish to inform themselves before choosing an instructor for a given class. The value of said sites is evidenced by the web traffic that they draw.

For example, ratemyprofessors.com features more than 10 million reviews of more than 1 million instructors as of April, 2010. The site features instructor reviews for more than 6,500 universities in the United States, Canada, and England. Such sites are important not only to students. Given their accessible nature, Otto, Sanford and Ross (2008) note that online evaluations may also influence the hiring decisions of faculty and administrators. Anecdotally, we are aware of at least two university faculty hiring processes that used online evaluations as an information point.

Otto, Sanford, and Ross further test the internal reliability of ratemyprofessors.com site reviews. They find the relationship between the different measures of instructor quality (helpfulness, clarity, and easiness) to be "consistent with our expectations under the assumption that the ratings reflected student learning" (p. 364). In another recent study of ratemyprofessors.com, Bleske-Rechek and Michels (2010) collected surveys and online SET scores to determine what motivations lead some students to provide online ratings. The authors find that online evaluations differentiate between such factors as how difficult an instructor is and the overall quality of the instructor. In yet another study of ratemyprofessors.com, Gonyea and Gangi (2010) develop a model to categorize and draw information from online student comments. Davison and Price (2009) find that student comments on ratemyprofessors.com are not independent of student rating. The authors also show a moderate correlation between instructor quality and instructor easiness and conclude that online SETs suffer from an anti-intellectual tone. Coladarci and Kornfield (2007) study the relationship between official SET quality and online SET quality. They find the latter variable to explain much of the variation in the former variable.

The present study's purpose is to analyze the ability of online SET quality to estimate the official (university-administered) instructional ratings of the same respective university instructors. Given self-selection among raters, we test whether more online ratings of instructors lead to better prediction of official ratings in terms of both R-squared value and root mean squared error. We also test and correct for heteroskedastic error terms in the regression analysis to allow for the first robust estimations on the topic. It is important to ascertain the validity of online evaluation scores vis-à-vis official SETs and whether said validity is dependent upon number of ratings, where official SET scores are designed to provide a representative student assessment of a professor's teaching performance.[1] Online evaluations are publicly available by design, whereas official SETs are almost never publicly disclosed. If it is the case that, in spite of their popularity, online evaluations do not provide a reliable measure of teaching quality or only do so given a sufficient number of online ratings per instructor, universities might consider publicly disclosing official SETs—perhaps at the discretion of each particular instructor. By observing how well online evaluations predict the more comprehensive official SETs, student and administrative users might learn to apply a realistic degree of confidence to composite online scores.

## Method

The data set incorporates eight semesters of evaluation data across 175 instructors at a four-year university in the southeastern United States (Auburn University Montgomery). The final data set was obtained in an anonymous format (i.e., with each instructor's name erased) and includes all instructors who a) taught at the university at some point between 2005 and 2008 and b) were rated at least five times on the website ratemyprofessors.com in course sections that took place over the same time period.[2] The teaching quality of each such instructor was estimated (in two ways) over a three-year period from Fall 2005 through Spring 2008 (i.e., Fall 2005, Spring 2006, Summer 2006, Fall 2006,…). The variable Official Quality represents an instructor's average official, university-administered SET score across all evaluating students during the time period. This average score is taken from a single question asking students to rate the instructor's overall quality in the course. For a given instructor, note that this quality measure represents the average student quality rating across all sections rather than the average class section quality rating.[3] This methodology was chosen to mirror the online rating system, in which each

[1] Official SETs may not be representative if response rates are low.

[2] The authors recorded all publicly available data for the project and sent the partial data set to the Auburn University Montgomery Office of Institutional Research. Confidential data was added by this Office. Further, instructor names were made anonymous before the data set was returned to the authors.

[3] In other words, each student of an instructor has an equal impact upon the instructor's overall quality score regardless of the student's section class size.

student, regardless of class size, has an equal opportunity to provide an online rating. The variable Online Quality represents an instructor's average teaching quality score from ratemyprofessors.com during the time period of the study. The variable nonline represents an instructor's number of ratemyprofessors.com online ratings over the sample period, and ninclass represents an instructor's number of official ratings over the sample period. There are other variables that are conceivably important to the study. For example, Ragan and Walia (2010) find differences in rating patterns between principles and non-principles courses. In the present study, course type variables were not revealed for reasons of confidentiality.

## Results

Table 1 summarizes all variables outlined in the previous section.

**Table 1:** Summary Statistics

| Variable | Obs* | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| *nonline* | 175 | 10.55 | 5.33 | 5 | 31 |
| *ninclass* | 175 | 281.30 | 153.60 | 9 | 781 |
| *Online Quality* | 175 | 3.79 | 0.90 | 1.5 | 5.0 |
| *Official Quality* | 175 | 4.16 | 0.44 | 2.77 | 4.83 |

*Each observation represents a different rated professor.

It is evident from Table 1 that *Online Quality* and *Official Quality* are distributed differently. Namely, *Official Quality* has a higher mean and lower variance than *Online Quality*. The observed differences do not, of course, preclude the latter variable from accurately predicting the former variable. If the distribution of *Online Quality* ratings represents something close to an ordinal transformation on the distribution of *Official Quality* ratings, then the model will be highly predictive in terms of ranking instructor quality. Table 1 also informs us that the average number of online ratings for an instructor over the sampled time period, *nonline*, is 10.55. Further, the number of ratings across the set of instructors has a large range and is skewed considerably to the right. The minimum value of *nonline* lies barely outside of the one standard deviation confidence interval from the sample mean, whereas the right tail of the distribution is several standard deviations above the mean. This variability and skewness may be symptomatic of differences in "rate-ability" across instructor, of differences in the number of students taught by each instructor, or of both factors.

### *The General Relationship between Official Quality and Online Quality*

We initially use OLS regression and corresponding inference to test the strength of relationship between *Official Quality* and *Online Quality* (see Table 2). It is important to note that this primary model is predictive rather than causal in its nature and purpose.

**Table 2:** Results of General OLS Regression with Robust Standard Errors

| Variable | Coefficients | Robust Standard Error |
|---|---|---|
| *Online Quality* | 0.351*** | 0.03 |
| Constant | 2.83*** | 0.14 |
| Observations | 175 | |
| R-squared | 0.521 | |
| Root MSE | 0.302 | |

*** indicates significance at the .01 level.

An application of the White test reveals that heteroskedasticity is very likely to be present (i.e., the variance of the error term appears to be dependent upon the value of *Online Quality*). The Chi-squared statistic corresponding to the White test is equal to 14.00 (p-value = .0009). We accordingly use robust standard errors to allow for valid inference (Greene 1999, p.506). Within the program Stata, the command "robust" scales the estimated error variance matrix to minimize bias. The first regression model specification and results appear as follows:

$$Official\ Quality_1 = \beta_0 + \beta_1\ Online\ Quaility_i + \varepsilon_i$$

Said estimates suggest that *Online Quality* has a positive and statistically significant relationship with *Official Quality*. The two variables have strikingly different distributions, as the graph in Figure 1 reveals.

If the two variables were identically distributed, the solid trend line would have a slope of one and an intercept of zero. However, the variable *Official Quality* is less variable than *Online Quality*, usually larger than *Online Quality* at relatively low values, and usually smaller than *Online Quality* at relatively high values. Despite said distributional differences, approximately 52 percent of the *sample* variation in *Official Quality* is predicted by

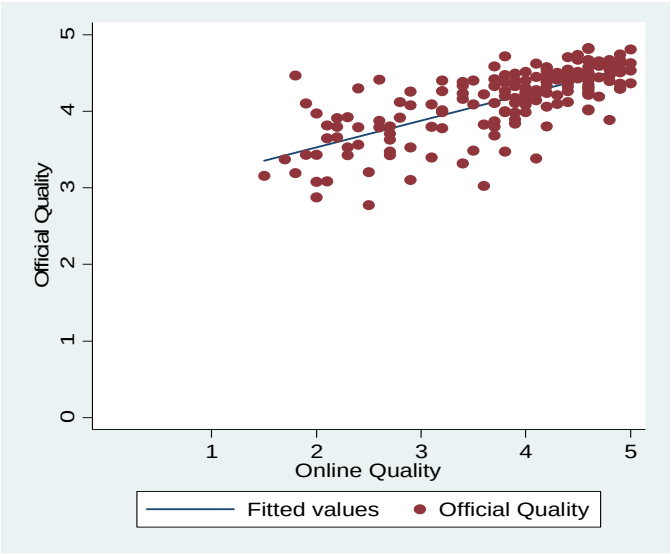


**Figure 1:** Plot of Regression Data

changes in *Online Quality*. Further, the root mean squared error shows that the average distance between an observation of *Official Quality* and the corresponding estimate of the same variable is .302 average (official) SET points. This value equals 0.69 standard deviations of *Official Quality*. To provide perspective, the margin represents roughly the true difference in the *Official Quality* rating of the 175th ranked instructor and the 172nd ranked instructor, the 124th ranked instructor and the 81st ranked instructor, or the 33rd and the 1st ranked instructor in the sample. Therefore, prediction of *Official Quality* from *Online Quality* is not a perfect science but does provide considerable insight.

We next explore *whether Online Quality* becomes a better estimator of *Official Quality* as we consider instructors with a larger number of online ratings. It is not clear, *a priori*, that this is the case. For example, the sampling method for online ratings may be sufficiently flawed to disallow such convergence. We first consider the sample correlation between the variables for different value ranges of *nonline* in Table 3.

Table 3 essentially splits the sample by quartile values of *nonline*. It is clear from the table that number of online ratings for an instructor influences the correlation between *Official Quality* and *Online Quality*. Namely, the correlation becomes more strongly positive as *nonline* increases. The analysis in Table 4 explores changes in the predictive and explanatory capabilities of the regression model, in terms of root mean squared error and R-squared, as *nonline* rises.

**Table 3:** Correlation between *Official Quality* and *Online Quality*

| Size of *nonline* | Correlation | Obs |
|---|---|---|
| [5,6] | 0.572 | 44 |
| [7,8] | 0.616 | 41 |
| [9.13] | 0.698 | 46 |
| [14,31] | 0.874 | 44 |

Total Obs = 175

From the regressions in Table 4, it is evident from the rise in R-squared values that *Online Quality* is better able to explain variation in *Official Quality* as *nonline* increases. For sample points in which *nonline* is at least 14, *Online Quality* explains more than 76 percent of the variation in *Official Quality*. Further, root mean squared error falls from the first regression to the second regression, from

**Table 4:** Summary Statistics and Sub-Sample OLS Regressions (with robust std. errors)

| | $5 \leq$ *nonline* $\leq 6$ | $5 \leq$ *nonline* $\leq 86$ | $9 \leq$ *nonline* $\leq 13$ | *nonline* $\geq 14$ |
|---|---|---|---|---|
| *Online Quality* | 0.295*** | 0.273*** | 0.346*** | 0.455*** |
| | (0.10) | (0.06) | (0.06) | (0.037) |
| Constant | 3.02*** | 3.13*** | 2.88*** | 2.47*** |
| | (0.43) | (0.27) | (0.23) | (0.15) |
| Observations | 44 | 41 | 46 | 44 |
| R-squared | 0.327 | 0.380 | 0.488 | 0.764 |
| Root MSE | 0.335 | 0.291 | 0.308 | 0.258 |
| $\bar{x}_{online\ quality}$ | 4.07 | 3.97 | 3.63 | 3.50 |
| $s_{online\ quality}$ | .78 | .82 | .86 | 1.01 |

Robust standard errors in parentheses. *** indicates significance at the .01 level.

the first regression to the third regression, from the first regression to the fourth regression, from the second regression to the fourth regression, and from the third regression to the fourth regression (to a value of 0.258 average, official SET points—roughly the true difference between the 175th and 173rd ranked instructors, the 124th and 91st ranked instructors, or the 1st and 24th ranked instructors in the sample). In general, then, it appears that *Online Quality* becomes more predictive of *Official Quality* as *nonline* rises. Overall, the analysis suggests that *Online Quality* becomes a better estimator of *Official Quality* as we consider instructors with a larger number of online ratings. From the summary statistics at the bottom of the table, we observe that more frequently rated professors obtain lower average ratings. In the following section, we explore this relationship in greater depth.

### *Explaining Variation in Number of Online Ratings across Instructor*

We next consider why some instructors are rated online more frequently than others (see Table 5). The most obvious explanation of this variation is that some instructors teach more students. However, we also consider whether there exists a relationship between *nonline* and average instructional quality (*Online Quality*) in the following regression:

$$nonline_i = \beta_0 + \beta_1\, ninclass_i + \beta_2\, OnlineQuaility_i + \varepsilon_i$$

There are many unobserved factors that cause variation in *nonline*. However, the model does inform us that *nonline* rises as an instructor teaches more students and as an instructor's *Online Quality* rating declines. An individual with an *Online Quality* rating of 3.0 is expected to receive two more ratings, *ceteris paribus*, than an instructor with an *Online Quality* rating of 4.15. Students are more likely to rate instructors that they view as relatively poor in quality. This suggests that venting one's frustrations may serve as a disproportionate motivation to leave online instructional ratings. It also suggests that reviewer (self-)selection bias may not be constant across the distribution of *Online Quality* ratings.

**Table 5:** Results of OLS Regression explaining *nonline* heterogeneity

| Variable | Coefficients | Robust Standard Error |
|---|---|---|
| *ninclass* | 0.010*** | 0.0003 |
| *Online Quality* | -1.74*** | 0.0451 |
| Constant | 14.20*** | 2.0900 |
| Observations | 175 | |
| R-squared | 0.153 | |
| Root MSE | 4.93 | |

*** indicates significance at the .01 level

## Conclusion

Within an OLS regression model that controls for heteroskedasticity, online ratings explain much of the variation in official ratings from one instructor to another. This conclusion strengthens as one considers instructors with a larger number of online ratings. Among instructors receiving at least 14 online ratings over the sample period, *Online Quality* explains 76.4 percent of variation in *Official Quality*, as compared to 52.1 percent in the regression of the general sample. Despite self-selected sampling in the case of *Online Quality*, the two variables correlate more highly as one considers instructors with a larger number of online ratings. The simple correlation coefficient between the two variables is 0.887 in the aforementioned sub-sample, as compared to 0.722 in the general sample. In another comparison of regressions, the root mean squared error falls from 0.335 average SET points for instructors with five to six online ratings to 0.258 average SET points for instructors with fourteen or more online ratings.

Lastly, we explore why some instructors are rated more frequently than others. This heterogeneity is found to be rooted in the number of students that an instructor teaches and the quality of the instructor, as perceived by students. Instructors who receive low average online SET scores are typically rated more frequently. This may suggest that venting one's frustrations serves as a disproportionate motivation to leave online instructional ratings. It also suggests that (self-)selection bias may not be constant across the distribution of Online Quality ratings. There are avenues for future study on the subject of online instructional reviews. For example, if online SET scores correlate positively to official SET scores and official SET scores correlate positively to student learning, it may be that online SET scores correlate positively to student learning. In such a case, online reviews would not only serve the superficial preferences of students but would also lead to better matching between student-type and instructor-type toward the

improvement of student learning. Whether there is a positive relationship between *online* SET scores and student learning is outside the scope of the present study but is certainly ascertainable.

## References

Bleske-Rechek, A. and K. Michels. 2010. RateMyProfessors.com: Testing assumptions about student use and misuse. *Practical Assessment, Research & Evaluation* 15(5): 1-12. Available: http://pareonline.net/getvn.asp?v=15&n=5.

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

Cohen, P. 1981. Student ratings of instruction and student achievement. *Review of Educational Research* 51(3): 281-30.

Feldman, K. 1989. The association between student ratings of specific instructional dimensions and student achievement: refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education* 30(6): 583-645.

Coladarci T. and I. Kornfield. 2007. RateMyProfessors.com versus formal in-class student evaluations of teaching. *Practical Assessment, Research and Evaluation* 12(6): 1-15. Available: http://pareonline.net/getvn.asp?v=12&n=6.

Davison, E. and J. Price. 2009. How do we rate? An evaluation of online student evaluations. *Assessment and Evaluation in Higher Education* 34(1): 51-65.

Gonyea, N. and J. Gangi. 2010. Reining in student comments: a model for categorising and studying online student comments . *Assessment and Evaluation in Higher Education* forthcoming.

Greene, W. 1999. *Econometric Analysis, Fourth Edition*. Prentice Hall: New Jersey.

Otto, J., Sanford, D. & D. Ross. 2008. Does ratemyprofessors.com really rate my professor? *Assessment and Evaluation in Higher Education* 33(4): 355-368.

Ragan, J. and B. Walia. 2010. Differences in student evaluations of principles and non-principles economics courses and the allocation of faculty across these courses. *Journal of Economic Education* forthcoming.

Wolfer, T. A. & Johnson, M. (2003) Re-evaluating student evaluation of teaching: the teaching evaluation form, *Journal of Social Work Education,* 39(1): 111–120.

## Citation:

Sanders, Shane, Walia, Bhavneet, Potter, Joel & Linna, Kenneth W. (2011). Do more online instructional ratings lead to better prediction of instructor quality? *Practical Assessment, Research & Evaluation*, 16(2). Available online: http://pareonline.net/getvn.asp?v=16&n=2.

## Authors:

Shane Sanders
Assistant Professor of Economics
Department of Economics and Decision Sciences
Western Illinois University
sd-sanders [at] wiu.edu

Bhavneet Walia
Assistant Professor of Economics
Department of Economics and Decision Sciences
Western Illinois University
b-walia [at] wiu.edu

Joel Potter
Assistant Professor of Economics
Mike Cottrell School of Business
North Georgia College and State University
jmpotter [at] northgeorgia.edu

Kenneth W Linna
Associate Professor
Department of Economics
Auburn University Montgomery
klinna [at] aum.edu