Assessment & Evaluation in Higher Education
Vol. 33, No. 4, August 2008, 355–368

Routledge
Taylor & Francis Group

Does ratemyprofessor.com really rate my professor?

James Otto*, Douglas A. Sanford, Jr. and Douglas N. Ross

Towson University, USA

This article explores the usefulness and validity of self-selected online student ratings of faculty. These ratings appear to be increasing in importance, particularly as students utilize them as guides in their choice of instructors. Validity is investigated by analyzing the pattern of relationships of online ratings for 399 randomly selected faculty. Analysis suggests that online ratings in their current form may be useful, even though possible abuses could limit validity in specific instances.

Introduction

The evaluation of faculty teaching by students has come to be one nearly ubiquitous measure of teaching effectiveness and, often, a major consideration for promotion, tenure and merit at most institutions of higher education.[1] Research on student evaluations of faculty has shown that evaluations typically are based on forms that are filled out anonymously by students in a classroom using formal, well-defined and controlled processes (Cashin 1995; Martin 1998; Read et al. 2001; Centra 2003).

We extend this research by exploring a new, exciting and quite different source of faculty rating data that has recently become available on the World Wide Web – online faculty rating sites. Our research question is: Does the pattern of relationships between self-selected online student ratings of faculty reflect student learning or a halo effect (Feeley 2002)? We also look at faculty feedback and student feedback to assess the use of online ratings for administrative selection of course offerings, as well as for promotion, tenure and merit deliberations, and in hiring decisions for new faculty.

Online faculty rating sites include RateMyProfessor.com, PassCollege.com, ProfessorPerformance.com, RatingsOnline.com and Reviewum.com (Foster 2003; Stone 2003). It may be argued that data from these sites are characterized by bias such as instructors' personality, charisma and grading leniency, and are therefore not of value as a measure for either faculty performance or student learning (see Cashin 1996, 1999; Greenwald and Gilmore 1997; Wilson 1998; Liaw and Goh 2003). In fact, this interpretation is most common among the faculty we meet. However, when one examines the amount of traffic at these sites, it becomes evident that online rating sites are popular with students. Some students use the data on these sites to develop expectations of their professors and set schedules, which indirectly affect faculty teaching loads and student expectations. Thus, these ratings probably should be taken seriously by both faculty and administrators.

Our analysis concentrated on the most popular site, in terms of usage: RateMy Professor.com (also known as RateMyProfessors.com). There has been a marked increase in activity. In May 2003, 2.7 million ratings of 478,000 faculty had occurred, and by August

*Corresponding author. Email: jotto@towson.edu

ISSN 0260-2938 print/ISSN 1469-297X online

DOI: 10.1080/02602930701293405
http://www.informaworld.com

2006 the numbers had risen to over 5.7 million ratings of about 770,000 professors in nearly 6000 schools (RateMyProfessor 2006a). However, it is also true that these ratings represent a small percentage of all students who register for courses taught by a professor. Many professors have only one rating, and even some of the most heavily rated professors, with 40 to 50 ratings, have had a great deal more students in their classes.

Only very recent research exists on the faculty rating data from these sites. Felton et al. (2004) analyzed overall professor quality ratings as a function of easiness and sexiness ratings. Their findings were that ratemyprofessor.com ratings did not reflect student learning, but were characterized by a halo effect (Feeley 2002). Felton et al. (2004) found that faculty characteristics such as easiness and sexiness (beauty and attractiveness), as perceived by students, tended to be associated with overall positive ratings. Otto et al. (2005) investigated the interrelationships between online rating variables and described a number of issues related to bias, availability of ratings and purpose of ratings. They had mixed results, with some patterns reflecting student learning and others, such as the correlation between easiness and clarity, suggesting halo effect.

We agree with leading researchers in this field that student learning is the most valid purpose of faculty ratings for performance in courses. It is more important for faculty to facilitate student learning than to cause student satisfaction. Student learning is a fundamental goal that faculty achieve in courses. Because the online ratings we use do not include a direct measure of student learning, we use student learning as a latent variable that may be causing the patterns of student ratings that we observe.

Online ratings may be biased. Ratings may be entered by anyone and at any time. They may be affected by emotion. And because only some students enter ratings, they have potential for selection bias. All these factors suggest that the ratings have a halo effect—that students who make ratings of their professors either give them universally high or low ratings without accurately reporting distinct aspects of faculty performance.

On the other hand, it is possible that online ratings may not be biased. Research on online ratings suggests that the online format does not lead to substantially biased ratings (Carini et al. 2003; Hardy 2003; McGhee and Lowell 2003). Students who post ratings may be regarded as experts who have had significant experience with the professors. They may also have consulted with a number of other students who share the same perspective, so that online ratings may represent a far larger and more representative sample of students than the numbers suggest. Even if some students give biased ratings, these may be balanced between those that are positive and those that are negative. For example, students who have had a professor may visit the site and see that the existing ratings are not what they experienced, so they might enter a rating that 'corrects' for the listed ratings. So the average rating for a professor across a number of students may reflect less bias than any particular rating.

It is arguable that the online student ratings could improve both student learning and instructor performance. Students may respond to this rating information by choosing instructors who are best suited to their learning styles. Online ratings are available to instructors too, and in response to poor online ratings instructors may work to improve their performance. These improvements, however, are dependent on the validity of the ratings as a measure of student learning. If, for example, the student ratings reflect instructor charisma, sense of humor, grading leniency or entertainment value rather than learning, then instructors will be motivated to achieve something not in the interest of higher education.

There is a substantial literature on the extent to which traditional faculty evaluations are a valid measure of student learning (Cashin 1995; Marsh and Roche 1997, 2000; McKeachie

1997; Harrison et al. 2004). These studies indicate that 'proving' bias or lack of bias in student evaluations of teaching is difficult, largely because of the challenge of reliably measuring student learning. However, research findings do consistently hold that learning is positively associated with instructor clarity and instructor helpfulness.

Research findings also hold that course difficulty is not linearly associated with student learning. Centra (2003) elaborated on this finding. His study confirmed a lack of linear relationship between course difficulty and student learning, but suggested a non-linear relationship. Students may learn best not when the course is too easy or too difficult, but when the difficulty is between these extremes, or 'just right'.

There is also a stream of research related to online versus paper surveys of faculty performance. In general, online ratings were found to be about the same as paper ratings. The online ratings were, if anything, slightly higher than paper ratings (Carini et al. 2003). Other researchers found that the properties of student ratings were similar for both paper and online responses (Hardy 2003; McGhee and Lowell 2003). Even though there was a slightly lower response rate for online as opposed to paper ratings, researchers conclude that with acceptance of online ratings by students and faculty, appreciation of the advantages of online ratings (thoughtful feedback, convenience and user-friendliness) and the development of online infrastructure, this response difference will diminish (Ballantyne 2003; Johnson 2003).

Objective and approach

The objective of this paper is to examine the pattern of association between components of online ratings and then to test whether they are more consistent with the pattern expected of valid measures of student learning or, alternatively, halo effect. We cannot definitively prove or disprove whether online student ratings are valid. But we can test the commonly held proposition that self-selected student online ratings are due to halo effect and therefore extremely suspect. We also offer a framework for further investigation.

We think that the growing use of online ratings justifies our efforts. We understand that online ratings are subject to possible bias, but also think that students may be providing their peers with accurate and dependable information. And we are sure that some of our students rely on ratemyprofessor.com ratings for selecting their professors. We want to know if we as faculty should ignore ratemyprofessor.com ratings, and even discourage students from relying on them, or if we should admit that the ratings may contain some useful information.

Previous research indicates that the ratings components of instructor clarity in the presentation of course material; instructor helpfulness as perceived by students; and course easiness will associate with student learning as shown in Figure 1. Our data source is from ratemyprofessors.com, and the definition of clarity, helpfulness and easiness, was provided by the website.[2] Students provide ratings on a scale of 1 to 5 as follows. For easiness, the rating is from 1 (hard) to 5 (easy). For both helpfulness and clarity, the rating is from an icon for 'frowny face' to an icon for 'smiley face'. Previous research demonstrates that both instructor clarity and instructor helpfulness are linearly positively correlated with learning, and we show this with the two positively sloped lines for clarity and helpfulness.

The relationships in Figure 1 imply that instructors who are perceived to be helpful to students and have clarity in their presentation of the course material will have increased learning in their classes. Therefore, if the online ratings reflect student learning, then we expect the following.

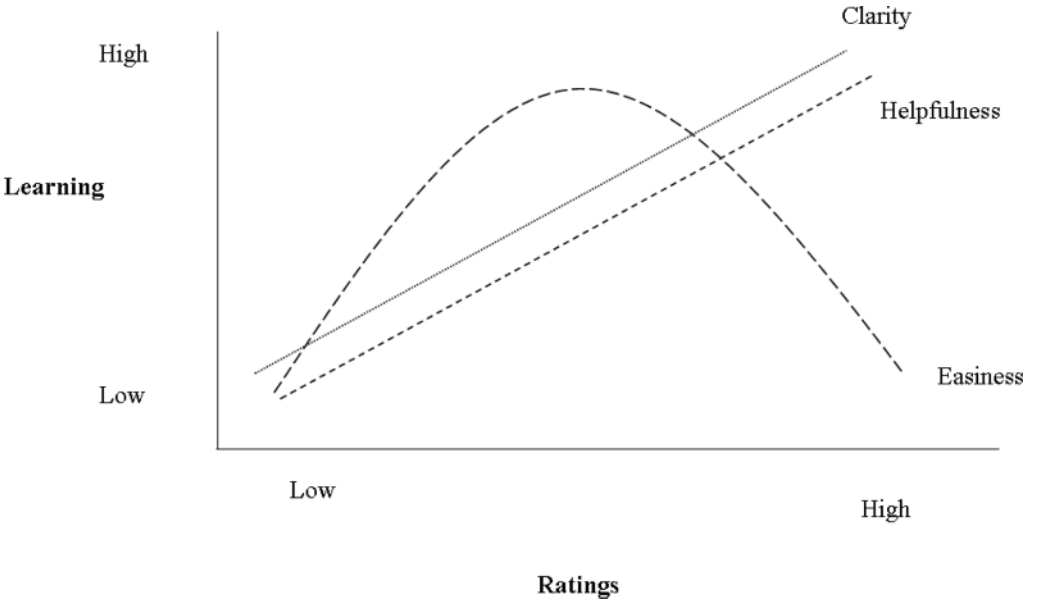

Figure 1. Relationships between learning and clarity, helpfulness, and easiness.

H1. Helpfulness and clarity will be positively correlated.

The association between learning and easiness found in the literature is less well demonstrated. Most studies show that course difficulty (the opposite of easiness) is either insignificantly or positively associated with student learning (Marsh and Roche 1997, 2000). However, Centra's (2003) suggestion of a non-linear relationship is logically compelling, as instructors can impede learning by either being too easy, and not challenging students, or too difficult, and frustrating them. On Figure 1, we show this non-linear relationship with a non-linear inverse 'U' relationship between student ratings for easiness and student learning. At low levels of student learning, rating for easiness can have either high or low values, as indicated by the fact that a horizontal line from low learning intersects the easiness line twice. Instructors who do not promote learning would tend to be either too easy or too difficult.

Similarly, with high levels of learning, easiness would tend to have middle values and a smaller variation. This pattern reflects the fact that instructors who promote learning tend to be helpful and clear, but neither too easy nor too difficult. The optimal level of learning will be associated with instructors who have levels of easiness in between the high and low ratings.

Figure 2 shows this pattern. Again assuming that the ratings reflect learning, at low levels of clarity and helpfulness, easiness may be either high or low, suggesting high variability. At high levels of clarity and helpfulness, easiness is expected to have a moderate value and a low variability.

H2a. The level of variability in easiness, measured by the squared difference from its mean, will be inversely related to the level of clarity.

H2b. The level of variability in easiness, measured by the squared difference from its mean, will be inversely related to the level of helpfulness.

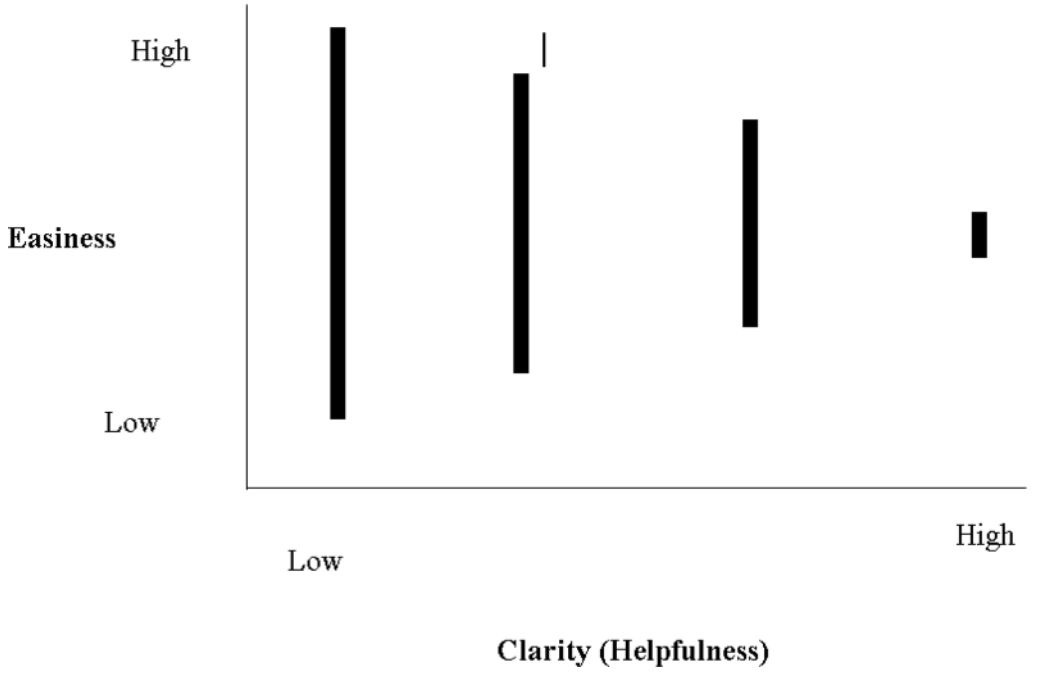

Figure 2. The relationship between clarity (helpfulness) and easiness
Note: Bars represent the expected range of responses in easiness at various levels of clarity (helpfulness). With variability measured as the squared difference between the value of easiness and the mean value of easiness, we expect that for instructors with lower levels of clarity and helpfulness, the average variability in easiness will be greater than for instructors with higher levels of clarity and helpfulness.

If the ratings reflect a halo effect, then we would expect a different pattern. Students would not distinguish between the concepts clarity, helpfulness and easiness. They would rate professors high or low on all three measures depending on some overarching characteristic such as popularity or charisma. Students making ratings reflecting a halo effect would be signaling their peers simply to seek out or avoid this professor, and not providing more nuanced information regarding the learning potential from taking the course. As a result, we would find the variance for all easiness to be similar for all levels of clarity and helpfulness.

Methodology

The first step was to randomly select about 400 institutions from the RateMyProfessor.com website total of 4077 educational institutions, as listed in late 2004. This procedure ensured that each of the institutions represented in the RateMyProfessor database had an equal chance of being selected. The second step was to randomly select from each of the chosen institutions one faculty member and to download his or her ratings (see Appendix 1 for further clarification of our sampling methodology). In the database, two institutions were selected three times and 22 institutions were selected twice. The resulting sample contained 399 unique faculty members from 373 institutions.

Each downloaded observation contained the number of times the faculty member was rated, the average rating for easiness, the average rating for helpfulness, the average rating

for clarity, the number of 'hotness' rankings, the faculty member's name, the name of the institution, the institution's location, and the academic field or department of the faculty member. The ratings for easiness, helpfulness and clarity ranged from 1 (low) to 5 (high). While student responses are based on perception, we assume that the data are interval level, or that the difference between a '1' and a '2' is approximately the same level of difference as between any other two ratings. This assumption justifies our use of correlation and regression analysis. The variables for the number of times a faculty member was counted and 'hotness' were counts, and the other variables were categorical.

These average ratings compensate for the 'errors-in-variables' problem associated with having only one rating per faculty member. Our unit of analysis for this research was the faculty member, not each student rating. With more student ratings per faculty member, the resulting average rating will be more reliable (Cashin 1995). Using individual student ratings is problematic, as they are not independent. Students have the opportunity of looking at other ratings for a professor before entering their own, and are likely to make ratings that are influenced by what they have seen for the professor.

We include information on the variables used in our analysis in Table 1, including variable mean, standard deviation and a correlation matrix. For example, in line 3 average easiness is 3.09 on a (ratemyprofessor) 1–5 scale, which means the average rating of the 399 faculty for easiness was middle of the road. The standard deviation shows the variation in the variable across faculty, and means that about 68% of ratings are between 2.02 and 4.16, or 3.09 plus or minus 1.07 (provided the variable has a normal distribution). In Table 1 line 4, the variability in easiness has an average value of 1.14, which shows for each faculty their tendency to have a rating different from the mean. (This variable enabled us to test hypothesis 2). In line 4, column 3, the correlation shows the statistic with variable 1 'average helpfulness' is –0.24, which means that variability and easiness tend to be lower with higher levels of average helpfulness.

In order to facilitate analysis, we created additional variables for the variability of easiness, faculty gender, Carnegie classification and academic field. The 'variability of easiness' is the squared distance from the mean value for the variable 'easiness.' We would expect that there would be on average larger deviations from the mean for low levels of learning—either too easy or too hard—and smaller deviations for higher levels of learning—or a 'just right' balance of easy–difficult.

We included a variable for faculty gender in our analysis to account for the possible impact that faculty gender has on student ratings (Baldwin and Blattner 2003). We interpreted female gender from the faculty member names. This procedure was likely to be slightly inaccurate because (a) names such as Carol, Trubnik and Kim could denote either gender, and (b) some names in the database included only surnames. We coded as female only those names that we could reasonably identify as female, leaving the undecided cases coded the same as males. While this variable may be slightly inaccurate, it does at least partially account for systematic gender effects.

For Carnegie classification, we looked up each institution on the Carnegie Foundation website (Carnegie Foundation 2005). We were able to categorize 360 of the 373 institutions into Carnegie classes, and found some differences in ratings for easiness, helpfulness and clarity for those classes with more than 10% of the sample. These classes included Master's colleges and Universities I, Baccalaureate colleges–general, and Associate's colleges.

For academic field, the sample included faculty in 48 fields. We created binary variables for faculty in fields with 20 or more ratings to control for field-specific effects. These were English, Mathematics, Business and Computer Science.

Table 1. Descriptive statistics and correlations among the variables in the analysis.

Variable	Mean	SD	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Average helpfulness	3.79	1.30	1.00												
2. Average clarity	3.69	1.30	0.87	1.00											
3. Average easiness	3.09	1.07	0.41	0.38	1.00										
4. Variability in easiness*	1.14	1.38	−0.24	−0.20	0.01	1.00									
5. Female gender	.33	.47	−0.08	−0.09	−0.05	0.06	1.00								
6. Master's I	.11	.31	−0.01	−0.01	−0.03	0.01	−0.05	1.00							
7. Baccalaureate—General	.10	.30	−0.03	−0.07	−0.03	−0.03	−0.01	−0.12	1.00						
8. Associate's colleges	.39	.49	0.08	0.09	0.16	0.01	0.07	−0.28	−0.27	1.00					
9. English	.10	.30	−0.02	−0.07	−0.06	−0.01	0.08	0.07	0.02	0.05	1.00				
10. Mathematics	.07	.25	−0.07	−0.11	−0.07	−0.05	0.05	−0.09	0.04	0.07	−0.09	1.00			
11. Business	.06	.24	0.05	0.07	0.03	−0.06	0.00	−0.05	0.01	0.00	−0.09	−0.07	1.00		
12. Computer Science	.05	.22	−0.02	−0.01	0.10	0.05	−0.09	−0.04	−0.08	0.14	−0.08	−0.06	−0.06	1.00	
13. Hotness rating	.17	.35	0.28	0.29	0.15	−0.01	0.01	0.05	−0.02	0.04	−0.04	0.05	0.00	−0.01	1.00
14. Multiple ratings	.52	.50	−0.02	−0.05	−0.05	−0.18	0.02	0.13	−0.02	−0.03	0.01	0.04	0.12	0.01	−0.14

Note: *We subtracted the value from the mean of easiness for the sample, then squared it; technically this gives positive numbers and enables us to distinguish observations that have high deviations from the mean, and therefore high variability, from observations that have low variability.

We also divided the 'hotness' rating by the number of ratings for each faculty member, which created a ratio of 'hotness' for each faculty member. Perceived 'hotness' may be a measure of a type of universal appeal, and has been shown to affect overall ratings (Felton et al. 2004).

We also used a variable for the number of ratings a faculty member might have, to control for any tendency for number of ratings to associate with high or low clarity, helpfulness or easiness. We transformed the number of ratings into a binary variable with a value of 1 for multiple ratings and 0 for only one rating. Forty-eight of the faculty in our sample had only one rating.

Analysis and results

Hypothesis 1 is that helpfulness and clarity will be positively correlated. Average helpfulness and average clarity are strongly correlated (0.87, $p < 0.000$) as shown in Table 1. We interpret this high correlation to reflect the fact that faculty who clearly explain course material also tend to be very helpful. While these variables have some cognitive overlap, the construction of the items in the website, and the fact that the correlation is significantly less than 1, indicate that they are cognitively distinct and not a single item construct. Other analyses, including regression of average helpfulness on average clarity with control variables, give similar results. These findings support hypothesis 1 and confirm similar results originally reported by Otto et al. (2005).

Hypothesis 2a is that the variability in easiness will be inversely related to clarity. We tested this hypothesis with multiple regression, using clarity as the dependent variable and variability in easiness as an independent variable. Our model included control variables for gender, Carnegie classification, academic area, hotness ratio and multiple ratings. As shown in Table 2, column 2, the coefficient for 'variability in easiness' is negative and significant as hypothesized.

Similarly, hypothesis 2b, that the variability in easiness will be inversely related to helpfulness, is tested with the regression model shown in Table 2, column 3. The negative and significant coefficient for the variable 'variability in easiness' supports the hypothesis.

The regression models have significant explanatory power as indicated by the R squared and F statistics.[3] Examination of the residuals showed no evidence of non-spherical disturbances. The variance inflation factors for each variable were under 1.3, indicating that the results are not affected by multicollinearity.

Table 1 also shows that average clarity and average helpfulness associate negatively with the variability in easiness. This bivariate analysis also supports Hypotheses 2a and 2b.

Discussion

The analysis shows statistical support for all three hypotheses. The variables clarity, helpfulness, easiness and variability in easiness demonstrate patterns of association that are consistent with the assumption that ratemyprofessor.com ratings reflect student learning. These findings are inconsistent with the assumption that the ratemyprofessor.com ratings reflect a halo effect. However, there is a positive and significant correlation between easiness and clarity/helpfulness as shown in Table 1, row 3. This correlation is not consistent with previous research showing low correlation between workload or course difficulty and student learning (Marsh and Roche 1997, 2000; McKeachie 1997).

With regard to easiness, some interpretation may be in order. First, the ratemyprofessor.com variable of easiness may have a positive interpretation. Easy could mean either low

Table 2. Regression analysis results.

Variable	Dependent variable Unstandardized coefficients	
	Average clarity (standard error)	Average helpfulness (standard error)
(Constant)	3.86** (.14)	3.92** (.14)
Variability in easiness	−.20** (.05)	−.22** (.05)
Female gender	−.21 (.13)	−.20 (.13)
Carnegie classification: Master's	−.02 (.21)	−.02 (.22)
Carnegie classification: Baccalaureate—General	−.21 (.21)	−.04 (.21)
Carnegie Classification: Associate's Colleges	.24 (.14)	.24 (.14)
Academic Area: English	−.28 (.20)	−.07 (.20)
Academic Area: Mathematics	−.70** (.25)	−.53* (.25)
Academic Area: Business	.25 (.25)	.13 (.26)
Academic Area: Computer Science	−.20 (.28)	−.21 (.29)
Hotness ratio	1.06** (.18)	1.00** (.18)
Multiple ratings	−.11 (.13)	−.05 (.13)
Adjusted R squared	0.145	0.131
F statistic	7.14**	6.45**
n	399	399

Note: *$p < 0.05$, **$p < 0.01$.

workload—not challenging or that the course material was explained so well that it was 'easy' to understand. Survey researchers have long noted that survey responses are extremely sensitive to negative and positive presentation and item wording (see Fowler 1995). Easy may thus elicit a different response than the standard questionnaire items that measure 'workload' or course 'difficulty', which may have more of a negative interpretation. It is entirely possible that students interpret easiness in a manner consistent with learning, such as 'this professor makes the course material easy to understand'. Further research into how students interpret their ratings could shed light on this issue.

There may be latent variables, correlated with student learning and easiness, that affect the ratings. There are many possibilities, such as professor charisma, congeniality, popularity or ability to hold students' interest in the classroom. Any such latent variable may associate with perceived clarity, perceived helpfulness and perceived easiness. Further research into a more complete set of ratings scales, including charisma, liking and instructor personality, could shed light on this issue.

We are aware of other limitations in our model. Our analysis relies on the assumption that student learning will generate a specific pattern among student ratings. The variables for gender, Carnegie classification and academic field or department may not be accurate. We interpreted gender imperfectly from faculty names. We did not find the Carnegie classification for all institutions, and the measure of academic field or department may not be the field of the course in which the faculty member was teaching. The insignificance of the coefficients for these variables may reflect their inaccuracy. We also analyzed the patterns for only a few institutions, primarily in North America, at one point in time. And of course there is the possible problem of selection bias because we analyzed only a few institutions out of the entire population. Addressing these limitations in our methodology is an objective of future research.

Conclusion

Our analysis of online ratings from ratemyprofessor.com showed a similarity with what might be expected if the ratings were valid measures of student learning. Our analysis of a random sample of 399 ratings demonstrated that students' ratings of instructor clarity and helpfulness were strongly correlated. In addition, we found that the variability in easiness was inversely associated with clarity and helpfulness. These findings were consistent with our expectations under the assumption that the ratings reflected student learning. Our findings were not consistent with what we would expect of ratings characterized by a halo effect. Overall, these results have potential implications for instructor evaluation and perhaps promotion, tenure and merit procedures.

Online ratings may not be a biased measure of student learning. Student responses may nonetheless reflect honest and true assessments of instructors. Our evidence is weak in this regard, but surprising given the high potential for bias. To the extent that future research can further demonstrate with greater confidence that online ratings are unbiased, it may be appropriate to consider using this information to supplement decisions with regard to faculty hiring and promotion, tenure and merit decisions. Of course, such a change would have significant implications for university employment practices and policies.

Further research is needed to investigate and improve the conditions under which online ratings would be valid measures of student learning. We suggest three approaches. First, student interpretations of the wordings of online survey items need to be clarified, especially with regard to 'easiness'. A variety of methods including focus groups, follow-up surveys, nominal group technique analyses, and other qualitative techniques can be used. Second, analysis of students' responses to open-ended questions can be incorporated into the analysis. In particular, student comments can suggest whether a rating is an objective assessment of instructor effectiveness or the result of personal reactions to an instructor's personality or teaching style, and not student learning. Third, online ratings could be compared with traditional end-of-course ratings of teaching effectiveness. Such an analysis could demonstrate the validity of online ratings as a measure of student learning, or perhaps demonstrate the conditions under which online ratings are most valid.

To limit bias in online ratings, websites such as ratemyprofessor.com can revise their surveys. For example, the definitions for easiness, clarity and helpfulness can be made more prominent. Access to the website can be limited, so that only students who have taken a course can submit a rating. And the websites can collect other information, such as the recent effort to include 'level of interest', that can tell other students whether a rating is based on an objective assessment of instructor competence or a comment on an instructor's personality. And timing can be factored in more explicitly. Old ratings may not be relevant, as professors sometimes do improve over time. Some method for emphasizing more recent, and relevant, ratings could improve accuracy.

Some genuine skepticism exists as to the efficacy of online ratings. A recent audience at a professional meeting, attended mostly by faculty (but including two students), recognized that ratemyprofessor.com ratings were used by many students as a source of information on their professors. Some comments:

- In the first class of a semester, a student raised his hand and said, 'But this isn't the course I had expected'. And indeed, the professor had just changed the course. The student's expectations were based on what he had read on ratemyprofessor.com.
- A department chair noted that a woman member was crestfallen because her ratings on ratemyprofessor were low, with comments like 'don't go to her for fashion tips'.

The department chair then went to ratemyprofessor.com and made a couple of very flattering entries designed to lift the spirits of the faculty member.

- A faculty member had about 20 ratings in ratemyprofessor. All these were low ratings except for one. The one high rating included comments in the first person that defended the faculty member's teaching.
- A professor mentioned that a student came to him for advising, and was not happy with the advice he received. At the end of the meeting, the student said, 'I'm gonna screw you on ratemyprofessor.' Sure enough, the next day there was a negative entry on the professor in ratemyprofessor.com.

These comments raise an issue that we think is critical if online ratings are to be used in faculty performance evaluation. Obviously, abuses of the integrity of the system happen. While anecdotes imply that some of the ratings at present may be invalid, the question is: Of the millions of ratings in ratemyprofessor.com, how many are biased? If there are few, or if the few are balanced between positive and negative ratings, then the existence of biased ratings may be as small a problem as with other surveys of instructor performance. Because we used average ratings for faculty members, the few (possibly) biased ones might have a limited impact on the average.

We think this research is timely and important. Ratemyprofessor.com ratings are growing in usage and popularity, and we think that they affect student decisions about which professors to take. In addition, students are developing expectations of instructors based on peer comments in online ratings websites. If the websites do not provide valid information, then the ratings would not provide accurate information and thus students would probably respond to correct the ratings.

On the other hand, to the extent that online ratings can be demonstrated as valid measures of instructors' abilities to inspire learning, online ratings have potential value. Increased transparency of the ratings should have virtuous effects for students to select professors who conform to their learning styles. And instructors should improve their teaching methods in order to improve their ratings. Further research and improvement of the system can clarify the conditions for improving online ratings.

Notes

1. A possible confusion resides in the terminology. An 'evaluation' implies a conclusion based on some direct definitive measure, whereas a 'rating' connotes data susceptible to interpretation. We use the term 'rating' in order to distinguish more clearly between data providers and data interpreters.
2. From ratemyprofesor web site (RateMyProfessor 2006b):

   - *Easiness*: This is definitely the most controversial of the three rating categories, which is why it is NOT included in the 'Overall Quality' rating. Although we do not necessarily condone it, it is certainly true that many students decide what class to take based on the difficulty of the teacher. When rating a teacher's easiness, ask yourself 'How easy are the classes that this professor teaches? Is it possible to get an A without too much work?'
   - *Helpfulness*: This category rates the professor's helpfulness and approachability. Is the professor approachable and nice? Is the professor rude, arrogant or just plain mean? Is the professor willing to help you after class?
   - *Clarity*: This is the most important of the three categories, at least to many people. How well does the professor convey the class topics? Is the professor clear in his/her presentation? Is the professor organized and does the professor use class time effectively?
   - *Overall quality*: The overall quality rating is the average of a teacher's helpfulness and clarity ratings, and is what determines the type of 'smiley face' that the Professor receives. Due to

popular demand, a teacher's easiness rating is NOT used when computing the overall quality rating, since an easiness of 5 may actually mean the teacher is TOO easy.
- *Hotness*: It's fun but has deeper significance. Beauty and attractiveness may have a payoff in some occupations.

3. R squared is the relative predictive power of a model. R squared is a descriptive measure between 0 and 1. The closer it is to 1, the better your model is. By 'better' we mean a greater ability to predict. A value of R squared equal to one, which almost never occurs, would imply that your quadratic regression provides perfect predictions. The F statistic (F) is a ratio of (estimated) population variances with the regression mean square variance divided by the error mean square variance. The F statistic tests whether all estimated variable coefficients are significantly greater than 0.

Notes on contributors

James Otto (CBD, Ph.D., University of Kentucky) is an Associate Professor of Management Information Systems in the Management Department at Towson University. His teaching interests include management information systems, electronic commerce and Internet programming. His current research interests include data analysis and artificial intelligence; electronic commerce; and Internet technologies.

Douglas Sanford (Ph.D., University of Michigan) is an Assistant Professor in the Management Department at Towson University where he has taught international business courses. He has published research on educational quality improvement, regional economics and international marketing.

Douglas N. Ross (Ph.D., University of Colorado) is Professor, Department of Management, Towson University. He has been a Fulbright scholar—senior specialist—in both Sweden and the Czech Republic. His teaching interests include undergraduate and graduate courses in business policy & strategic management, international business, business and society, and business strategy and the Internet. He has been awarded Teacher of the Year. Current research interests include: comparative management and culture, integration of technology and teaching, and business case studies focused on the biotechnology industry.

References

Baldwin, T., and N. Blattner. 2003. Guarding against potential bias in student evaluations: what every faculty member needs to know. *College Teaching* 51, no. 1: 27–32.

Ballantyne, C. 2003. Online evaluations of teaching: an examination of current practice and considerations for the future. *New Directions for Teaching and Learning* Winter (96): 103–112.

Carini, R.M., J.C. Hayek, G.D., Kuh, J.M. Kennedy, and J.A. Ouimet. 2003. College student responses to web and paper surveys: does mode matter? *Research in Higher Education* 44, no. 1: 1–19.

Carnegie Foundation. 2005. Carnegie Classification of Institutions of Higher Learning. Available online at: http://www.carnegiefoundation.org (accessed August 2006).

Cashin, W. 1995. *Student ratings of teaching: the research revisited.* Idea Paper No. 32, Center for Faculty Evaluation and Development, Kansas State University.

———. 1996. *Developing an effective faculty evaluation system.* Idea Paper No. 33, Center for Faculty Evaluation and Development, Kansas State University.

———. 1999. *Appraising teaching effectiveness: beyond student ratings.* Idea Paper No. 36 Center for Faculty Evaluation and Development, Kansas State University.

Centra, J.A. 2003. Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education* 44, no. 5: 495–518.

Feeley, T.H. 2002. Evidence of halo effects in student evaluations of communication instruction. *Communication Education* 51, no. 3: 225–236.

Felton, J., J. Mitchell, and M. Stinson. 2004. Web-based student evaluations of professors: the relations between perceived quality, easiness and sexiness. *Assessment & Evaluation in Higher Education* 29, no. 1: 91–108.

Foster, A. 2003. Picking apart pick-a-prof: does the popular online service help students find good professors, or just easy A's? *Chronicle of Higher Education* 49, no. 26: A33–A34.
Fowler, F., Jr. 1995. *Improving survey questions: design and evaluation.* Applied Social Research Methods Series, Vol. 38 Thousand Oaks, CA: Sage Publications.
Greenwald, A.G., and G.M. Gillmore. 1997. Grading leniency is a removable contaminant of student ratings. *American Psychologist* 52: 1209–1217.
Hardy, N. 2003. Online ratings: fact and fiction. *New Directions for Teaching and Learning* Winter (96): 31–38.
Harrison, P.D., D.K. Douglas, and C.A. Burdsal. 2004. The relative merits of different types of overall evaluations of teaching effectiveness. *Research in Higher Education* 45, no. 3: 311–323.
Johnson, T. 2003. Online student ratings: will students respond? *New Directions for Teaching and Learning* Winter (96): 49–59.
Liaw, S., and K. Goh. 2003. Evidence and control of biases in student evaluations of teaching. *International Journal of Educational Management* 17, no. 1: 37–43.
Marsh, H.W., and L.A. Roche. 1997. Making students' evaluations of teaching effectiveness effective: the critical issues of validity, bias, and utility. *American Psychologist* 52, no. 11: 1187–1197.
———. 2000. Effects of grading leniency and low workload on students' evaluations of teaching: popular myths, bias, validity, or innocent bystanders? *Journal of Educational Psychology* 92, no. 1: 202–228.
Martin, J.R. 1998. Evaluating faculty based on student opinions: problems, implications, and recommendations from Deming's theory of management perspective. *Issues in Accounting Education*November: 1079–1094.
McGhee, D., and N. Lowell. 2003. Psychometric properties of student ratings of instruction in online and on-campus courses. *New Directions for Teaching and Learning* Winter (96): 39–48.
McKeachie, W.J. 1997. Student ratings: the validity of use. *American Psychologist* 52: 1218–1225.
Otto, J., D. Sanford, and W. Wagner. 2005. Analysis of online student ratings of university faculty. *Journal of College Teaching & Learning* 2, no. 6: 25–30.
RateMyProfessor. 2006a. *Statistics.* Available online at: http://www.ratemyprofessor.com/index.jsp (accessed August 2006).
RateMyProfessor. 2006b. *Rating categories.* Available online at: http://www.ratemyprofessors.com/categories.jsp. (accessed August 2006).
Read, W., D. Rama, and K. Raghunandan. 2001. The relationship between student evaluations of teaching and faculty evaluations. *Journal of Education for Business* 76, no. 4: 189–193.
Stone, M. 2003. EC Professors fare (somewhat) well on ratemyprofessors.com. *The Leader,* 18 March 2003, available online at: http://www.elmhurst.edu/~leader/archive/2003/cultue_03_18/ratemyproffessor.htm.
Wilson, R. 1998. New research casts doubt on value of student evaluations of professors. *Chronicle of Higher Education* 44, no. 19: A12–A14.

Appendix 1: Sampling methodology

Our sampling procedure was designed to be more representative of the population of faculty members than the sample frame. The sample frame of the faculty members already listed in RateMyProfessor.com includes most of those in North America. According to the Carnegie Foundation, there are 3941 higher education institutions that it classifies, so presumably most of these are included in the 5200 schools listed on the website in October 2005. In our classification of these schools into Carnegie classifications, we were able to find 360 of the 373 schools (96.5%).

Nonetheless, the RateMyProfessor ratings are clustered in certain schools. The 10 schools with the most ratings had ratings for 18,037 faculty, or an average of 1803.7 per school. Data from the schools' websites, however, taken during 2004–2005, indicate an average of 1481.2 full- and part-time faculty at each school. Thus, RateMyProfessor.com data include ratings for more than the number of faculty stated by these universities. This difference may be because universities may not count adjunct, visiting and temporary instructors. Since, in October 2005, there were 600,000 rated faculty in 5200 universities or 115.4 faculty per school (600,000/5200 = 115.4) there must be many schools that have very few ratings.

If we used a random sampling procedure where all faculty rated in RateMyProfessor.com could be chosen with equal probability, then we would have a high proportion of faculty in some schools that are very popular with the website. From the population of faculty, there would be over-sampling for those from these popular schools and under-sampling for the rest. We compensated for this possible bias in the RateMyProfessor sample frame by giving each school an equal chance of being chosen in the first step. We note that our sampling procedure is only ideal if all schools have equal numbers of faculty. But our procedure does have the benefit of reducing selection bias by preserving heterogeneity among institution types, and is not affected by institution-specific biases.