Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 16, Number 18, November 2011 ISSN 1531-7714

Student Consensus on RateMyProfessors.com

April Bleske-Rechek and Amber Fritsch
University of Wisconsin-Eau Claire

At the same time as some faculty committees and corporations are appealing to the use of online ratings from RateMyProfessors.com to inform promotion decisions and nationwide university rankings, others are derogating the site as an unreliable source of idiosyncratic student ratings and commentary. In this paper we describe a study designed to test the assumption that students' ratings are unreliable. The sample included 366 instructors with 10 or more student ratings. Contrary to the assumption that students' ratings are unreliable, variance in students' ratings about a given instructor was similar across number of raters, with 10 raters showing the same degree of consensus as 50 or more raters. Students showed the most consensus about instructors who were among the top third of the distribution in quality, and this effect occurred even among instructors rated as the most difficult. Taken alongside other investigations of RateMyProfessors.com and the broad literature on student evaluations of teaching, our findings suggest that students who use RateMyProfessors.com are likely providing each other with useful information about quality of instruction.

RateMyProfessors.com is an online forum on which students rate and comment on their instructors. The site launched in 1999 and grew rapidly: The site touted millions of ratings on hundreds of thousands of instructors within its first 10 years. Moreover, the site is not without influence. Forbes Magazine has recently included RateMyProfessors ratings to create its rankings of top colleges and universities in the United States (Steinberg, 2009); and anecdotal evidence suggests that faculty tenure and promotion committees have begun to attend to ratings and commentary on RateMyProfessors (Sanders, Walia, Potter, & Linna, 2011). Skepticism about the website is high, however, as revealed by comments such as the following: "Information provided by the RMP website is not valid" (Davison & Price, 2009, p. 61) and "…high-quality ratings may have more to do with an instructor's appearance and how easy he or she makes a course than with the quality of teaching" (Felton, Mitchell, & Stonson, 2004, p. 106). *Should* instructional and administrative staff pay any heed to ratings from students who post on the site? RateMyProfessors.com is certainly unique because students visit the site voluntarily, and only a subset of students who visit the site actually post ratings. In this paper, we briefly review some of the systematic trends documented thus far in analyses of ratings from the site. Then, we describe the results of an investigation that we designed to test the reliability of students' ratings.

RateMyProfessors.com was launched as a public outlet for students to rate and voice commentary on their instructors. On the site, students use five-point Likert-type scales to rate their instructors' *easiness* ('How easy are the classes that this professor teaches?' 'Is it possible to get an A without too much work?'), *helpfulness* ('Is the teacher approachable and nice?' 'Is s/he willing to help you after class?'), and *clarity* ('How well does the teacher convey the class topics?' 'Is s/he clear in his presentation?' 'Is s/he organized and does s/he use class time effectively?'). The elements of helpfulness and clarity are frequently measured in student evaluations of instruction and also show up in reports of university students' perceptions of

effective university instructors (Slate, LaPrairie, Schulte, & Onwuegbuzie, 2011). On RateMyProfessors.com, the helpfulness and clarity scores are averaged to provide a *quality* score for each instructor. Students on the site can also give instructors a "chili pepper" if they find them attractive and can post written commentary about the instructor. As of 2010, RateMyProfessors.com held over 10 million ratings on over a million instructors from the United States, England, and Canada. Although some instructors have only one or two student posts, there are thousands on the site with 10 or more posts (Felton et al., 2004).

Researchers have voiced concerns about the reliability and validity of ratings on RateMyProfessors.com. One concern is that students who post ratings and comments on the site are different from other students (Davison & Price, 2009; Felton, Koper, Mitchell, & Stinson, 2008; Posillico, 2009). Because students go to RateMyProfessors on their own time and not as part of an end-of-semester, class-wide evaluation, it is possible that students who choose to post ratings and commentary are those with extreme views. In potential support of the possibility that students who post are those who have something strongly negative to say, researchers have documented a weak, negative association between instructors' overall mean quality rating and the number of ratings they have received (Davison & Price, 2009); that is, more difficult instructors have more ratings. However, that link is inconsistent (Riniolo, Johnson, Sherman, & Misso, 2006), and when researchers controlled for the number of semesters that instructors have taught since the website launched, thereby creating a measure of rating *frequency*, the association did not replicate (Bleske-Rechek & Michels, 2010).

Other findings suggest that students who choose to post on RateMyProfessors are not different from other students. First, instructors' ratings on RateMyProfessors correlate strongly with the ratings they receive on in-class student evaluations (Coladarci & Kornfield, 2007; Sanders et al., 2011). In Sanders et al.'s (2011) research, the correlation between online score and in-class student evaluation score was .57 even for those instructors with as few as five or six online ratings. Second, students who report that they have posted ratings on the site (between 20 and 30% of students) are similar to those who have not posted, in their grade point average, learning goals, and grade orientations (Bleske-Rechek & Michels, 2010). These trends coincide with research on traditional student evaluations of teaching, which has documented that students' own personality traits do not predict their ratings of their instructors (Patrick, 2011); in fact, students' personality traits do not predict differential response to researcher-manipulated variation in instructor expressiveness. Instead, students of widely varying personality traits consistently view expressive instructors who cover a lot of content more favorably than they view less expressive instructors who cover little content; they also learn more from those instructors (Abrami, Perry, & Leventhal, 1982).

Another concern is that students' ratings are not valid but instead biased, such that they give high quality ratings to easy instructors (Davison & Price, 2009; Felton et al., 2004). This concern stems from various analyses of RateMyProfessors data that have revealed strong, positive associations ($r \sim .4$-$.6$) between ratings of instructor easiness and ratings of instructor quality (Coladarci & Kornfield, 2007; Davison & Price, 2009; Felton et al., 2004). The inference of causal bias from this correlational finding is analogous to that with traditional student evaluations of teaching. On traditional student evaluations of teaching, students' expected grade is typically positively associated with their instructor ratings, and one interpretation of that correlation is that students are biased in favor of instructors who are lenient graders (e.g., Greenwald & Gillmore, 1997) or who give them good grades. In the domain of student evaluations of teaching, however, a massive literature supports other interpretations, including that highly effective instructors facilitate learning (and hence, good grades; Remedios & Lieberman, 2008), as well as the possibility that student factors such as interest, effort, and motivation drive both good grades and perceptions of the instructor as effective (Marsh, 1984; Heckert, Latier, Ringwald-Burton, & Drazen, 2006). Specifically, multi-section validity studies show that

students of instructors with higher ratings score higher on end-of-semester tests (Cohen, 1981; Feldman, 1989); controlled manipulations of instructor expressiveness and content coverage have a large impact on students' ratings of their instructors, whereas manipulations of instructor grading standards have small and inconsistent effects (Abrami, Dickens, Perry, & Leventhal, 1980); and links between students' grades and ratings of their instructor can be accounted for by student interest and perceived amount of learning in the class (Marsh & Roche, 2000; Patrick, 2011). Because RateMyProfessors.com ratings closely coincide with traditional student evaluations of teaching (Coladarci & Kornfield, 2007; Sanders et al., 2011), the consensus that students generally provide useful ratings of instructor quality on traditional student evaluations would imply that students probably do so on RateMyProfessors.com, as well.

The primary objective of the current research is to investigate reliability in students' ratings on RateMyProfessors.com by analyzing variance in their ratings of instructor quality and easiness. On one hand, if students' postings reflect idiosyncratic responses to their instructors, then having a limited number of student raters about a given instructor might limit the interpretability of those ratings. That is, if each student's ratings about an instructor include varied sources of error, then only by aggregating many students' responses should there be consensus about an instructor. According to this logic, the number of ratings should be negatively associated with degree of variance in the ratings – that is, it should take many ratings to get consensus around the mean. On the other hand, if students are consistent in their assessments, then the number of ratings should not be related to variance in ratings – that is, there should be consensus around the mean with just a small number of ratings. It is possible, however, that student consensus could reflect consistency in students' responses to an educationally-irrelevant aspect of their experience with an instructor. We suggest that if consistency in student ratings reflects consistency in students' perceptions of quality, then students should be distinguishing between easiness and quality when making their judgments, and variance in students' perceptions of quality of instruction should be tied more to instructors' mean level of quality than to instructors' mean level of easiness.

METHOD

Sample

We analyzed ratings that are publicly available on the RateMyProfessors.com website. We included all instructors from a large, public university who had 10 or more ratings on the site. The value of 10 was chosen as a cutoff because it corresponded with the cutoff previous researchers have used (Felton et al., 2004); it also allowed for omission of instructor names that may have shown up in error due to misspellings or confusion about instructor status. After omission of six outlier instructors (3*SD* away from the mean), the final dataset included 366 instructors (207 male, 159 female) who had between 10 and 86 ratings. The omitted instructors, who were outliers, had 89 or more ratings. Instructors represented the major disciplines on campus, with 37% of instructors from Arts and Humanities, 18% from Social Sciences, 28% from Math and Natural Sciences, and 17% from Pre-professional majors.

Procedure

We first created a single data set for each instructor. Each instructor's data included helpfulness, clarity, easiness, and quality ratings of each student who had rated that instructor; we computed summary statistics (*M*, *SD*, and variance) for each of those variables. Then, we compiled an umbrella dataset that included all instructors. For each instructor, we noted their sex and discipline, their means, standard deviations, and variances for easiness and quality, the number of student raters that had contributed to those summary statistics (range = 10 to 86, $M = 28.57$, $SD = 16.93$), and the number of years the instructor had been teaching at the university since fall of 1999, when RateMyProfessors was launched (range = 1 to 11, $M = 8.35$, $SD = 3.20$). Because instructors who had been at the university longer had more ratings, $r(366) = .34$, $p < .001$, we computed a weighted index of the frequency with which instructors had received student posts (number of raters/number of years) and labeled it, "rating frequency." Rating

frequencies ranged from 0.91 to 19.00 (M = 4.03, SD = 2.76).

RESULTS

Summary statistics for easiness and quality are displayed in Table 1. The typical instructor had a mean quality rating of 3.58 and a mean easiness rating of 3.14. Consensus around the mean can be described using either variance or its square root, standard deviation.

Table 1: Summary Statistics for Ratings of Instructor Quality and Easiness

	Range	Mean	Std. Dev.
Instructor's Overall Quality Rating	1.41-4.98	3.58	0.84
Variance in Instructor's Quality Ratings	0.01-2.81	1.23	0.62
Standard Deviation in Quality Ratings	0.10-1.68	1.06	0.33
Instructor's Overall Easiness Rating	1.39-4.90	3.14	0.78
Variance in Instructor's Easiness Ratings	0.10-2.65	1.09	0.43
Standard Deviation in Easiness Ratings	-1.06-1.63	1.01	0.24

Although variance is more commonly used to denote degree of consensus, standard deviation values are in the original rating scale units and can provide a benchmark for the degree of overall spread of scores around the mean. For example, the typical instructor had a standard deviation in quality ratings of 1.06, such that 68% of that instructor's ratings were between 2.52 and 4.64 (3.58 +/- 1.06). On the five point scale, then, the majority of ratings for a typical instructor fell within three points of one another (between 2 and 4 for an instructor with an average of 3). Instructors with smaller variance (and standard deviation values) have a narrower spread of measurements around the mean and therefore usually have comparatively fewer high or low values. Notably, instructors with a low mean rating can still have a larger variance if they have several high ratings from individual students; instructors with a high mean rating can also have a larger variance if they have several low ratings from individual students; and instructors with a moderate mean rating can still have a small variance if most of the individual ratings are moderate.

Mean Quality and Easiness Ratings

Figure 1 displays the distribution of mean quality ratings (upper panel) and the distribution of mean easiness ratings (lower panel). As shown in the upper panel, instructors differed widely in how high or low in quality they were rated, but the distribution of means was slightly negatively skewed, such that instructors' mean quality ratings were more positive than negative overall. As shown in the

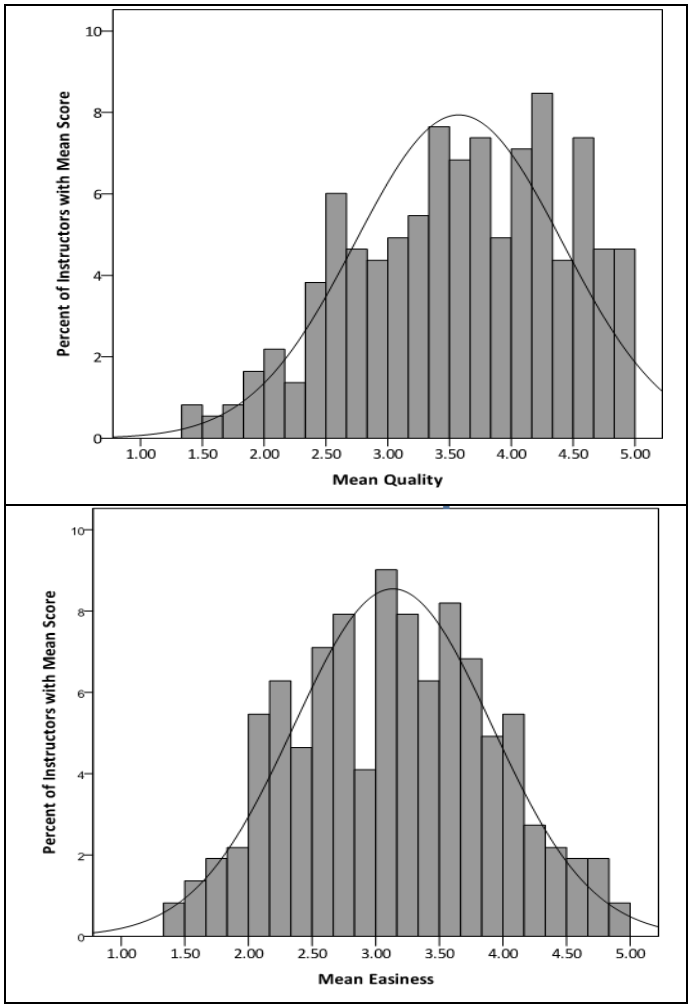

Figure 1. Distribution of mean quality ratings (upper panel) and distribution of mean easiness ratings (lower panel).

bottom panel, instructors differed widely in how low or high in easiness they were rated, and those ratings follow a normal distribution. As in previous studies (e.g., Felton et al., 2004), instructors' easiness and quality ratings were positively correlated, $r(366) = .57$, $p < .001$ (see below). On average, however, instructors were rated as higher in quality than in easiness, paired samples $t(365) = 11.41$, $p < .001$, $d = 0.60$.

Mean Ratings and Rating Frequency

Instructor quality was not associated with number of ratings ($r(365) = -.08$, $p = .105$) or with number of years at the university since RateMyProfessors began ($r(365) = -.05$, $p = .373$). Instructors who were rated as more difficult (less easy) had more student ratings, $r(365) = -.11$, $p = .04$; however, more difficult instructors also had been teaching at the institution for more of the years since RateMyProfessors began, $r(365) = -.13$, $p = .01$. As such, we conducted an additional analysis between easiness and rating frequency (number of ratings weighted by number of years at the university since RateMyProfessors launched). As displayed in the upper panel of Figure 2, instructors with higher quality ratings were not rated any more or less frequently than other instructors were, $r(365) = -.01$, $p = .852$. As shown in the lower panel of Figure 2, instructors rated as easier were not rated any more or less frequently than other instructors were, $r(365) = .03$, $p = .575$.

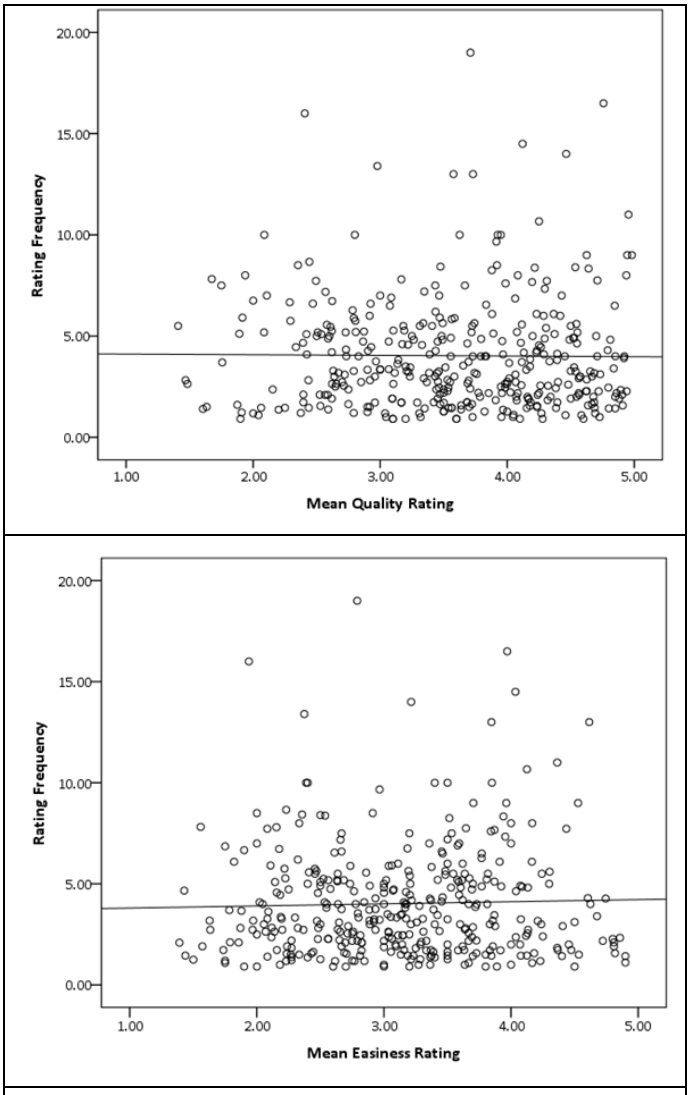

Figure 2. Association between mean quality rating and rating frequency (upper panel), and association between mean easiness rating and rating frequency (lower panel). Instructors' mean quality and easiness ratings did not predict how frequently they were rated.

Consensus and Number of Raters

The primary objective of the current study was to determine the reliability of students' ratings. If students who choose to post ratings on RateMyProfessors.com are at the extremes or are rating according to idiosyncratic perceptions, then variance in student ratings should be high when there are fewer ratings, and student consensus should increase (variance should decrease) with number of student raters. That is, more student raters should aggregate signal of instructor process and filter out students' emotional biases. Figure 3 (upper panel) displays the association between number of quality ratings and degree of consensus in those ratings. Each dot represents a given instructor's number of ratings and degree of variance in those ratings. Counter to concerns about the reliability of student ratings, degree of variance in a given instructor's ratings was not associated with how many students had rated them. In other words, instructors with 10 ratings showed the same degree of consensus in their quality ratings as did instructors with 50 ratings, $r(366) = .03$, $p = .573$. Figure 3 (lower panel) shows the association between rating frequency and degree of consensus in those ratings. Each dot represents the frequency

with which a given instructor was rated and the degree of variance in the ratings received. Rating

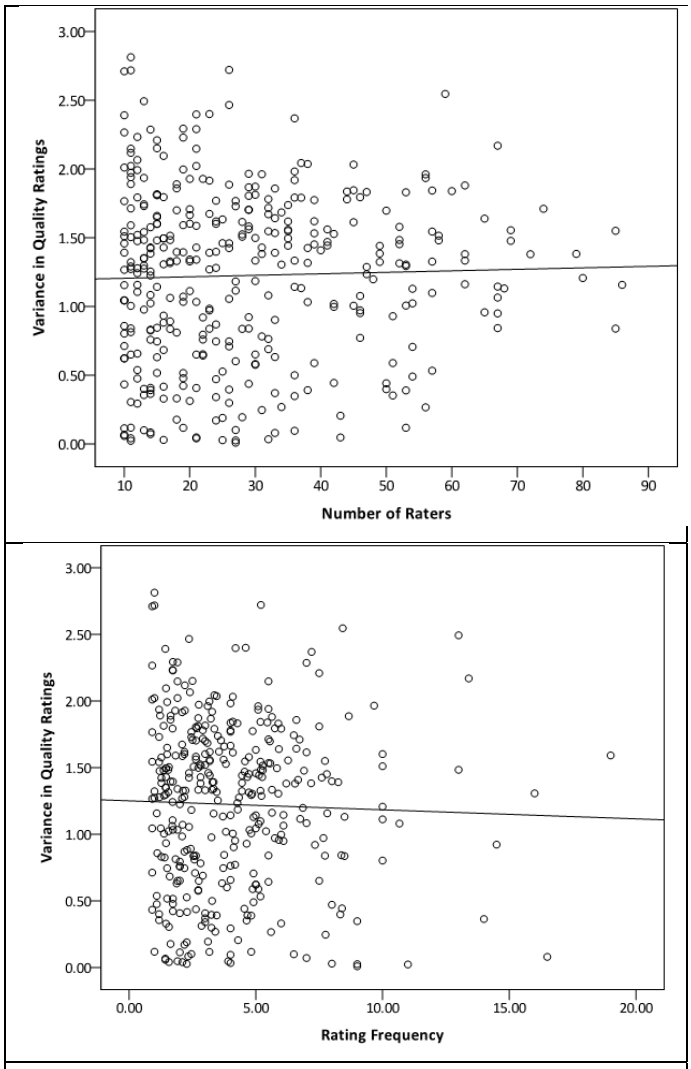

Figure 3. Association between number of raters and degree of consensus around the mean quality rating (upper panel) and association between rating frequency and degree of consensus around the mean quality rating (lower panel). Instructors with fewer ratings had as much consensus in their ratings as did instructors with more ratings.

frequency was not associated with degree of student consensus, $r(365) = -.03$, $p = .566$. Within each sex and within each discipline, number of raters was not associated with degree of variance in instructor quality ratings (all ps > .09, values for r ranged from -.18 to +.15), nor was number of raters (or rating frequency) associated with student consensus when we used standard deviation as the unit of measurement, all ps > .24. These non-significant associations between consensus and number of raters run counter to the suggestion that students are unreliable judges of instructor quality.

Consensus and Quality

Although degree of variance in a given instructor's ratings was not tied to how many students had provided the ratings, degree of variance in instructors' quality ratings was tied to the overall perception of instructors' quality (quadratic $F(2, 364) = 424.86$, $p < .001$, $R^2 = .70$) and degree of variance in instructors' easiness ratings was tied to the overall perception of instructor's easiness (quadratic $F(2, 364) = 66.25$, $p < .001$, $R^2 = .27$). Instructors with low and (especially) high mean ratings had the least variance in the ratings they received. These associations are displayed in the upper and lower panels of Figure 4. The effect was very robust for quality, and also replicated by sex and discipline of instructor, all ps < .001, R^2 values ranging from .66 to .74. Instructors with very high mean quality ratings showed very little variance (i.e., strong consensus) in students' ratings– in some cases essentially no variance at all.

Consensus about Instructor Easiness and Instructor Quality

As noted above, instructors who were rated as easy also received higher quality ratings, $r(366) = .57$, $p < .001$. If instructor easiness leads to high quality ratings, then instructors who are very easy should be consistently rated as high in quality. To look at the competing links of easiness and quality with variance in student ratings, we first used percentile values to place instructors into quality and easiness thirds. Instructors in the bottom third of the distribution for quality had means ranging from 1.41 to 3.20, those in the middle third had means ranging from 3.21 to 4.07, and those in upper third for quality had means ranging from 4.08 to 4.98. Instructors in the bottom third of the distribution for easiness (the most difficult instructors) had means ranging from 1.00 to 2.74, those in the middle third had means ranging from 2.75 to 3.50, and those in upper third for easiness had means ranging from 3.51 to 4.90.

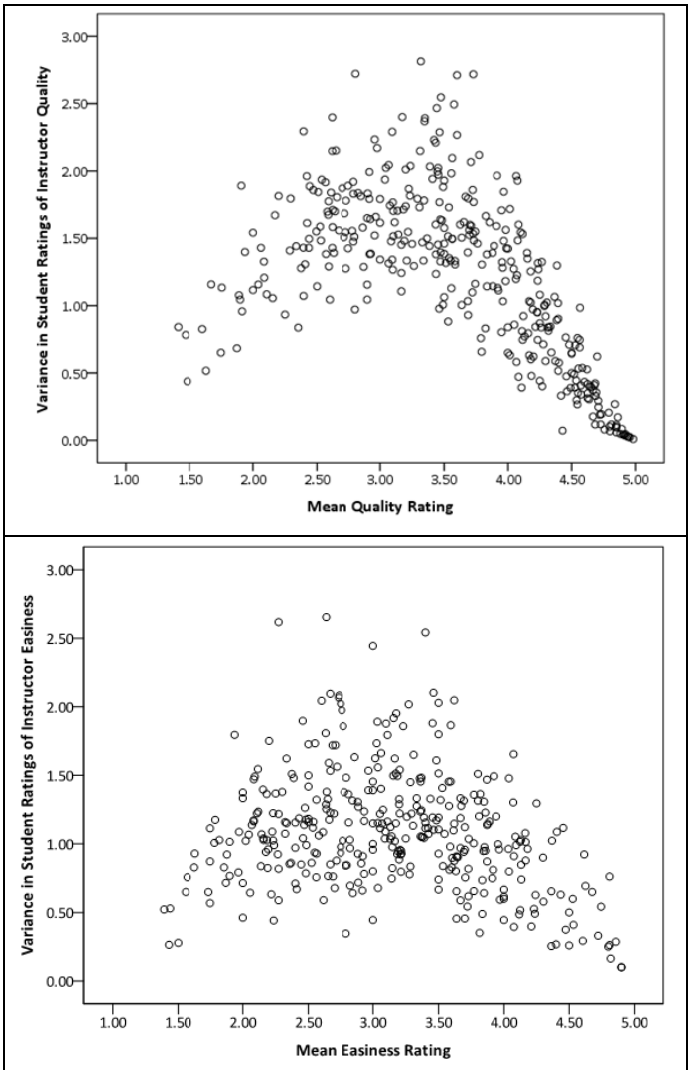

Figure 4. Variance in student ratings of quality, as a function of mean quality rating (upper panel); and variance in student ratings of easiness, as a function of mean easiness rating. Students showed the most consensus about instructors with low mean ratings or high mean ratings.

Our analysis of variance for quality ratings revealed that easy instructors were not rated with any more consistency than moderate or difficult instructors were. As shown in Figure 5, students did not show any more consensus overall about the quality of easy versus less easy instructors, $F(2, 357) = 0.98$, $p = .378$, partial $\eta^2 = .005$. Instead, students showed the most consensus (least variance) about instructors who were among the top third of the distribution in quality, $F(2, 357) = 120.52$, $p < .001$, partial $\eta^2 = .403$. As revealed by the interaction *F*-test, this effect of quality on variance occurred even among instructors rated as the most difficult (Figure 5, left set of columns) and was magnified among instructors rated as easy (Figure 5, right set of columns), $F(4, 357) = 4.48$, $p = .002$, partial $\eta^2 = .05$. Thus, student consensus about instructor quality was not a direct function of how easy the instructor was.

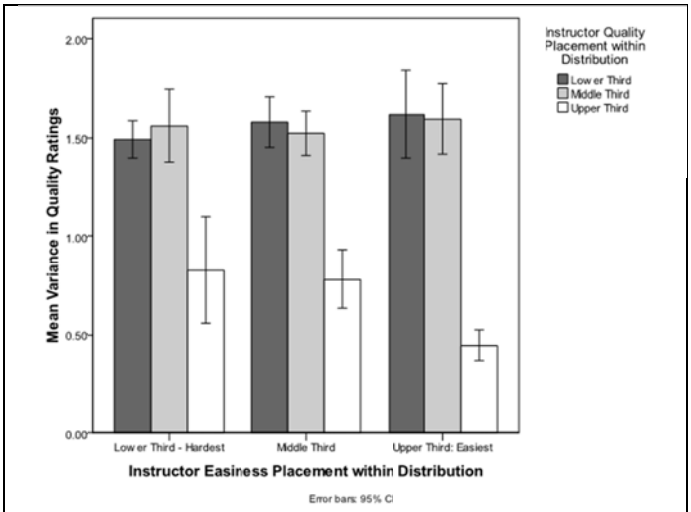

Figure 5. Variance in student ratings of instructor quality, as a function of instructor's mean level of easiness and quality. A total of 122 instructors were in the lower third for easiness (76 lower third in quality, 33 middle third in quality, and 13 upper third in quality), 124 were in the middle third (26 lower third in quality, 60 middle third in quality, and 38 upper third in quality), and 120 were in the upper third for easiness (20 lower third in quality, 29 middle third in quality, and 71 upper third in quality).

We ran one more set of analyses to address the possibility that students are biased in their judgments of instructor quality. We reasoned that there may be comparatively more room for bias in two of the four major disciplines. First, because instructors in Math & Natural Sciences courses are rated as more difficult overall than are instructors in other disciplines (Aleamoni, 1999; Bleske-Rechek & Michels, 2010), students may be more varied in their response because of the more varied grade

distributions that coincide with more students struggling to achieve the desired grades. Second, because courses in the Arts & Humanities may be perceived, at least by students, as having more room for subjectivity in opinion and belief, there may be stronger emotional responses to instructors in those departments. To test these ideas, we analyzed mean ratings and variance in ratings as a function of discipline. As we expected on the basis of previous research, Math & Natural Sciences instructors were rated as less easy than were instructors in each of the other disciplines (Arts & Humanities, Social Sciences, and Pre-professional), $F(3, 363) = 6.21$, $p < .001$, partial $\eta^2 = .05$, all pair-wise ps $\leq .01$. However, instructors in Math & Natural Sciences were rated similarly in quality, $F(3, 363) = 0.21$, $p = .89$, partial $\eta^2 = .002$. This finding implies that students distinguish between easiness and quality in their ratings. In further support of students as reliable judges of quality of instruction, Figure 6 shows that variance in students' ratings of easiness and quality did not differ by discipline. That is, students showed similar consensus about their instructors, regardless of the type of discipline those instructors were in (easiness $F(3, 363) = 0.68$, $p = .56$, partial $\eta^2 = .006$; quality $F(3, 363) = 0.10$, $p = .96$, partial $\eta^2 = .001$).

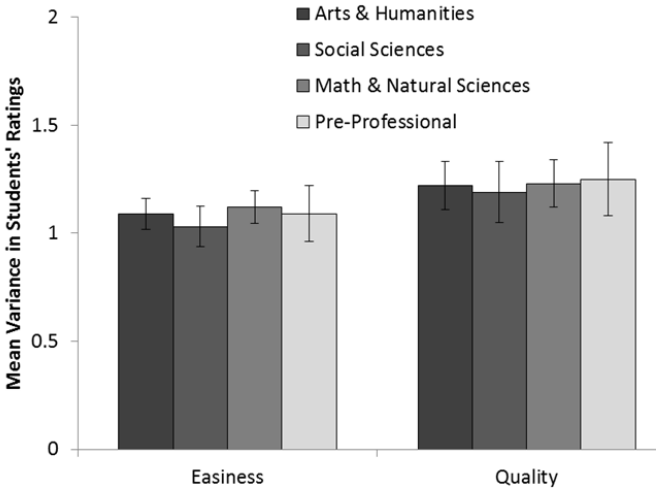

Figure 6. Degree of variance in students' ratings of instructor easiness and quality, by discipline. Although students rated instructors in Math & Natural Sciences as more difficult, on average (see text), they did not rate them as lower in quality and did not demonstrate any more or less consensus about the quality of those instructors.

instructors were rated was not tied to how favorably they were rated, again implying that students are not rating just the instructors with whom they are most displeased or upset.

DISCUSSION

In this study we documented a number of findings that could be used to inform faculty and researchers' judgments about online rating sites such as RateMyProfessors.com. First, the instructors in our sample varied in how many students had taken the time to go online and rate them; some were rated far more frequently than others were. Second, some instructors were rated more favorably than other instructors were, a finding that coincides with Riniolo et al.'s (2006) observation that "ratings are widely dispersed and not just clustered at the extremes on the 5-point student evaluation scale, indicating a wide distribution of input that is not solely targeted at evaluating professors rated as very poor (i.e., motivated to "slam" professors) and outstanding (i.e., motivated to praise professors) or both." (p. 33). Third, the frequency with which

As documented in previous investigations of online rating sites (Silva, Silva, Quinn, Draper, Cover, & Munoff, 2008), we also found that instructors received more favorable than unfavorable ratings, and tended to be rated as higher in quality than in easiness; these trends run contrary to the assumption that students who post are especially displeased or those looking for the easy "A." The finding of more favorable ratings overall is especially important in light of past research showing that students tend to give more negative ratings when they are anonymous (Feldman, 1979), because RateMyProfessors.com is anonymous. Another past study found that students who had posted on RateMyProfessors did not differ from other students in their learning goals, focus on grades over learning, or GPA (Bleske-Rechek & Michels, 2010). The evidence thus far, then, indicates that students who do post do not differ from students who do not post, at least on the

variables that have been measured. Yet, only a minority of students post on RateMyProfessors (Bleske-Rechek & Michels, 2010; Davison & Price, 2009), which implies that students who post must be different somehow from students who do not. Are they more conscientious, or do they feel a stronger sense of obligation to inform their fellow students? If so, why?

Another finding of the current study was a consistent degree of consensus among students about a given instructor's quality, regardless of how many students had provided the ratings. In other words, degree of consensus about an instructor, on average, was the same if that instructor had 10 ratings as if the instructor had 50 ratings. We did not include instructors who had fewer than 10 ratings, and it seems that students and instructors should be cautious in interpreting a small number of posts or any individual post taken on its own. However, our findings suggest that with at least 10 ratings instructors may be able to extract crude judgments -- exceptional, adequate, or unacceptable (McKeachie, 1997) -- of students' perceptions of their clarity and helpfulness.

The current study also documented that students agree about low quality instructors and, in particular, high quality instructors. In fact, differences in mean quality ratings accounted for 70% of the variance in degree of consensus about instructor quality. Moreover, while we replicated previous research with student evaluations of teaching that has shown instructors in math and natural sciences are rated as more difficult, we also showed that students (a) did not rate math and natural sciences instructors as any lower in quality than other instructors and (b) showed as much consensus in their ratings of math and natural sciences instructors as they did about instructors in other disciplines. In the aggregate, students agreed about which instructors were highly effective (in terms of clarity and helpfulness) and which instructors were not, across instructor sex and discipline. This strong student consensus about quality coincides with research showing that manipulating teacher expressiveness and content coverage can have a large impact on students' ratings of instructor effectiveness regardless of students' own personalities (Abrami et al., 1982). Despite faculty doubts about the ability of students to appreciate good teaching, "we now know that students can evaluate teaching effectively" (McKeachie, 1990, p. 197).

Implications and Limitations

The trends documented thus far on data from RateMyProfessors closely parallel those seen in the literature on student evaluations of teaching. Indeed, instructors' student evaluation scores correlate strongly with their online ratings (Coladarci & Kornfield, 2007; Sanders et al., 2011), and in general students who post on RateMyProfessors do not appear to be different from other students in ways relevant to the ratings (such as learning goals; Bleske-Rechek & Michels, 2010). However, the RateMyProfessors website has plenty of room for improvement, and we recommend cautious interpretation of any single instructor's ratings. As with student evaluations of instruction, there is the concern of misuse of RateMyProfessors ratings. Administrators should not make decisions about faculty development and promotions on the basis of these ratings. Instead, as with any personnel decision, multiple sources of information must contribute.

The findings of the current study are potentially limited by several factors. First, we chose to analyze all instructors (with 10 or more ratings) from one institution rather than a few instructors from many institutions. In this way, we could guarantee access to records of instructors' number of years at the university since RateMyProfessors began and we could easily access instructors from a variety of disciplines. Although future research ideally would expand our analyses to instructors of multiple institutions, our university is a typical, four-year liberal arts based institution. Moreover, the distributions of mean ratings and the associations we documented between mean quality and mean easiness ratings coincided with those of other studies (e.g., Felton et al., 2004; Silva et al., 2008) and thus imply our findings are robust.

Other concerns about our data stem from the nature of the RateMyProfessors.com website. Student ratings on the site are not entirely

independent, and we have no way of distinguishing independent ratings from dependent ratings. For example, a single student may rate more than one instructor. A single student may also rate the same instructor more than once, but for different courses. Each instructor receives a single mean quality (and easiness) rating, yet for some instructors that composite may come from ratings of a single course and for others that composite may come from ratings of five or more different courses. It is likely that these related concerns are relatively minor, because research on student evaluations of teaching suggests that students in the same course with two different instructors differ widely, on average, in their ratings, whereas students enrolled in two different sections of the same instructor's course or in two different courses provided by the same instructor provide very similar ratings (Marsh, 1984). However, we recognize that if we knew which students were rating which instructors, we could have computed interrater reliability coefficients as measures of student consensus. Instead, we proceeded from a statistically conservative assumption that different students rated each instructor (instead of that the same group of students rated the same set of instructors).

CONCLUSION

RateMyProfessors.com is widely used by students. Its growth has been accompanied by skepticism, but the site sits on the wave of social networking and likely will expand in use and influence. If faculty do not want students to use the site, one option is to make the results of traditional student evaluations of instruction easily accessible – i.e., public – so students have no apparent need for the site (Felton et al., 2004). However, analyses thus far suggest that RateMyProfessors is an easily accessible public forum on which students are providing each other with not just emotional reactions but serious information about instructor process (Bleske-Rechek & Michels, 2010; Otto, Sanford, & Ross, 2008; Silva et al., 2008). Thus, another option may be to work with RateMyProfessors administrators to include more questions that are obviously linked to instructor pedagogy (Felton et al., 2004). For example, the site could ask students about their perceptions of instructors' use of class time, provision of opportunities to master course material, use of assignments and assessments that are tied to course objectives, and use of objective grading criteria. Relatedly, another option may be to work with RateMyProfessors.com administrators to remove questions that detract from the face validity of the site. For example, although it is possible that attractiveness is tied to student learning (e.g., via student attention) or to desirable instructor personality traits (e.g., confidence, intellect) that facilitate effective instruction, public display of the chili pepper and students' comments about instructors' personality and appearance will probably continue to cause skepticism that will outweigh any statistical evidence of the site's utility. Our study has added to that statistical evidence. We demonstrated strong student consensus about instructor quality, which did not hinge on instructor easiness. Trends in student ratings on RateMyProfessors mirror those found on traditional student evaluations of teaching (Coladarci & Kornfield, 2007; Sanders et al., 2011). In the aggregate, RateMyProfessors.com is providing useful feedback about instructor quality.

REFERENCES

Abrami, P. C., Dickens, W. J., Perry, R. P., and Leventhal, L. (1980). Do teacher standards for assigning grades affect student evaluations of instruction? *Journal of Educational Psychology, 72:* 107-18.

Abrami, P. C., Perry, R. P., and Leventhal, L. (1982). The relationship between student personality characteristics, teacher ratings, and student achievement. *Journal of Educational Psychology, 74:* 111-25.

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13:* 153-66.

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

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of

multisection validity studies. *Review of Educational Research, 51:* 281-309.

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

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

Feldman, K. A. (1979). The significance of circumstances for college students' ratings of their teachers and courses. *Research in Higher Education, 10:* 149-72.

Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement. *Research in Higher Education, 30:* 583-645.

Felton, J., Koper, P. T., Mitchell, J., and Stinson, M. (2008). Attractiveness, easiness, and other issues: Student evaluations of professors on ratemyprofessors.com. *Assessment & Evaluation in Higher Education, 33:* 45-61.

Felton, J., Mitchell, J., and Stinson, M. (2004). Web-based student evaluations of professors: The relationships between perceived quality, easiness, and sexiness. *Assessment & Evaluation in Higher Education, 29:* 91-108.

Greenwald, A. G., and Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52:* 1209-17.

Heckert, T. M., Latier, A., Ringwald-Burton, A., and Drazen, C. (2006). Relations among student effort, perceived class difficulty appropriateness, and student evaluations of teaching: Is it possible to 'buy' better evaluations through lenient learning? *College Student Journal, 40:* 588-96.

Kindred, J., and Mohammed, S. (2005). "He will crush you like an academic ninja!": Exploring teacher ratings on Ratemyprofessors.com. *Journal of Computer-Mediated Communication, 10,3,* article 9: http://jcmc.indiana.edu/vol10/issue3/kindred.html

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76:* 707-54.

Marsh, H. W., and Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology, 92:* 202-28.

McKeachie, W. J. (1990). Research on college teaching: The historical background. *Journal of Educational Psychology, 82:* 189-200.

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52:* 1218-25.

Otto, J., Sanford, D. A., and Ross, D. N. (2008). Does ratemyprofessor.com really rate my professor? *Assessment & Evaluation in Higher Education, 33:* 355-68.

Patrick, C. L. (2011). Student evaluations of teaching: Effects of the Big Five personality traits, grades, and the validity hypothesis. *Assessment & Evaluation in Higher Education, 36:* 239-49.

Posillico, F. (2009, April 1). *Rate my professor's...changing scheduling decisions?*. Retrieved from http://www.sbstatesman.com/2.873/rate-my-professors-changing-scheduling-decisions 1.35665 (Stony Brook, NY).

Remedios, R., and Lieberman, D. A. (2008). I liked your course because you taught me well: The influence of grades, workload, expectations and goals on students' evaluations of teaching. *British Educational Research Journal, 34:* 91-115.

Riniolo, T.C., Johnson, K. C., Sherman, T. R., and Misso, J. A. (2006). Hot or not: Do professors perceived as physically attractive receive higher student evaluations? *The Journal of General Psychology, 133,* 19-35.

Sanders, S., Walia, B., Potter, J., and Linna, K. W. (2011). Do more online instructional ratings lead to better prediction of instructor quality? *Practical Assessment, Research & Evaluation, 16(2):* 1-6. Retrieved from: http://pareonline.net/getvn.asp?v=16&n=2

Silva, K. M., Silva, F. J., Quinn, M. A., Draper, J. N., Cover, K. R., and Munoff, A. A. (2008). Rate my professor: Online evaluations of psychology instructors. *Teaching of Psychology, 35:* 71-80.

Slate, J. R., LaPrairie, K. N., Schulte, D. P., and Onwuegbuzie, A. J. (2011). Views of effective faculty: A mixed analysis. *Assessment & Evaluation in Higher Education, 36:* 331-346.

Steinberg, J. (2009, August 6). *As Forbes sees it, West Point beats Princeton (and Harvard, too)*. Retrieved from http://thechoice.blogs.nytimes.com/2009/08/06/westpoint/

Citation:

Bleske-Rechek, April & Amber Fritsch (2011). Student Consensus on RateMyProfessors.com. *Practical Assessment, Research & Evaluation*, 16(18). Available online: http://pareonline.net/getvn.asp?v=16&n=18

Authors:

April Bleske-Rechek,Associate Professor of Psychology
University of Wisconsin-Eau Claire
105 Garfield Avenue
Eau Claire, WI 54702-4004
bleskeal [at] uwec.edu
www.uwec.edu/psyc/Who/bleske-rechek.htm

Amber Fritsch, Student
University of Wisconsin-Eau Claire
105 Garfield Avenue
Eau Claire, WI 54702-4004
fritscar [at] uwec.edu