Assessing faculty performance using student evaluations of teaching in an uncontrolled setting
Clifford Nowell*, Lewis R. Gale and Bruce Handley

John B. Goddard School of Business and Economics, Weber State University, Utah, USA

This paper provides new evidence on the disparity between student evaluation of teaching (SET) ratings when evaluations are conducted online versus in-class. Using a multiple regression analysis, we show that after controlling for many of the class and student characteristics not under the direct control of the instructor, average SET ratings from evaluations conducted online are significantly lower than average SET ratings conducted in-class. Further, we demonstrate the importance of controlling for the factors not under the instructor’s control when using SET ratings to evaluate faculty performance in the classroom. We do not suggest that moving to online evaluation is overly problematic, only that it is difficult to compare evaluations done online with evaluations done in-class. While we do not suppose that one method is ‘more accurate’ than another, we do believe that institutions would benefit from either moving all evaluations online or by continuing to do all evaluations in-class.

Keywords: online evaluations; multiple regression

Introduction
For more than 75 years, academicians have been examining student evaluation of instructor performance, and after thousands of pages have been written on the subject, the usefulness of the data collected is still debated. Almost all institutions of higher education collect data on student perception of faculty performance. These data are used for different purposes, both evaluative and developmental. Clearly, significant agreement exists among researchers as to what makes a useful teaching evaluation. Students and faculty, however, have concerns regarding how the data are collected, how the data are used and ultimately the value of the process (Shao, Anderson, and Newsome 2007). Whereas researchers often focus on the consistency of data collected in large studies, faculty members focus on how student evaluations relate to them individually. Perhaps it is this disconnect between the broad conclusions of research and the individual application to instructors that keeps this debate alive. What may be most surprising about research into the data collected through student evaluation of teaching (SET) is how persistent faculty members are in expressing their doubts about the value of the exercise.

With this paper we hope to add to the discussion on the usefulness of teaching evaluations in two ways. First, because a growing number of universities are conducting student evaluations online rather than in-class, we analyse the impact this approach may have on student evaluations. This is not a new undertaking, indeed many
researchers (Layne, DeCristoforo, and McGinty 1999; Dommeyer et al. 2002, 2004) have looked at the differences between the ratings of in-class and online evaluations. With a few exceptions, significant differences have not been found. Our contribution centres on how the data are analysed. Most analyses that have been conducted to date fail to control for many of the variables which have been found to impact instructor evaluations, such as expected grades, class size and level of class (Krautmann and Sander 1999). We show that although a simple test of means does not reveal differences between online and in-class evaluations, significant differences between online and in-class evaluations do emerge after controlling for instructor and class characteristics.

Our second contribution goes to the heart of faculty hesitancy regarding the evaluation process. We show how student evaluations of faculty can be adjusted so that only the factors under the instructor’s control are considered in evaluation of faculty members. This is important because substantial evidence suggests that students evaluate faculty based on a myriad of factors, not all of which are controlled by the faculty (Isley and Singh 2005). Resistance to administering teaching evaluations online rather than in-class likely arises from the fact that this new method adds one more variable not under control of the faculty into the overall process. In order to gain increased acceptance of the use of teaching evaluations by faculty, we must explicitly recognise, and account for the factors which are not under the control of the instructor, but are used by students to evaluate faculty.

Our paper has six sections. Immediately following this introduction we present a brief literature review summarising the benefits and concerns of evaluating faculty online. The third section describes the methods we used to analyse data. Data are presented in the fourth section, and a statistical analysis is presented in the fifth section. Finally, the last section presents conclusions.

Background

Many factors influence how students evaluate their instructors. These factors may generally be broken down into three components: student characteristics, instructor attributes and institutional and classroom settings. Past research has shown that some of the most important student characteristics that contribute to predicting student evaluations include expected grades, individual grade point average and student major (Nowell 2007). Classroom characteristics include time of day, class size, subject matter and class level (Millea and Grimes 2002). Of course, instructor organisation, clarity, availability and knowledge of subject (Marsh and Roche 1999) have also been demonstrated to be important. The difficulty in using student ratings to evaluate instructor performance is that instructors control only a few of these items. When conducting faculty evaluations, it would be appropriate for administrators to control for differences not directly under faculty members’ direct influence.

Recently a debate has been taking place on whether the method of administering faculty evaluations is one of the institutional settings that influence student ratings of faculty. Two questions have been paramount: the first question is whether mean ratings of faculty performance differ when using online and in-class surveys. The second question is how the lower response rate of online surveys impacts the evaluation process. Most findings regarding the comparability of online and in-class evaluations of faculty generally identify few significant differences in mean ratings (Layne, DeCristoforo, and McGinty 1999; Dommeyer et al. 2004; Avery et al. 2006;
Ardalan et al. 2007), although a few exceptions do exist (Kulik 2005). These studies present valuable evidence on the comparability of online and face-to-face evaluation methods. Still, all of these studies were conducted with either a one-way analysis or by controlling for very limited factors. None of these studies control for many of the well-documented factors that instructors do not directly control. As to the question of response rates, it is clear that when evaluations are conducted online, response rates are lower (Sax, Gilmartin, and Bryant 2003).

Based in part on the findings that mean ratings do not differ between online and in-class evaluations, the use of online evaluations has grown rapidly. Anderson, Brown, and Spaeth (2006) reported that the percentage of US universities conducting evaluations online increased from 2% to 33% between the years 2000 and 2005. This growth reflects the many advantages of online evaluations. Dommeyer et al. (2004) note that significant cost savings are achieved when moving to online evaluations, but perhaps more importantly, online evaluations may reduce some of the influence that faculty have over in-class evaluations. Standardisation of the evaluation process may reduce the variation of responses due to who administers the evaluations in class, and the activities that take place in class just prior to the evaluation.

Despite the growing popularity of online evaluations, faculty members appear to be hesitant about the idea. This hesitancy stems from the lower response rates (Sax, Gilmartin, and Bryant 2003), and the concern that lower response rates have the potential to create bias if the students who complete the online evaluations are significantly different from the students who are enrolled in the course. When evaluations are conducted in-class, the cost of completing evaluations for students in terms of time and effort is almost zero. When evaluations are moved online, students must take additional time out of the classroom to log in and complete the evaluation. Students will only complete the evaluation when they feel the benefits of completion outweigh the time and effort it takes to complete the evaluation. As a result, one would predict that students who have strong positive or negative feelings towards the professor will be more likely to complete the survey instrument. Based on this, we hypothesise that although differences in the mean ratings are not predictable, the variance of responses in the data collected online will be greater than the variance of data collected in-class, and the sample size will be smaller.

The increased variance of individual responses and the decreased sample size are extremely important when student evaluations are used for the evaluation of faculty. In order to differentiate faculty based on their teaching, small differences in overall ratings may be explained by differences in the quality of instruction, but may also be due to the sample variation in the data collected. Generally if we focus on statistically significant differences in average evaluations, we are focussing on differences due to factors other than sampling variation. Online evaluations increase the variance of individual responses and decrease the sample size, both of which increase the variance of the sample mean. Both of these make it much more difficult to find significant differences between sample means. The predictable result of moving to online evaluations, therefore, is that it will become more difficult to use numerical evaluations to differentiate instructors based on quality of instruction. The error component of the evaluations will increase relative to the component explained by teaching quality. This is why controlling for as many possible sources of variation is paramount. By controlling for multiple sources of variation, the error variance is reduced, making it less difficult to focus on statistically different differences.
Methodology
During the last week of the spring semester 2008, we asked students at a large public university for information about their classroom experiences. The instructor evaluation was conducted in 28 separate courses, taught by 15 different faculty members in economics, finance and quantitative analysis. A faculty member went to each class and read a set of instructions on how the evaluation would be conducted. Approximately half of the classes were told that their evaluation would be done online and were given a URL which linked to the evaluation. In order to increase response rates every student was sent three reminders to his or her email address explaining how to complete the evaluation. The message contained a link to the actual evaluation. In classes where the evaluation was conducted in-class, the identical instructions were given and the exact same form was used.

Our data differ from much of the published work in this area because our unit of observation is the individual student rather than the individual class. We sampled 28 individual classes with 562 enrolled students. Online evaluations were done in 15 classes, with 288 enrolled students, and in-class evaluations were done in 13 classes with 274 enrolled students. Similar to other published research, our response rate using online evaluations was significantly less than the response rate when conducting evaluations in-class. The overall response rate for evaluations conducted online was 28.1% and for evaluations conducted in-class it was 72.2%. These response rates are well within the bounds of prior studies reported by Nulty (2008). The low response rate in the online courses obviously limits our sample sizes in some courses and for some instructors. Detecting significant differences can be a challenge when sample sizes are small. Still, compared with prior published research in this area (Dommeyer et al. 2002; Watt et al. 2002) our sample size of 279 students is certainly adequate.

The first page of the instructor evaluation asked students about five dimensions of instructor quality: organisation, willingness to respond to students, availability, respect for students and the overall contribution of the instructor. Students ranked their instructors on a scale of 1 (low) to 7 (high) in each category. The second page of the evaluation form asked students to rate the quality of their instructor on a scale of 1 (low) to 5 (high). We analyse the data in two different ways. First, similar to McPherson (2006), we used the average of the five responses (SET) as the measure of the students’ satisfaction with their instructor. We use linear regression to analyse these data. Second, we estimate the models using only the responses to the question regarding the overall quality of the instructor using an ordered logit regression suitable for these types of data (Greene 2003). We found little differences between the two types of analysis.

Data
Based on prior research into the factors that influence SET ratings, we collected data on average expected grades (EXGRADE), class size (SIZE) and level of class (LEVEL). The variable LEVEL was coded so that a value of 0 indicates a lower division class and a value of 1 indicates an upper division class. Lower division classes are taken by freshman and sophomore students, whereas upper division classes are taken by junior and senior students. We also collected information on the number of times the class met per week: once, twice or thrice (MEET), and the subject matter taught: economics, finance or quantitative analysis. We used two dummy variables,
one for economics (ECON) and one for finance (FIN). Courses taught in quantitative analysis were used as the baseline comparison group. The variables ECON and FIN take on a value of 1 if the class was in the identified field and 0 otherwise.

In order to control for instructor characteristics, we use two approaches. First we use two broad, albeit incomplete, measures of faculty characteristics. This allows us to indicate whether the faculty member is a permanent employee of the university or is an adjunct faculty member. The variable ADJUNCT takes on a value of 1 if the faculty member is part-time faculty and a value of 0 if the faculty member is a permanent employee. The second measure is the number of years the faculty member has been teaching at the university where the study was conducted (YEARS).

We recognise that these two measures are inadequate to describe the multiple dimensions of instructor attributes. A far better approach would have been to gather data on a wider array of experience measures, information on teaching styles and many other inputs that more accurately describe individual variations in teaching. Because this is likely an impossible task, we perform a separate analysis which simply uses indicator variable for each of the 15 faculty members involved in the study. Rather than trying to measure all the inputs into teaching, we simply control for the style of each instructor by including variables to reflect the 15 instructors in the study. We use the following notation: T1 stands for Instructor 1, T2 stands for Instructor 2 and so forth till T15 for Instructor 15. Although this approach does have certain advantages, disadvantages also arise using these indicator variables. One disadvantage of this approach is that it limits the number of observations for each faculty member. Combining our average response rate of approximately 50%, with the fact that upper division classes typically enrol only 5–10 students, the number of observations for some of the faculty will be very small. Although this does make it difficult to search for significant differences between the faculty, it does not lessen the value of results when significance is found.

We use these variables to explain two alternative measures of teaching effectiveness: the average of the five instructor characteristics (SET1), and the overall instructor evaluation (SET2). Means and variances of all data collected are given in Table 1. This table gives information on both quantitative variables, such as SET1 and SET2, and categorical variables, such as ONLINE, LEVEL and T2 through T15. The mean of the categorical variable ONLINE is 0.30 or 30%. This indicates that 30% of the 279 observations are from online evaluations. The mean of T15 is 0.043 or 4.3%. This indicates that 4.3% of the data comes from Instructor 15.

Analysis

Prior to a detailed analysis of the data we perform three simple tests on the dependent variables. Using SET1 we test if the mean scores from the online evaluations differ from the mean scores from in-class evaluations. The mean of SET1 for evaluations conducted in-class was 5.71 with a variance of 1.23. The mean of SET1 for online evaluations was 5.46 with a variance of 1.80. Testing $H_0: U_{\text{online}} = U_{\text{in-class}}$ against a two-sided alternative with a simple $t$-test yields a test statistic of $t = 1.47$ which is not significant at $\alpha = 0.10$. In this test, where we assume the only explanation for difference in average SET rankings is due to the method of evaluation, we fail to reject $H_0$ and conclude no differences exist between online and in-class evaluations. These findings are similar to the vast majority of past studies on this topic.
The second test we perform is a test for equality of variances on SET1. We hypothesised that variance of the evaluations conducted online would be larger than the variances of the evaluations conducted in-class. Using an $F$-test for the equality of variances we test $H_0: \sigma^2_{\text{online}} = \sigma^2_{\text{in-class}}$, once again using a two-sided alternative. The test statistic is $F = 1.46$, which is significant at $\alpha < 0.05$. As expected, the variance of evaluations collected in an online setting is greater than the variance of evaluations collected in an in-class setting.

The third test we conduct at the onset is a test of independence using the variable SET2. The purpose of this test is to examine whether the distribution of responses to the question asking students to evaluate the overall quality of their instructor on a one to five scale is independent of method of evaluation: in-class and online. This test is conducted with the $\chi^2$ distribution. The null hypothesis of the test is that responses and method of data collection are independent. Table 2 presents the data used for the test. The data show that regardless of the method of data collection, a large percent of respondents evaluate their instructors favourably. In the face-to-face evaluations 76% of students give their instructors a score of 4 or 5 out of a possible 5, and in the online evaluations, 74% of students give their instructors a grade of 4 or 5. As would be expected from the data, the resultant $\chi^2$ test statistic of 1.211 is not significant at $\alpha = 0.10$.
This preliminary data analysis leads one to believe that the method of evaluation does not influence average SET ratings or the overall distribution of SET ratings. Only the variance of responses is significantly impacted. Although this makes it more difficult to distinguish meaningful SET rankings between faculty, we are still left with the conclusion that no bias is created through the method of administering evaluations, online or face-to-face.

We now turn to the more complicated question of explaining SET ratings using regression analysis, which lets us control for many of the factors that influence SET ratings. We first analyze average teaching evaluations using student, class and instructor characteristics. We use the data in Table 1 to estimate the following equation:

\[ \text{SET1} = B_0 + B_1 \cdot \text{ONLINE} + B_2 \cdot \text{EXGRADE} + B_3 \cdot \text{SIZE} + B_4 \cdot \text{LEVEL} + B_5 \cdot \text{MEET} + B_6 \cdot \text{ECON} + B_7 \cdot \text{FIN} + B_8 \cdot \text{ADJUNCT} + B_9 \cdot \text{YEARS} + U \]

where \( U \) is the error term, \( B_i \) are the coefficients estimated in the regression analysis for each of the explanatory variables defined previously. Estimated coefficients represent the predicted change in the average SET rating for a one-unit change in the explanatory variable when all other explanatory variables are held constant. The estimated coefficients and their associated \( t \)-statistics are presented in Columns 2 and 3 of Table 3.

We analyze a second equation where, rather than attempting to control for instructor characteristics, we use an indicator variable for each of the instructors. We randomly omitted a single instructor so that all estimated coefficients are interpreted as the difference between Instructor 1 and the instructor identified in the table. We estimate the following equation:

\[ \text{SET1} = B_0 + B_1 \cdot \text{ONLINE} + B_2 \cdot \text{EXGRADE} + B_3 \cdot \text{SIZE} + B_4 \cdot \text{LEVEL} + B_5 \cdot \text{MEET} + B_6 \cdot T2 + B_7 \cdot T3 + B_8 \cdot T4 + B_9 \cdot T5 + B_{10} \cdot T6 + B_{11} \cdot T7 + B_{12} \cdot T8 + B_{13} \cdot T9 + B_{14} \cdot T10 + B_{15} \cdot T11 + B_{16} \cdot T12 + B_{17} \cdot T13 + B_{18} \cdot T14 + B_{19} \cdot T15 + U \]

As with the prior equation, \( U \) is the error term, \( B_i \) are the coefficients estimated in the regression analysis for each of the explanatory variables. Estimated regression coefficients are interpreted in the same manner as in the first equation. Results from this equation are presented in the fourth and fifth columns of Table 3.

<table>
<thead>
<tr>
<th>Response</th>
<th>Face-to-face evaluations</th>
<th>Online evaluations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ineffective</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Less than effective</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Satisfactory</td>
<td>36</td>
<td>14</td>
</tr>
<tr>
<td>Effective</td>
<td>85</td>
<td>36</td>
</tr>
<tr>
<td>Very effective</td>
<td>66</td>
<td>24</td>
</tr>
</tbody>
</table>

Table 2. Student responses by method of survey.
As expected, all else being equal, grades and student SET ratings are positively correlated. Students do reward instructors for higher grades. Controlling for instructor and student characteristics, the students who completed evaluations online gave their instructors significantly lower ratings as compared with students who completed their evaluations in-class. Ratings for classes where the evaluation was conducted online are predicted to be 0.69 points lower in the equation using instructor characteristics and 0.81 points lower in the equation identifying instructors individually.

As class size increased, all else being equal, evaluations significantly declined. The impact of class level on SET ratings is not clear. When we fail to completely control for the individual instructor differences as shown in Columns 2 and 3 of Table 3, upper division classes are rated significantly higher, but as shown in Columns 4 and 5 of Table 3, that difference is not significant when we control for individual instructors. The same result is found for the relationship between the number of times a class meets per week and SET rankings. When we do not control for individual faculty differences, a significant negative correlation exists between weekly class frequency

<table>
<thead>
<tr>
<th>Variable</th>
<th>Equation 1</th>
<th>Equation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimated coefficient</td>
<td>t-ratio</td>
</tr>
<tr>
<td>Constant</td>
<td>3.02**</td>
<td>4.75</td>
</tr>
<tr>
<td>ONLINE</td>
<td>-0.69**</td>
<td>-4.04</td>
</tr>
<tr>
<td>EXGRADE</td>
<td>0.58**</td>
<td>7.19</td>
</tr>
<tr>
<td>SIZE</td>
<td>-0.02**</td>
<td>-2.06</td>
</tr>
<tr>
<td>LEVEL</td>
<td>0.48**</td>
<td>2.09</td>
</tr>
<tr>
<td>MEET</td>
<td>0.24*</td>
<td>1.93</td>
</tr>
<tr>
<td>ECON</td>
<td>0.26</td>
<td>0.99</td>
</tr>
<tr>
<td>FIN</td>
<td>-0.21</td>
<td>-0.91</td>
</tr>
<tr>
<td>ADJUNCT</td>
<td>0.12</td>
<td>0.58</td>
</tr>
<tr>
<td>YEARS</td>
<td>0.04**</td>
<td>2.22</td>
</tr>
<tr>
<td>T2</td>
<td>-0.63</td>
<td>-0.79</td>
</tr>
<tr>
<td>T3</td>
<td>-1.48**</td>
<td>-3.04</td>
</tr>
<tr>
<td>T4</td>
<td>-0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td>T5</td>
<td>-0.55</td>
<td>-0.96</td>
</tr>
<tr>
<td>T6</td>
<td>-1.16</td>
<td>-1.1</td>
</tr>
<tr>
<td>T7</td>
<td>-0.40</td>
<td>-0.55</td>
</tr>
<tr>
<td>T8</td>
<td>-0.65</td>
<td>-1.48</td>
</tr>
<tr>
<td>T9</td>
<td>-1.33</td>
<td>-1.44</td>
</tr>
<tr>
<td>T10</td>
<td>-0.05</td>
<td>-0.12</td>
</tr>
<tr>
<td>T11</td>
<td>-0.94**</td>
<td>-1.96</td>
</tr>
<tr>
<td>T12</td>
<td>-1.25**</td>
<td>-2.06</td>
</tr>
<tr>
<td>T13</td>
<td>-1.03</td>
<td>-1.25</td>
</tr>
<tr>
<td>T14</td>
<td>-0.51</td>
<td>-0.80</td>
</tr>
<tr>
<td>T15</td>
<td>-0.96</td>
<td>-1.29</td>
</tr>
</tbody>
</table>

Significance level Chi-squared $p < 0.001$

Notes: *significant at the 10% level in a two-tailed t-test; **significant at the 5% level in a two-tailed t-test.
and SET ratings. This difference becomes insignificant as we control for individual faculty differences.

In the fourth and fifth columns of Table 3, where we control for faculty differences, each coefficient measures the average difference between the faculty member and Faculty member number 1, who had the highest predicted SET ratings. The coefficient on the variable T3 equals 1.48 indicating that on a scale of 0–7, Instructor 3 had average scores 1.48 points lower than Instructor 1. Notice that there are only three instructors (T3, T11 and T12) that have significantly lower scores than the top rated instructor, T1.

Now we turn to the analysis of SET2 using ordered logit. Remember, SET2 is measured on a scale of 0–4, where increasing SET ratings correspond to higher teaching ratings. We estimate these equations using the same explanatory variables as we did in the linear regression equation, and present the results in Table 4. Estimated coefficients in the ordered logit model cannot be interpreted in a similar manner as in the multiple regression equation. Estimated coefficients do not represent the change in the dependent variable for a one-unit change in the explanatory variable; however, it is generally the case that the sign of the estimated coefficient indicates the direction of the change in the dependent variable as the explanatory variable changes.

The results in Table 4 are similar to the results presented in Table 3. Expected grades and years spent teaching are directly related to higher SET rankings. Class size and having evaluations administered online are negatively related to SET ratings. Regardless of exactly how we measure student ratings of teaching, online evaluations result in significantly lower SET ratings than when evaluations are done in-class.

The reason why, holding other variables constant, SET rankings are lower when they are conducted online is not answered with this analysis. At least two explanations deserve further investigation. First, it may be possible that when additional time is required outside of class to complete the evaluation, students who are less satisfied are more likely to complete the evaluation. Second, it is also possible that the method of evaluation may actually influence how students complete the evaluation. An interesting area for future research would be to compare in-class paper evaluations with in-class online evaluations to determine if such differences exist.

**Conclusion**

Faculty are concerned about how SET ratings are used for evaluation purposes. Suspicions exist as to how accurately mean rankings measure teaching quality. Many instructors feel these ratings reflect many other factors besides the quality of their efforts. Instructors are correct in supposing that a multitude of factors determine SET ratings, and that their actual performance is only one of those factors. The multiple regression analysis shows that many variables influence SET scores. An interesting question, which is left unanswered at this point, is whether after controlling for all of the factors not under the instructor’s control, ‘will a different ranking of faculty result’?

To answer this question, we use the regression analysis conducted with average teaching evaluation, SET1, as the dependent variable. The estimated coefficient for each instructor variable represents the estimated difference between the identified instructor and the base instructor, Instructor 1. These estimated coefficients, found in Columns 4 and 5 of Table 3, show how SET rankings differ if all instructors had the same class size, had their evaluations conducted in the same manner (online or face-to-face), had
students with the same expected grades, taught classes at the same level, and had the same number of meetings per week. Notice all estimated coefficients on the instructor variables in Columns 4 and 5 of Table 3 are negative, indicating the data have been arranged so that Teacher 1, who was excluded from the regression equation, has the highest predicted ratings.

Table 5 shows how dramatically relative instructor ratings can change when we remove the impacts of factors not under the instructor’s control. The first column of Table 5 lists all of the 15 faculty members. The second column gives their raw SET ranking and the third column gives their adjusted SET ranking holding all variables constant. The highest rated teacher before adjustment is Teacher 15. After adjustment, Teacher 15’s ranking falls to 10th.

Consider Instructor 1, who was ranked 11th out of the 15 instructors based on the raw evaluation scores. Instructor 1 taught larger classes than average which lowered her evaluations. Instructor 1 was regarded as a tough grader, the average expected grades of her students were below average, and she had a larger percentage of her
evaluations completed online than average. Instructor 1’s relatively low unadjusted rating resulted from these factors over which she had no control. When we ask how her relative SET rating would change if she taught classes that were of average size, had students with average grade expectations and had the same percentage of her evaluations conducted online as the other instructors we see that she would have become the top rated instructor.

Instructor 15 on the other hand was teaching small classes at the upper division level and had a larger portion of his evaluations conducted in-class. Instructor 15’s high raw scores include the benefits derived from these advantages. His evaluations cannot be directly compared with those of Instructor 1. In order to compare the performance of these two instructors we must know what would happen if they were compared on equal terms. Table 5 clearly demonstrates that if we control for all these factors, a different picture emerges. Instructor 15’s relative ranking falls to the extent that he is no longer in the top half of all instructors.

Table 5 demonstrates the importance of why a detailed analysis of SET ratings is important in faculty evaluation. By simply using raw means, which are often what is included in tenure and evaluation files, an incomplete picture is drawn. A univariate analysis is inadequate in such an uncontrolled environment. Using a multiple regression analysis, it is simple to set all uncontrollable variables constant for all instructors and then recalculate SET ratings.

The benefits from this type of analysis should help ease the concerns of faculty members who claim that the use of teaching evaluations encourages grade inflation. Of course, if we use unadjusted mean ratings to evaluate faculty, those who give higher grades will likely improve numerical teaching evaluation scores; however, we can adjust for these affects and remove the incentive to give higher grades.

This paper provides new evidence that conducting evaluations online does influence SET scores. We underscore the importance of controlling for factors not under the instructor’s control when using SET ratings for faculty evaluation. We do not suggest that moving to online evaluation is problematic, only that it is difficult to
compare evaluations done online with evaluations done in-class. We do not suppose that one method is ‘more accurate’ than the other. We believe that institutions would benefit from either moving all evaluations online or by continuing to do all evaluations in-class.

Notes on contributors
Clifford Nowell is a professor of economics and associate dean for faculty development in the John B. Goddard School of Business and Economics, Weber State University, USA. His research interests include the economics of education and recreation economics.

Lewis R. Gale is the dean of the John B. Goddard School of Business and Economics, Weber State University, USA. He has more than a decade of teaching experience in business and economics. His research interests are in international and behavioural economics.

Bruce Handley is the chair of the Department of Business Administration in the John B. Goddard School of Business and Economics, Weber State University, USA. He has been chair for more than two decades and has been involved in faculty evaluation for his entire tenure as department chair.

References

