Predictors of LMU Student Evaluations of Teaching Scores

Bryce Mason, Ph.D. and Ryan Johnson, Ph.D.

Is high-quality teaching the only thing measured by student evaluations of teaching (STE)? What other factors partially explain them? Late in 2011, the Committee on the Comprehensive Evaluation of Teaching (CCET) commissioned Institutional Research (IR) to study various hypotheses about influences on ratings. This paper summarizes results of analyses that address questions posed by the CCET regarding the new course evaluation form implemented in fall 2009. In brief, we identified a number of factors with moderate to large influences on ratings, some of which may have policy implications.

Method

IR compiled a course-level data set using multiple sources. First, average ratings were constructed for all questions on the new course evaluation form for each of the first four terms it was used (fall 2009 to spring 2011). Then, instructor characteristics were added to these ratings, including rank (assistant, associate, full, visiting/clinical, or part-time) and demographics (age, gender, and ethnicity). For roughly 85% of instructors, Ratemyprofessors.com also provided a proxy for physical attractiveness; their website asks students to assign attractive professors a chili pepper. In addition, using the student information system, characteristics of the courses were connected to the ratings (e.g., time, location, college, subject code, class size, average grade given, and percentage of enrollment by males). This set contained 4205 evaluated undergraduate courses.

To simplify the analyses, we limited the outcome studied to the global “overall instructor effectiveness” item. The new form contains eight questions that measure teaching quality, but the global item correlated with each of the other measures from 0.78 (instructor being accessible) to 0.90 (constructive interactions). Such correlations suggested to us that results would not fundamentally change had we studied additional outcomes.

Ordinary least squares regression models were the primary method to test hypotheses. First, however, we explored the data to identify general predictive factors that could be used as controls in a model of overall instructor effectiveness. Then, with general covariates set, we tested each of the committee’s hypotheses, which were “Do ratings differ across: (1) gender; (2) college; (3) instructor ethnicity; (4) rank for tenure-line faculty or for term/part-time faculty; (5) courses with a higher proportion of students fulfilling a core requirement; (6) courses with a higher proportion of students fulfilling a major requirement; and (7) courses with a higher proportion of students in that major?”

Results

IR found six strong, statistically significant covariates as given in the table (ordered by strength). Average student interest in the course was by far the most powerful predictor. A 1.0-point increase in the class’s average interest suggested an increase in the overall instructor effectiveness rating of 0.39 points. Being physically attractive—according to users of Ratemyprofessors.com—suggested an average 0.20-point increase in the outcome. Instructors teaching a course for the first time at LMU had a negative association of -0.17 points on their overall instructor effectiveness rating. Age was also negatively correlated. Each additional year
of age predicted a little less than a 0.01-point decline. The average grade assigned to students in the class was a significant positive predictor of ratings of effectiveness of instruction. For each one-point (e.g., a B to an A) increase in the average grade, effectiveness of instruction was predicted to rise by 0.16 points. Finally, course duration (single session) was negatively correlated, with each additional hour of meeting time predicting a -0.07 change in the outcome. Combined, these six variables explained 33% of the variance in the outcome.

Covariates used as hypothesis test controls

<table>
<thead>
<tr>
<th>Factor</th>
<th>Coef.</th>
<th>t</th>
<th>p</th>
<th>Beta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Interest</td>
<td>0.394</td>
<td>17.22</td>
<td>0.00</td>
<td>0.44σ</td>
</tr>
<tr>
<td>Chili Pepper</td>
<td>0.200</td>
<td>6.49</td>
<td>0.00</td>
<td>0.17σ</td>
</tr>
<tr>
<td>Age</td>
<td>-0.007</td>
<td>-5.84</td>
<td>0.00</td>
<td>0.17σ</td>
</tr>
<tr>
<td>New Preparation</td>
<td>-0.168</td>
<td>-5.43</td>
<td>0.00</td>
<td>0.12σ</td>
</tr>
<tr>
<td>Class GPA</td>
<td>0.157</td>
<td>-5.31</td>
<td>0.00</td>
<td>0.13σ</td>
</tr>
<tr>
<td>Course Duration</td>
<td>-0.074</td>
<td>-5.15</td>
<td>0.00</td>
<td>0.11σ</td>
</tr>
</tbody>
</table>

To provide more clarity about the relative strength of these factors, IR calculated Beta for each, the product of its coefficient and that factor’s standard deviation, divided by the standard deviation of overall instructor effectiveness. Interest appeared about twice as strong as age and attractiveness. New preparation, the average grade assigned, and course duration were less powerful.

The following sections present model results for each of the CCET’s questions. The interested reader may find summaries of all variables used in the study, regression output, and tables used to make conclusions in the appendix.

Gender

The results did not support a difference in ratings between male and female instructors. Simple means of average instructor effectiveness showed that men scored roughly a tenth of a point lower than women. However, after accounting for age, this relationship vanished. Among all instructors in the study, the average age of men and women was 51.3 and 44.8, respectively.

The literature on gender differences in student teacher evaluations sometimes finds effects within the sciences, so the same analysis was conducted just among the SCSE, but no differences were found.

College

Using BCLA course ratings as a baseline, SCSE courses averaged 0.11 points lower ($p = 0.01$). At a lower level of statistical significance, SFTV courses averaged 0.09 points lower ($p = 0.10$). Other colleges showed no statistically significant differences.
Ethnicity

These data and methods suggest ratings differences by ethnicity. Among all faculty (tenure-line, visiting/clinical, and part-time), average course ratings for Asian and Black/African-American instructors were 0.24 and 0.29 points lower than White/Caucasian instructors, respectively (top figure). Among tenure-line faculty, average course ratings for Asian and Black/African-American instructors were 0.18 and 0.34 points lower than White/Caucasian instructors (bottom figure).

Rank and Faculty Type

The results suggest a small relationship to ratings, but only for one group. Associate professors averaged 0.09 points less than assistant professors ($p=0.03$). Simple group means of average course ratings among the five groups showed some differences on the surface, with assistant professors earning the highest marks (4.48), followed by visiting/clinical faculty (4.40), part-time faculty (4.33), associate professors (4.32), and finally full professors (4.24). However, accounting for instructor age eliminated all but one of these differences.

Core Fulfillment

There appears to be a small positive association between the proportion of students taking a course for a core requirement and the rating. The difference between a completely non-core class (0%) and a fully-core class (100%) predicted a 0.15-point increase in the average rating of effectiveness of instruction. However, it should be noted that many students appear to misinterpret this question as “core for my college” as opposed to “university core.” For example, in many business and accounting courses there were improbably high levels of students indicating they were taking it to fulfill a core requirement.

Major Fulfillment

We did not identify a statistically significant relationship between overall instructor effectiveness and the proportion of the class taking it to fulfill a major requirement.

Students in the Major

Courses with more students in the major subject of the course appear, on average, to have lower overall instructional effectiveness ratings. The difference between a course with no majors in it and a course entirely composed of majors predicted a 0.08-point decline ($p=0.02$) in overall effectiveness of instruction.
Discussion

Many things appear to be related to ratings of instructor effectiveness, but it is important to keep in mind that this method only establishes correlations and does not imply causation. In such a model, it is impossible to know whether the factor showing a relationship to the outcome is truly the cause, or whether some other underlying variable is responsible for both. For example, despite RateMyProfessors.com asking to students to judge their instructor’s appearance, it may also measure other attractive qualities like a positive attitude or joie de vivre. Alternatively, it could have the reverse relationship in that students may find excellent instructors more charismatic and attractive.

Numerous stories could be told to explain each of the findings in this analysis, with some factors yielding many competing hypotheses. Some of the factors identified are probably more likely to have clearer relationships to ratings, such as duration of the class and whether the course was a new preparation. They were measured carefully, had the relationship that IR expected, and we could not think of other factors that explain them away. However, other factors such as age, student interest, and the average grade assigned are likely more complicated.

Probably the most complicated finding to interpret is the one regarding ratings and ethnicity. Fortunately, other scholars have provided some thinking on this topic as many universities have had similar experiences. Collaborating with the Office of Intercultural Affairs, we conducted a literature review of other institutions’ experiences with race/ethnicity inequities in STE.

Although there were a limited number of studies that explored the relationship between race/ethnicity and STE, they generally reported a negative main effect for faculty of color. Hamermesh and Parker (2005), who were mainly interested in estimating the effect of beauty on student ratings of professors, found a negative 0.5σ impact on STE scores for being a female minority instructor vs. all White faculty.¹ Smith (2007), who analyzed STE scores for 190 tenure-line instructors in a college of education, found that students rated Black/African-American faculty 0.5σ and 0.6σ lower than white faculty in overall course value and overall teaching ability, respectively. Similarly, DiPietro and Faye (2005) noted that students tended to rate white faculty higher than Asian and Hispanic/Latino faculty. Although limited to economics classes, McPherson et al. (2009) observed a negative 0.3σ impact on STE scores for non-White faculty. Finally, in a recent large-scale study across multiple selective liberal arts colleges using STE scores from RateMyProfessors.com and student perceptions of race/ethnicity using publicly available information about the instructor, Reid (2010) found that Asian and Black/African-American faculty were rated lower than White/Hispanic/Latino faculty on three dimensions (overall quality, helpfulness, and clarity) in the range of 0.1σ-0.2σ (Asian) to 0.4σ (Black/African-American).

Some reasons hypothesized in the literature for inequities in STE scores for faculty of color include student prejudices, student preconceptions, language differences, and expectations of students regarding minority faculty members. Many researchers have noted that students may have communication issues with faculty whose native language is not English. Most recently, Hamermesh and Parker (2005) found a negative 0.5σ effect on STE scores for non-native English speakers, and other research has found similar results (Rao, 1995; Rubin, 1998).

¹ Minority was not defined.
However, most of the literature speaks to possible prejudices and expectations of students with respect to faculty of color. One study performed controlled experiments using a fake course syllabus to tease out student preconceptions and found that students rated Anglo women professors more capable than Latina professors (Anderson & Smith, 2005). Another experiment involved an Indian instructor giving two identical courses with either traditional or Western dress and found substantial STE score differences (Chowdhary, 1988). Further, Smith (2007) observed that among 26 measures of instructor behavior that presumably would be consistent with a separate, global assessment of teaching ability (e.g., “Is this instructor effective?”), Black/African-American faculty were rated lower than all of the behavior-specific means. This is supportive of the idea that students may be letting things other than instructor behavior dictate their ratings. Lastly, in a series of student interviews, Hendrix (1998) found that students in a predominantly White university did not believe that a professor’s race/ethnicity influenced their perceptions of instructor credibility; however, the students simultaneously described different criteria for evaluating Black/African-American instructors for courses on certain topics. Students gave more credibility to Black/African-American instructors when they taught courses that had an ethnic or racial focus, and were more willing to question their credibility for courses that lacked such a focus.

Ways to address inequities in STE scores for faculty of color were rarely addressed in the literature reviewed, but many scholars pointed out that the stakes for bias in STE are high, because they tend to be used as a part of merit and promotion decisions (e.g., Boatright-Horowitz & Soeung, 2009; McPherson et al., 2009). McPherson et al. (2009) were the only team in the literature reviewed who outlined a framework for a policy remedy by creating adjusted ratings based on predicted vs. actual values—essentially giving credit for the average differences observed for many key factors, including race/ethnicity. Researchers that have focused on the role of faculty members’ gender on STE scores, however, have recommend other actions such as including measurements of teaching techniques on STEs (Feldman, 1993; Ludwig & Meachum, 1997; Crombie et al., 2003). Chamberlin & Hickey (2001) recommend having an open dialogue about different teaching styles, how these styles are translated into STEs, and how STEs are used to make career decisions.

This literature parallels some findings at LMU, but with key differences. For example, many other institutions reported similar average score differentials (roughly 0.2σ to 0.6σ, for various measures) for Black/African-American faculty. Some research also found negative impacts for Asian faculty and, similar to LMU, the size of the differential appeared smaller than for Black/African-American faculty. However, none of the literature reviewed explored these differences among tenure-line vs. other types of faculty, which for LMU appeared to be an important moderating factor. A final key difference is that, while some studies found negative relationships for Hispanic/Latino faculty, LMU did not experience this type of relationship.

Conclusion

Addressing questions put forth by the CCET, this analysis has identified several variables that partially predict SET ratings. The present research should promote richer faculty and administrator discussion about the appropriate use of ratings in the rank and tenure process.
References


