NORMING OF STUDENT EVALUATIONS OF INSTRUCTION: IMPACT OF NON-INSTRUCTIONAL FACTORS

Satish Nargundkar
Department of Managerial Sciences
Robinson College of Business
Georgia State University
Atlanta GA 30302
678-644-6838
snargundkar@gsu.edu

Milind Shrikhande
Department of Finance
Robinson College of Business
Georgia State University
Atlanta GA 30302
404-406-8556
mshrikhande@gsu.edu

ABSTRACT

We analyzed Student Evaluations of Instruction (SEIs) from about 6000 sections over four years representing over 100,000 students at Georgia State University's Robinson College of Business, to look for factors other than instructional attributes that might have an impact on the scores. We looked at environmental factors like semester, time of day, location, and instructor attributes like gender, and rank. These were analyzed across four segments created by course level (graduate and undergraduate) and course type (core and non-core). We found that several of the factors had varying degrees of impact on overall evaluation of an instructor. Summer semester ratings were higher than the spring semester ratings, which in turn were higher than those in the fall semester. Time of day and classroom location mattered, though not in a consistent way across the segments. Undergraduates rated female instructors higher, while graduates rated male instructors higher on average. Instructor rank mattered, with non-tenure track faculty generally outperforming others (although there were small variations in results). Overall impact of all the non-instructional factors studied was relatively low in terms of explaining the variation in the rating of instructor effectiveness.

Keywords: Instructional Innovation, Student Evaluation, Norming, Non-Instructional Factors, Gender Bias, Rank, Environment, Faculty Performance.

INTRODUCTION

Student Evaluations of Instruction (SEIs) are now commonplace among universities as a key mechanism for getting feedback regarding teaching practices. These SEIs also form a key component of evaluations of faculty teaching performance by the administration, and impact promotion and tenure decisions. As such, there is always debate about the validity and appropriate use of these instruments. Brightman (2005) has argued that to be useful, an instrument must first be valid, and that norming procedures must be in place to aid comparative interpretation of the data. This requires an understanding of systematic biases in the results due to factors that go beyond the teaching aspect.

A clear understanding of the impact of non-teaching related factors is necessary to ensure that evaluation of faculty is fair. Researchers have examined the impact of various factors on SEI results to look for systematic biases in various fields, from Psychology (Greenwald, 1997) to Economics (Isely & Singh, 2005) and Business (Peterson, et al, 2008; Isely & Singh, 2007; Liaw & Goh, 2003). The non-teaching related factors can be classified as Student related, Instructor related, Course related, and Environmental (Peterson, et al, 2008). Student related factors include the initial motivation of the student for the subject, grade expectation, grade point average, and gender. Instructor related factors include the instructor's rank and gender, while course characteristics include type of course (qualitative vs. quantitative, core vs. non-core), course level. Environmental factors that might influence SEI ratings include class size, location, classroom structure and equipment, time of day.

Research Questions

While many researchers have been examining the impact of non-teaching related factors on instructor ratings in different disciplines, there is a need to conduct more such studies to look for consistent patterns across universities and disciplines, or examine the differences as they appear. The non-instructional factors, especially environmental ones, are likely to be different in each institution, and a fair evaluation requires examination of the data in various institutions.

Our study focuses on SEIs from the Robinson College of Business at a Georgia State University. Our data spans across four years and 10 different departments, and we examine instructor, course, and environmental factors. While many researchers agree that student grade expectations are positively correlated with SEI ratings, we did not include that in our study, since grade expectations are sufficiently intertwined with teaching ability that it is difficult to separate the two. In other words, is it merely "easy grading" or is it excellent teaching that results in the higher grade expectation?[1] Our focus in this paper is on elements that are more clearly separable from the quality of teaching.

[1] It is a long-standing belief among many professors that "easy does it." Zangenehzadeh (1988) concluded that student ratings of faculty have resulted in changing teachers' grading behavior. Marsh and Roche (2000) debunk popular myths that student evaluations of teaching (SETs) are substantially biased by low workload and grading leniency. Using structural equation models the studies confirmed that perceived learning and prior characteristics (course level, prior subject interest) account for much of the grade-SET relation. A nonlinear relation also indicates that high grades are unrelated to SETs indicating bias claims are not tenable. Centra (2003) analyzed more than 50,000 college courses controlling for class size, teaching method, and student perceived learning outcomes in the course. Learning outcomes turned out to have a large positive effect. After controlling for learning outcomes, expected grades did not affect student evaluations.

We examine the following research questions:

1. Are overall ratings for instructor effectiveness different for Core and Non-core classes?
2. Does Course Level (graduate/undergraduate) affect SEI ratings for overall instructor effectiveness?
3. Do Instructor rank and gender affect SEI ratings for overall instructor effectiveness?
4. Do environmental factors like semester, time of day and location affect ratings for overall instructor effectiveness?
5. How strong is the overall impact of the non-instructional factors on student ratings of instruction?

We are interested in knowing if these factors are significant, and if the impact is large enough to be concerned about when comparing faculty performances.
The rest of the paper is organized into the following sections: literature review, methodology, results and discussion. We discuss the implications of our results specifically for our college and consider which of those results are broadly generalizable.

LITERATURE REVIEW

There is debate in the literature about the appropriateness of using student evaluations of instruction for assessment of teaching. Since the goal of teaching is to improve student learning, that is what must be measured, according to some researchers (McKenzie, 1975; Kelley, 1972). Student evaluations measure the characteristics of the intervention, rather than its effectiveness. There are practical problems with measuring student learning too, however, and associating the learning with a specific instructor's ability. While not perfect, student evaluations do provide at least the student's perception of how well an instructor was able to help them learn. This assumes, of course, a valid instrument that has questions that relate to student learning. Another major argument against SEIs is that they can be potentially biased and give a distorted view of the teaching ability of an instructor. We examine the research regarding some of factors that might bias student responses.

Course Related Factors

Do course type or level affect student ratings systematically? Costin et al (1971) studied ratings by class designation from freshman to senior, and noted that seniors gave higher ratings to instructors than did the less experienced freshmen. It could be because better instructors are selected to teach higher level classes, indicating a selection bias of sorts. It could also be because the poorer students drop out in the first couple of years, and better students make it to the senior year, which also affects instructor ratings.

Peterson et al (2008) find the senior-level students giving better ratings than sophomores and also better ratings than students taking graduate courses. Given that the 400- or senior-level courses are (a) in the discipline concentration, (b) student-selected electives, or (c) the required business capstone, one possible explanation for their significantly better student evaluations is what might be termed a "familiarity effect." Students become the more familiar with the professors from whom they have taken earlier classes and therefore have a reduced anxiety.

Aigner & Thum (1986) also conducted a detailed study of various factors that affect ratings, and among other things, noted that student ability has an impact. Courses aimed at students of high ability get higher ratings, and those aimed at students with low ability get lower ratings. Some of that may translate to non-core classes getting higher ratings, since those courses are selected by students that presumably believe that they have some ability in that subject. Brightman et al (1993) used four categories to norm SEI data – undergraduate core, undergraduate non-core, graduate core and graduate non-core, arguing that ratings for graduate courses would be higher than undergraduate, and non-core higher than core. Our first hypotheses therefore address the issue of course type and course level.

> **Hypothesis 1a:** Non-core classes will have higher ratings compared to core classes
> **Hypothesis 1b:** Graduate classes will have higher ratings compared to undergraduate classes.

Instructor Related Factors:

Is there a gender bias in the ratings for overall instructor effectiveness? Gender differences in performance evaluations in various fields have been studied extensively in the literature (Arvey, 1979; Dobbins, Cardy & Truxillo, 1988; Mobley, 1982). Most of the studies of gender differences regarding student evaluations of instruction have focused on the gender of the instructor rather than the student. Del Boca & Ashmore (1980) discussed general gender stereotypes, and found that positive characteristics of stereotypical men include rationality, competence and assertiveness, while for women warmth and expressiveness were seen as the main positive traits. How does this translate into perception of instruction by male and female instructors? The results in the literature are mixed. Sprague & Massoni (2005) found that students do expect different things from male and female instructors – for instance, they expect a woman to be caring, while they expect a man to be funny and intelligent. They argue that the burden on female instructors is more labor intensive, since the interpersonal relationship with students cannot be carried over from one semester to the next. Lackritz (2004) found that burnout for female faculty was related to lower ratings. Bauer & Baltes (2002) found that not all students who held traditional gender stereotypes rated women less accurately and more negatively. Heckert et al (2006) and Tatro (1995) found that women receive less favorable ratings than their male colleagues. However, other studies showed no difference between male and female instructors (Blackhart et al, 2006; Feldman, 1993). Reid (2010) studied race and gender of instructors and found that while racial stereotypes exist in student evaluations, gender did not seem to have any effect. Mohan (2011), on the other hand, in a study at a catholic business school, reports that female instructors received lower ratings than males.

Fewer studies have reported differences based on the gender of students. Feldman (1993) found that students rated same sex instructors a little higher. Hancock et al (1993) found no consistent pattern of gender difference across different colleges. Kohn & Hatfield (2006) found that female students rated male instructors higher than did male students, but did not see that difference for female instructors.

Based on all the mixed results in the literature, we hypothesize that there will be no difference in ratings due to gender.

Hypothesis 2a: Overall ratings by students will not be significantly different for male and female faculty members.

How does the rank or status of instructor affect the overall ratings given by students? Among the instructors' attributes that potentially influence the ratings are the instructors' positions or ranks, how demanding they are perceived to be, as well as experience, training, communication skills, and age (Blackburn & Lawrence, 1986). Isley and Singh (2007) studied various factors that influence student evaluations, including grade expectations, class size, the difficulty of the class, the percent of students responding, and the length of class. They found that while higher expected grades result in more favorable student evaluations, this relationship is significantly different depending upon faculty rank. The interaction effects showed that adjunct faculty ratings are most affected by student grade expectations, followed by tenured faculty, and lastly by tenure track faculty. They make the case that this is because of the differential emphasis on student evaluations in the overall evaluation of the faculty by administration. Adjunct faculty were judged solely on student evaluations, while tenure track faculty were judged on a wide portfolio of activities. Although we exclude expected grades from our analysis, the influence of faculty rank by itself is germane to our study. Mohan (2011) also reports that non-tenure track faculty get higher ratings than tenure track faculty, although the effect can be altered, she argues, by inflating grades. Peterson et al (2008) did not find any difference in ratings received by full-time faculty versus ratings received by adjunct faculty.

At our college, besides Tenure track (TT) and Tenured faculty, we have Part Time Instructors (PTI), Graduate Teaching Assistants (GTA), and Full time Non-tenure Track faculty (NTT). The last group is evaluated primarily on teaching, while Tenure Track faculty are evaluated primarily on research, with less emphasis on teaching. As such, we expect some differences in these groups.

Hypothesis 2b: Overall ratings of instructors will be different for different faculty ranks.

Environmental Factors

Do environmental factors like semester, time of day and location affect ratings for overall instructor effectiveness? Peterson et al (2008) in their study within their Managerial Sciences department find no evidence of any difference between Spring and Fall semester ratings. There is some evidence in the literature that class size does affect student ratings, with lower class sizes yielding higher ratings (Feldman, 1984; Liaw and Goh, 2003; Isley & Singh, 2007). Bedard & Kuhn (2008) find that for class sizes under 80, there is a relatively steep price to be paid for each additional student in terms of loss of ratings. The difference in ratings per additional student is not so great in larger class sizes (80-150 students). Since class sizes in the summer semesters typically are smaller than those in Fall or Spring in most colleges, one would expect ratings for summer classes to be higher. There is, however, no reason to expect any differences between the Fall and Spring semester ratings.

Hypothesis 3a: Ratings for the summer semester will be higher than those for Fall and Spring semesters.

Does time of day matter? Peterson et al (2008) suggest that daytime classes get better ratings than evening classes, which they attribute to either higher expectation from students who work during day and taking evening classes, or because these students resent being given homework that adds to their several preoccupations. In general, people describe themselves as a "morning person" or a "night person" depending on when they are at their most alert. The afternoon times typically are not described by people as their best time to focus on learning tasks. We divided the day into four segments – morning, afternoon, early evening, and evening. We expect to see some impact of the time on the ratings.

Hypothesis 3b: Ratings for morning or evening classes will be higher than for afternoon or early evening classes.

Finally, does the quality of the classroom itself matter? Some classes are taught in modern facilities with stadium seating, spacious rooms, ports for student laptops, internet connections, while others are still taught in fairly old, cramped rooms with students on chairs with a large arm on which to write. Little research has looked into this aspect. Based purely on anecdotal data about student complaints regarding old facilities, it might be that some of that frustration would lead to an overall reduction of positive experience, and perhaps affect ratings of instruction as well.

Hypothesis 3c: Classes taught in better quality classrooms will have higher ratings.
It is also important to know the overall effect size of all of these factors taken together. How much of the overall variation in the student ratings of instruction is explained by these non-instructional factors? The overall effect should be relatively small, compared to the effect of teaching related factors, for evaluation instruments to be meaningful at all.

METHODOLOGY

We collected data on all student evaluations filled out between 2005 and 2009 in the Robinson College of Business (RCB) at Georgia State University. For this study, the data were aggregated at the section level. Each row in the dataset after aggregation represented a section and the average scores for that section for each of the questions in the SEI instrument, along with information about the course, location, instructor, and other variables. Over 6000 sections of various courses were taught during this period. We eliminated all PhD classes from our analysis, since they tend to be very small in size, and are sufficiently different from typical undergraduate or graduate courses.

The response rates in each of the four segments (graduate/undergraduate, core/non-core) were all in the 60-70% range, which is par for most Universities. Richardson (2005) surveyed the literature on student evaluation instruments, and indicates that response rates of around 60% are common and that a 70% response rate would be considered good.

We computed average scores on the global question regarding instructor effectiveness (Q34 on the instrument used by RCB) for various segments based on our hypotheses, and conducted 2-sample t-tests and ANOVAs as appropriate to test for significant differences across the various

subgroups. If ANOVAs were significant, we used Tukey's two-way comparisons to determine specific differences among subgroups.

We then tested the overall impact of all the non-instructional factors taken together by creating dummy variables to indicate various subgroups for time of day, location, rank, gender, course type and course level, and then doing a regression analysis with Q34 as the dependent variable, and the dummies as the independent variables.

RESULTS

To test the first hypothesis, we compared the mean Q34 scores (Likert scale, 1=low, 5=high) for core and non-core classes using 2-sample t-tests, and then for graduate and undergraduate classes. The table below shows the results.

Course Type			**Course Level**	
Core classes	4.239 n=2490		Graduate classes	4.315 n= 2165
Non-Core classes	4.320 n= 3334		Undergraduate classes	4.268 n=3659
	$p < 0.001$			$p < 0.01$

Table 1: Ratings by Type (Core vs NC) and Level (Grad vs UG) overall

In both cases, there was a significant difference. Ratings for non-core classes were significantly higher than those for core classes, while graduate classes got higher ratings than undergraduate classes.

For the next hypothesis, we decided to create four segments based on course level and type, rather than looking at each one independently as above. Thus we compared ratings for Core and Non-core classes separately for Undergraduate and Graduate courses, and similarly, compared Graduate and Undergraduate ratings separately for Core and Non-core classes. The results are shown in Table 2 below.

	Undergrad	**Graduate**	
Core	4.228 n=1668	4.260 n=822	$p > 0.10$
Non-Core	4.301 n=1991	4.349 n=1343	$p < 0.05$
	$p < 0.001$	$p < 0.001$	

Table 2: Ratings by segment - Course Level by Type and Course Type by Level

As seen above, when it comes to Core classes, the ratings are not significantly different for Undergraduate and Graduate classes. For Non-core classes, however, ratings for graduate classes are significantly higher than for undergraduate classes. When it comes to the difference between Core and Non-core classes, ratings for non-core are higher in both the undergraduate and

graduate segments. For each of the four segments described above, we broke down the sections by the gender of the faculty members teaching those sections, and compared ratings for Male and Female faculty. Table 3 below summarizes our findings.

	Undergrad	**Graduate**
Core		
Female	4.237 (n=929)	4.285 (n=217)
Male	4.217 (n=719)	4.243 (n=572)
	P > 0.10	P > 0.10
Non-Core		
Female	4.355 (n=688)	4.286 (n=244)
Male	4.278 (n=1273)	4.365 (n=1086)
	p < 0.01	P<0.05

Table 3: Ratings by Instructor Gender within each segment.

Interestingly, as with the previous hypothesis, no significant differences were found between the genders for Core classes. For the Non-core segment, the ratings for females were higher than for males among undergraduate students, while the reverse was true among graduate students. When all four segments were combined, there was no difference overall between male and female instructors.

Faculty Status (rank) was the next variable we tested. We looked at various classifications of faculty, and eliminated sections with adjunct faculty or where faculty status was missing. This resulted in about 200 sections being eliminated overall (about 3.3% of the sections). The analysis was again conducted within each of the four broad segments. Table 4 below summarizes the results of faculty ratings by status.

	Undergrad	**Graduate**
Core	1. Tenured 4.32 (n=134) 2. NTT 4.28 (n=703) 3. GTA 4.25 (n=322) 4. PTI 4.19 (n=381) 5. TT 4.15 (n=27) 1,2 > 3,4,5 and 3 > 5 p < 0.05	1. NTT 4.36 (n=332) 2. Tenured 4.26 (n=248) 3. TT 4.14 (n= 55) 4. PTI 4.04 (n=144) 1 > 3,4 and 2 > 4 p < 0.05
Non-Core	1. NTT 4.35 (n=618) 2. PTI 4.31 (n=341) 3. TT 4.28 (n=166) 4. Tenured 4.25 (n=547) 5. GTA 4.15 (n=149) 1 > 4,5 and 2 > 5 p < 0.05	1. NTT 4.41 (n=362) 2. Tenured 4.38 (n=628) 3. PTI 4.20 (n=150) 4. TT 4.13 (n=144) 1, 2 > 3, 4 p < 0.05

Table 4: Ratings by Faculty Status within each Segment (based on Course Type and Level)

In each of the four segments, the ANOVA was significant at $p < 0.001$ overall, meaning that the scores for all faculty status groups were not equal; there were some differences somewhere. Tukey's two-way comparisons showed the specific differences as shown in the table above. For instance, for the Undergraduate Core segment, "1,2 > 3,4,5" means that the first two groups (Tenured and NTT) were not different from each other, but each of them was significantly better than groups 3, 4, and 5 (PTI, GTA and TT). Further, "3>5" means that group 3 (PTI) was significantly better than group 5 (TT).

To study the environmental factors, we compared ratings for sections taught in the Fall, Spring and Summer semesters with each other. The results are shown in Table 5 below.

Fall	4.2479 n=2320
Spring	4.2998 n=2348
Summer	4.3853 n=1322
Overall $p < 0.001$ Summer > Spring > Fall, $p < 0.05$ pairwise	

Table 5: Ratings by Semester across all years

Student evaluations were found to be significantly higher during Summer compared to Spring, and likewise significantly higher for Spring compared to Fall. Broken down by segment, the mean ratings for each semester are as shown in Table 6 below.

	Undergrad			**Graduate**		
Core	Fall	4.188	n=652	Fall	4.240	n=355
	Spring	4.212	n=671	Spring	4.244	n=283
	Summer	4.337	n=345	Summer	4.326	n=184
	Summer>Spring, Fall; $p<0.05$				$p>0.05$	
Non-Core	Fall	4.229	n=732	Fall	4.295	n=508
	Spring	4.312	n=795	Spring	4.359	n=530
	Summer	4.397	n=464	Summer	4.422	n=305
	Summer>Spring>Fall, $p<0.05$			Summer > Fall, $p<0.05$		

Table 6: Ratings by semester for each of the four segments

The largest differences between semesters were found for undergraduate classes in general, specifically undergraduate non-core, where summer ratings were higher than spring, which were higher than fall. The same pattern was observed in each segment, although in the graduate core, the differences were not significant. This may again have to do with class size, as non-core classes are typically smaller than core, and summer classes are smaller than the other semesters.

We next tested for differences in ratings for sections taught at various times during the day. The day was divided into four time segments. All classes that began before noon were in the Morning group Those that began at or after noon but before 4:30 PM were classified as afternoon. Those that began at 4:30 PM but before 7:15 PM were classified as Early Evening, while those that started at 7:15 PM or later were the Evening classes. The results are shown in Table 7 below.

	Undergrad	**Graduate**
Core	1. Morning 4.2229 (n=675) 2. Afternoon 4.2260 (n=338) 3. Early Evening 4.2123 (n=300) 4. Evening 4.2229 (n=355) $p > 0.10$	1. Morning 4.4117 (n=184) 2. Afternoon 4.3332 (n=31) 3. Early Evening 4.1844 (n=303) 4. Evening 4.2305 (n=291) $p<0.001$; Pairwise: 1 > 3,4
Non-Core	1. Morning 4.3479 (n=340) 2. Afternoon 4.2239 (n=630) 3. Early Evening 4.3019 (n=569) 4. Evening 4.2908 (n=339) $p < 0.05$; Pairwise: 1,3 > 2	1. Morning 4.3413 (n=85) 2. Afternoon 4.3160 (n=53) 3. Early Evening 4.2992 (n=549) 4. Evening 4.3947 (n=656) $p < 0.05$; Pairwise: 4 > 3

Table 7: Time of day by segment (based on Course Type and Level)

With all segments combined, $p > 0.10$ for time of day. In other words, there was no difference between scores at various times of the day overall. When broken down by segment, the results are mixed. Graduate core classes score better in the mornings, while Graduate non-core classes (which are mostly taught Early Evening or Evening) score better in the Evening compared to Early Evening. Undergrad core classes show no difference overall, while Undergrad non-core do better in the morning and early evenings.

Finally, we looked at the location of the classes. The results of our analysis are shown in Table 8.

	Undergrad	**Graduate**
Core	1. Aderhold 4.2102 (n=939) 2. Alpharetta 4.2197 (n=129) 3. Classroom South 4.2944 (n=242) 4. General Classroom 4.2988 (n=212) 5. Brookhaven 4.2599 (n=43) 6. Sparks Hall 4.0378 (n=57) p< 0.01; Pairwise: 3,4 > 6	1. Aderhold 4.1925 (n=261) 2. Alpharetta 4.3091 (n=183) 3. Classroom South 4.3069 (n=89) 4. General Classroom N/A (n=0) 5. Brookhaven 4.2412 (n=162) 6. Sparks Hall N/A (n=0) p > 0.10
Non-Core	1. Aderhold 4.2684 (n=1145) 2. Alpharetta 4.4155 (n=27) 3. Classroom South 4.3202 (n=250) 4. General Classroom 4.2779 (n=331) 5. Brookhaven 4.6667 (n=1) 6. Sparks Hall 4.2824 (n=87) p > 0.10	1. Aderhold 4.3256 (n=603) 2. Alpharetta 4.3050 (n=129) 3. Classroom South 4.2579 (n=171) 4. General Classroom 4.3132 (n=106) 5. Brookhaven 4.2846 (n=59) 6. Sparks Hall 4.3531 (n=31) p > 0.10

Table 8: Location by segment (based on Course Type and Level)

Some classroom buildings are new (Aderhold), with modern facilities, many rooms with stadium seating, plugin ports for student laptops, while other classroom buildings (Sparks Hall, General Classroom) are very old, with relatively cramped classrooms. No significant differences by location were found in any of the segments except Undergraduate Core classes, with Classroom South and General Classroom building scoring higher than Sparks Hall. Interestingly, classes in the most modern buildings (Aderhold, Alpharetta) do not show the highest scores.

Besides studying the various factors individually as shown in the above tables, we wanted to see the overall effect of all the non-teaching factors on the global rating of instruction. A regression was performed with Q34 as the dependent, and dummy variables were created for the independent variables to represent location, time of day, instructor gender, and so forth. For the dummy variables in the regression, note that Spring and Summer are compared against Fall as the baseline (neutral group), Time of Day uses Evening as baseline, Graduate Non-core and Undergraduate Non-core combined is the baseline for course level and type, and Tenure Track (but not tenured) is the baseline for faculty status. Table 9 shows the results below.

Regression Statistics	
Multiple R	0.19529
R Square	0.03814
Adjusted R Square	0.03637
Standard Error	0.52843
Observations	5996

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	4.33194	0.02365	183.18051	0.00000
Spring	0.04839	0.01550	3.12158	0.00181
Summer	0.13217	0.01827	7.23576	0.00000
Morning	-0.06765	0.02004	-3.37586	0.00074
Afternoon	-0.12268	0.02052	-5.97786	0.00000
Early Evening	-0.10362	0.01760	-5.88731	0.00000
UC	-0.06846	0.01754	-3.90210	0.00010
GC	-0.09389	0.02083	-4.50689	0.00001
Tenured	0.04601	0.02278	2.01986	0.04344
NTT	0.07643	0.02232	3.42427	0.00062
PTI	-0.06445	0.02542	-2.53514	0.01127
GTA	-0.13126	0.03174	-4.13582	0.00004

Table 9: Regression of Q34 on non-instructional factors. Highlighting is to show groups of dummies for a given variable together.

None of the variables for location were significant in the regression. The overall R-square value for the regression was only 0.038 (adjusted value of 0.036), indicating that less than 4% of the total variation in Q34 was explained by all of these non-teaching variables.

DISCUSSION

Instructor ratings were found to be significantly different for Course related factors like the course level and type. As expected, ratings were higher for non-core classes compared to core classes. This is consistent with our expectations as well as the findings of Brightman et al (1993) and Aigner & Thum (1986), and Peterson, et al. (2008). It seems to be fairly well established that initial liking for a course does in fact affect the ratings of an instructor. Similarly, we found that graduate classes get better ratings than undergraduate classes overall. This may have to do with the fact that graduate students are already among the higher achieving students when they were undergraduates. Thus they are generally expected to be better prepared and have a greater liking for the subject than a typical undergraduate might.

Since our college already norms the SEI reports for instructors by separating the four segments – undergraduate core, undergraduate non-core, graduate core and graduate non-core, based on

Brightman et al's (1993) study, we decided to test the differences in each of those categories. We found that among core classes alone, there was no difference in ratings for undergraduate and graduate classes. There was a difference between the two among non-core classes, however. This indicates that the initial dislike for core classes trumps any difference between undergraduates and graduates. Seen from the other direction, among undergraduate classes alone, the difference between core and non-core classes was significant. This was also true for graduate classes.

In terms of instructor related factors, we found that gender has an effect on ratings, but only for non-core classes. There was no significant difference in ratings between male and female instructors for core classes, whether undergraduate or graduate. However, we found an interesting effect in the non-core classes. The traditionally accepted view in the literature is that female instructors tend to be rated lower in general. We found that to be true only among graduate students. Undergraduate students rated female instructors higher than male instructors in non-core classes, while graduate students rated male instructors higher than female instructors. In general, undergraduate students are in the age range 18-22, while graduate students are older, and typically have work experience. According to the literature, the expectation from female instructors is that they be nurturing, show warmth and expressiveness (Del Boca & Ashmore, 1980). This expectation is perhaps more pronounced among younger students.

Instructor rank or status also had an impact on the ratings. In all four segments, Non-Tenure Track (NTT) instructors consistently got higher ratings than Tenure Track (TT) faculty. However, faculty members that were already tenured performed very well, especially in graduate classes. Among undergraduate students, part time instructors (PTIs) also consistently did better than TT faculty. Mohan (2011) found similar differences among NTT and TT faculty in her research. This is in our opinion consistent with the incentive structure in place. NTT faculty members are primarily charged with teaching classes, with a lower research requirement. TT faculty members, on the other hand, are required primarily to publish, and their focus is less on teaching. However, when they do get tenured, the focus on research is reduced, giving them more time to focus on teaching.

The influence of environmental factors like semester, time of day and location (classroom quality) on instructor ratings was mixed. Summer semester ratings were consistently higher than the Spring or Fall semesters, with the only exception being graduate core classes where there was no significant difference between any of the semesters. The higher summer ratings are consistent with expectations for smaller class sizes. Summer classes on average have around 20-25 students, while fall and spring classes have 30+ students on average. Also, students take fewer classes during summer, and are more focused as a result on the classes they do take. Further, more frequent meetings during summer perhaps helps build a better rapport with the instructor.

As for time of day, morning classes received a higher rating than evening for graduate core classes, where not many classes were offered in the afternoon. Also, many of these morning courses are offered on Saturday mornings, when the graduate students are relatively free from work related pressures. Among undergraduate core classes, we found that morning and early evening classes scored higher than afternoon classes, consistent with our expectation based on tiredness/sleepiness after lunch. Finally, among graduate non-core classes, evening classes score higher than early evening (there are few classes taught in the morning or afternoons). This is also

consistent with our expectations. After a long day at work, the students are typically tired for the early evening class, but get a second wind post dinner for the evening classes. Location had a rather unexpected result. Only among undergraduate core classes did we see any significant effect at all. Even among them the best facilities did not get the highest ratings for the instructors. However, in a cluster of three classroom buildings, the one that is the oldest and least equipped did get the lowest ratings for instruction.

The regression analysis using all of the above factors as dummy variables showed an R-square value of only 0.038. In other words, less than 4% of the variation in instructor ratings was explained by all of the non-instructional factors we studied. This is good news for instructors in general, since it indicates that while extraneous factors do have some statistically significant effects on their ratings by students, these effect overall are relatively low compared to the impact of the instruction related factors. Nargundkar & Shrikhande (2012) found that about 73% of the variation in the overall instructor rating was explained by the six underlying instruction related factors on this instrument.

CONCLUSION

It must be remembered that ultimately SEIs should be a mechanism for improving student learning. As Brightman (2005) points out, this can happen only if the SEI is a valid instrument with questions that are based on factors that have been shown to be related to student learning. Further, the results of the SEIs should be appropriately normed for fair feedback to faculty, eliminating the effects of extraneous non-instructional factors. Finally, there must be a faculty development mechanism in place to help faculty actually improve their teaching over time. Mere feedback is not enough.

The validity of the instrument used at Georgia State's Robinson College of Business was shown by Brightman et al (1989) and the instrument was revalidated in recent times by Nargundkar & Shrikhande (2012). The four segments currently used for comparison (undergraduate core/non-core, graduate core/non-core) still seem appropriate given the results of this study. Peterson et al (2008) suggest the possibility that instructors may try to game the system by using these results to improve their ratings without necessarily improving student learning. The fact that the overall impact of all these factors was very low makes that less of an issue here.

For instructors, the positive finding of our study is that while extraneous factors may influence their ratings a little, the impact overall is small enough to be ignored, especially with the norming in place. In other words, the non-instructional factors are weak enough to keep intact the validity of the SEI process. Still, many researchers provide ways of guarding against potential bias in student evaluations of instruction (Baldwin and Blattner, 2003). They recommend using alternative approaches such as portfolios, peer feedback sessions, and informal student surveys to combat or circumvent these potential biases.

REFERENCES

Aigner & Thum (1986) "On Student Evaluation of Teaching Ability" *Journal of Economic Education*, Fall 1986, pp. 243-265.

Arvey, R. D. (1979) Unfair discrimination in the employment interview: Legal and psychological aspects. *Psychological Bulletin*, 86, 736-765.

Baldwin & Blattner (2003) "Guarding Against Potential Bias in Student Evaluations: What Every Faculty Member Needs to Know", *College Teaching*, Vol. 51, Issue 1, 2003.

Bauer, C. B., & Baltes, B. B., (2002) "Reducing the Effects of Gender Stereotypes on Performance Evaluations of College Professors," *Sex Roles: A Journal of Research*, 47, 465-476,.

Bedard & Kuhn (2008) "Where class size really matters: Class size and student ratings of instructor effectiveness" *Economics of Education Review*, 27, 253-265.

Blackburn, R. T., & Lawrence, J. H. (1986). "Aging and the quality of faculty job performance." *Review of Educational Research, 56*, 265–290.

Blackhart, G. C., Peruche, B. M., DeWall, C. N., & Joiner, T. E. J. (2006). "Factors influencing teaching evaluations in higher education." *Teaching of Psychology, 33,* 37–39.

Brightman, H.J., Bhada, Y., Elliott, M., & Vandenberg, R. (1989) "An Empirical Study to Examine the Reliability and Validity of a Student Evaluation of Instructor Instrument," GSU College of Business Administration Internal Working Document, prepared by the Faculty Development Committee (FDC).

Brightman, H., Elliott, M., Bhada, Y., (1993) "Increasing the Effectiveness of Student Evaluation of Instructor Data through a Factor Score Comparative Report", *Decision Sciences*, Jan/Feb, p. 192-199.

Brightman, H. J. (2005). "Mentoring faculty to improve teaching and student learning." *Decision Sciences Journal of Innovative Education*, 3, 191–203.

Centra,J., (2003) "Will teachers receive higher student evaluations by giving higher grades and less course work?" *Research in Higher Education*, Volume 44, Number 8, pp. 495-518.

Costin, F., Greenough, W. T., & Menges, R. J. (1971). "Student ratings of college teaching: Reliability, validity, and usefulness." *Review of Educational Research, 41*, 511–535.

Del Boca, F. K., & Ashmore, R. D. (1980) "Sex stereotypes and implicit personality theory. II. A trait-inference approach to the assessment of sex stereotypes." *Sex Roles*, 6, 519-535.

Dobbins, G. H., Cardy, R. L., & Truxillo, D. M. (1988) "The effects of purpose of appraisal and individual differences in stereotypes of women on sex differences in performance ratings: A laboratory and field study." *Journal of Applied Psychology*, 73, 551-558.

Heckert, T. M., Latier, A., Ringwald, A., & Silvey, B. (2006). "Relation of course, instructor, and student characteristics to dimensions of student ratings of teaching effectiveness." *College Student Journal, 40,* 195–203.

Isely, Paul & Singh, Harinder (2007) "Does Faculty Rank Influence Student Teaching Evaluations? Implications for Assessing Instructor Effectiveness" *Business Education Digest,* Issue XVI May 2007,47-59.

Feldman, K. A. (1984). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Hi^er Education,* 22(1), 45-116.

Feldman, K. A. (1993) "College students' views of male and female faculty college teachers: Part II – Evidence from students' evaluations of their classroom teachers," *Research in Higher Education, 34,* 151-211.

Hancock, Gregory R., David M. Shannon, and Landa L Trentham (1993) "Student and Teacher Gender in Ratings of University Faculty: Results from Five Colleges of Study," *Journal of Personnel Evaluation in Education*, 6(3): 235-248, 1993.

Kelley, Allen C. (1972) "Uses and abuses of course evaluations as measures of educational output." *Journal of Economic Education* 3(Fall) 13-18.

Kohn, Jonathan & Hatfield, Louise (2006), "The Role Of Gender In Teaching Effectiveness Ratings Of Faculty," *Academy of Educational Leadership Journal*, September.

Liaw, S-H., & Goh, K-L. (2003). "Evidence and control of biases in student evaluations of teaching" *The International Journal of Educational Management*, 17(1), 37-43.

Marsh, H. W. and L. A. Roche (2000) "Effectiveness of Grading leniency and Low workload on Students' Evaluation of Teaching: Popular Myths, Bias, Validity or Innocent Bystanders?" *Journal of Educational Psychology,* Volume 92, Number 1, pp. 202-228.

McKenzie, R. B. (1975). "The economic effects of grade inflation on instructor evaluations: A theoretical approach." *Journal of Economic Education* 6(Spring): 99-105.

Mobley, W. H. (1982) "Supervisor and employee race and sex effects on performance appraisals: A field study of adverse impact and generalizability," *Academy of Management Journal*, 25, 598-606.

Mohan (2011) "On the Use of Non Tenure Track Faculty and the Potential Effect on Classroom Content and Student Evaluation of Teaching" *Journal of Financial Education*, Spring/Summer 2011, 29-42.

Nargundkar, S., & Shrikhande, M. (2012) "An Empirical Investigation of Student Evaluations of Instruction – The Relative Importance of Factors", *Decision Sciences Journal of Innovative Education* Volume 10, Issue 1, pages 117–135, January 2012.

Peterson, Richard L., Berenson, Mark L., Misra, Ram B. and Radosevich, David J. (2008) "An Evaluation of Factors Regarding Students' Assessment of Faculty in a Business School," *Decision Sciences Journal of Innovative Education,* Volume 6 Number 2, pp. 375-402.

Reid, Landon, D. (2010) "The Role of Perceived Race and Gender in the Evaluation of College Teaching on RateMyProfessors.com" *Journal of Diversity in Higher Education*, Vol. 3, No. 3, 137-152.

Sprague, Joey, Massoni, Kelley (2005) "Student Evaluations and Gendered Expectations: What We Can't Count Can Hurt Us**,"** *Sex Roles,* Volume 53, Numbers 11-12, December, pp. 779-793(15).

Tatro, C. N. (1995). "Gender effects on student evaluations of faculty." *Journal of Research & Development in Education, 28,* 169–173.

Zangenehzadeh, H. (1988). "Grade inflation: A way out." *The Journal of Economic Education, Vol. 19,* 217–226.