1-2014

Norming of Student Evaluations of Instruction: Impact of Non-Instructional Factors

Satish Nargundkar
Georgia State University, snargundkar@gsu.edu

Milind Shrikhande
Georgia State University, mshrikhande@gsu.edu

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

*Corresponding author

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

ABSTRACT

Student Evaluations of Instruction (SEIs) from about 6000 sections over four years representing over 100,000 students at the college of business at a large public university are analyzed, to study the impact of non-instructional factors on student ratings. Administrative factors like semester, time of day, location, and instructor attributes like gender and rank are studied. The combined impact of all the non-instructional factors studied is statistically significant. Our study has practical implications for administrators who use SEIs to evaluate faculty performance. SEI scores reflect some inherent biases due to non-instructional factors. Appropriate norming procedures can compensate for such biases, ensuring fair evaluations.

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

INTRODUCTION

Student Evaluations of Instruction (SEIs) are now commonplace among universities as a key mechanism for getting feedback regarding teaching practices. According to Seldin (1993), 86% of US colleges and universities use SEIs to make key decisions about faculty. 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 a 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 norming procedures must be in place to aid comparative interpretation of the data. Norming requires the identification of systematic biases in the ratings of overall instructor effectiveness (OIE) due to non-instructional factors.

A clear understanding of the impact of non-teaching related factors is necessary to ensure fair evaluation of faculty. For example, if a factor like class size significantly affects overall ratings on an SEI for an instructor, then there should be a norming process used by administrators which compensates for class size differences when evaluating faculty. 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 (Isley and Singh, 2005) and business (Peterson, Berenson, Misra & Radosevich, 2008; Isley and Singh, 2007; Liaw and Goh, 2003). The non-teaching related factors can be classified as student related, instructor related, course related, and administrative or situational (Peterson et al., 2008; Pounder, 2007). 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), and

course level (graduate vs. undergraduate). Administrative factors influencing SEI ratings include class size, location, classroom and equipment, and time of day.

Some researchers believe that student grade expectations are positively correlated with SEI ratings (Zangenehzadeh, 1988), while others argue the opposite (Marsh and Roche, 2000). 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 on SEIs. After controlling for learning outcomes, expected grades did not affect student evaluations.

Studies on teaching innovations demonstrate that a good innovation leads to improved student motivation and engagement, resulting in better student performance (Snider and Eliasson 2013; Bergquist and Maggs, 2011). Better student performance is in turn positively correlated with higher instructor effectiveness ratings (Davis, 2009). It is therefore plausible that improved teaching results in an increase in grade expectations as well as better student evaluation of teaching effectiveness.

The focus of this paper is on the impact of non-instructional factors on student evaluations. We therefore exclude grade expectation from the study, since it is sufficiently intertwined with teaching ability to be considered a non-instructional factor.

Research Question

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 integrative studies to look for

consistent patterns across universities and disciplines, or examine the differences as they appear. The non-instructional factors, especially administrative ones, are likely to be different in each institution, and a fair evaluation requires examination of the data at various institutions. This study focuses on SEIs from the College of Business at a large research university spanning across four years and 10 different departments.

We examine the following key research question:

> Do the non-instructional factors (such as course type and level, instructor rank and gender, semester, time of day) have a significant effect on the overall instructor effectiveness (OIE) ratings?

If these factors are significant, and if the impact is large enough, they should be used for norming purposes when comparing faculty performances. The rest of the paper is organized into the following sections: literature review, methodology, discussion of results, and reflections.

LITERATURE REVIEW

There is a debate in the literature about the validity of using student evaluations of instruction (SEI) for assessment of teaching. As some researchers argue, the goal of teaching is to improve student learning. Therefore, the learning must be measured, not the intervention. However, according to recent surveys of research on SEIs, most variables that correlate with student ratings of instruction are also related to instructional effectiveness and student learning (Benton and Cashin, 2012). Benton, Douchon & Pallett (2013) found self-ratings of student learning to be positively correlated with student performance. Students who rate instructors higher also perform

better on exams, and are better able to apply course material and show greater interest in pursuing the subject in later years (Davis, 2009).

One question goes beyond the validity of the instrument to ask if there are systematic biases due to factors that are extraneous to the student evaluation instrument. Michael Scriven (2011) argues that an evaluation instrument must be credible as well as valid, with credibility referring to the audience's estimate of the validity. He states,

> "… *evaluation design must sometimes involve considerations that go beyond validity. This must not be viewed as pandering to prejudice, but as of the essence of certification, of accountability, in a more general sense of the educational and social obligations of the evaluations. ("It is not enough that justice be done, it must also be the case that it must be seen that justice is done*.")."

In the context of higher education, norming of teaching effectiveness scores obtained from SEIs is the way to ensure that justice is done (and seen to be done) in evaluating faculty. If there are factors that bias the teaching effectiveness scores, then such biases must be compensated for. The factors causing such biases can be broadly categorized as Course Related, Instructor Related, and Administrative (Peterson et al., 2008; Pounder, 2007; Feldman, 2007).

Course Related Factors

Davies, Hirshberg, Lye, Johnson and McDonald (2007) studied the impact of several non-instructional factors on instructor ratings in a study of undergraduates in Australia. They found course related factors such as the quantitative nature of a subject to have a significant effect.

Costin, Greenough and Menges (1971) studied ratings by class designation and found instructors receiving higher ratings from seniors than from 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 more familiar with the professors from whom they have taken earlier classes and therefore have reduced anxiety.

Student ability and initial liking for the subject have an impact on instructor ratings (Aigner and Thum, 1986). 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. Feldman (2007) found that students in major courses rated instructors higher than students in non-major courses. Also, students in elective courses rated instructors higher than those in required courses. Expecting ratings for graduate courses to be higher than undergraduate, and non-core higher than core, Brightman, Elliott and Bhada (1993) used four categories – undergraduate core, undergraduate non-core, graduate core and graduate

non-core – based on course level (undergraduate, graduate) and course type (core, non-core) to norm SEI data.

Instructor Related Factors:

Gender differences in performance evaluations in various fields have been studied extensively in the literature (Arvey, 1979; Dobbins, Cardy and 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. Positive characteristics of stereotypical men include rationality, competence and assertiveness, while for women warmth and expressiveness were seen as the main positive traits (Del Boca and Ashmore, 1980). Sprague and Massoni (2005) 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. Table 1 below summarizes the conflicting findings regarding the ratings of male and female instructors:

Rated higher than male instructors	Centra (2009) – attributed to reasons other than bias. Feldman (1993) – rated higher by female students.
Rated lower than male instructors	Lackritz (2004) Heckert, Latier, Ringwald and Silvey(2006) Tatro (1995) Mohan (2011)
No gender difference found	Bauer and Baltes (2002) Blackhart, Peruche, DeWall and Joiner(2006) Centra and Gaubatz (2000) Reid (2010) Hancock, Shannon and Trentham (1993) Kohn and Hatfield (2006)

Table 1: Gender differences in student ratings

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 and Lawrence, 1986). Isley and Singh (2007) found that while higher expected grades result in more favorable student evaluations, this relationship is significantly different depending upon faculty rank. Adjunct faculty ratings are most affected by student grade expectations, followed by tenured faculty, and lastly by tenure track faculty. 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. Feldman (2007) reports higher ratings for higher ranked faculty compared with those of lower ranked faculty.

Administrative Factors

Several researchers have documented an absence of relationship between class timing and student ratings of instruction (Benton and Cashin, 2012; Aleamoni, 1981; Feldman,1978). However, Peterson et al. (2008) found better ratings for daytime classes than evening classes. They attribute the finding to either higher expectation from students who work during day and taking evening classes, or to these students resenting being given homework that adds to their several preoccupations. They also found no evidence of any difference between spring and fall semester ratings.

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. Anecdotal data suggest that there might be a relationship between the quality of classroom facilities and the ratings of instruction. No research has looked into this aspect.

There is some evidence in the literature indicating a relationship between class size and student ratings, with lower class sizes yielding higher ratings (Feldman, 1984, 2007; Liaw and Goh, 2003; Isley and Singh, 2007). For class sizes under 80, there is a relatively steep price to be paid for each additional student in terms of loss of ratings (Bedard and Kuhn, 2008). The difference in ratings per additional student is not so great in larger class sizes (80-150 students). On the other hand, some research finds U-shaped ratings with small and large class sizes yielding higher ratings than class sizes in between, due to a selection bias where teachers known to be good are assigned the really large classes (Wood, Linsky and Straus, 1974; Marsh, Overall and Kesler, 1979). In general, instructors believe smaller class sizes are easier to engage, and therefore result in higher ratings.

METHODOLOGY

We collected data on all student evaluations filled out between 2005 and 2009 in the college of business at a large public university. About 6000 sections of various courses were taught during this period at the undergraduate and graduate levels. Table 2 shows the number of sections taught in each year, segmented into four categories based on course type and course level – graduate non-core (GN), graduate core (GC), undergraduate non-core (UN), and undergraduate core (UC).

Year	GN	GC	Grad Total	UN	UC	UG Total	Grand Total
2005	131	74	205	199	151	350	555
2006	323	225	548	489	406	895	1443
2007	346	199	545	494	416	910	1455
2008	303	200	503	516	437	953	1456
2009	240	124	364	293	258	551	915
Grand Total	**1343**	**822**	**2165**	**1991**	**1668**	**3659**	**5824**

Table 2: Number of sections taught in the business school by year and by category

Data from four academic years starting 2005-06 and ending with 2008-09 was analyzed. Roughly 1450 sections were offered every year, with about a third of them being graduate classes. PhD classes were eliminated from our analysis, since they tend to be very small in size, and sufficiently different from typical undergraduate or graduate courses. The average enrollment per section was 28.36, and the average number of responses to the SEIs per section was 18.20. The response rate for the SEIs overall across the four year span was roughly 64%, 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. Table 3 shows the number of student responses to the SEIs by year and by category.

Year	GN	GC	Grad Total	UN	UC	UG Total	Grand Total
2005	1805	1163	2968	3535	3561	7096	10064
2006	4383	3374	7757	8425	9613	18038	25795
2007	4290	3198	7488	8828	10211	19039	26527
2008	3786	3295	7081	9500	10450	19950	27031
2009	2955	2042	4997	5130	6430	11560	16557
Grand Total	**17219**	**13072**	**30291**	**35418**	**40265**	**75683**	**105974**

Table 3: Number of responses to the SEIs by year and by category.

The Student Evaluation of Instruction (SEI) instrument used at this college is a modified version of one developed and originally validated at UC Berkeley. The modified version was validated at this college over 20 years ago by Brightman, Bhada, Elliott and Vandenberg (1989). More recently, Nargundkar & Shrikhande (2012) found the instrument to still be valid. The instrument consists of 33 question items pertaining to various teaching related factors, and question 34 addresses the overall instructor effectiveness (OIE). In this study we use the OIE ratings (based on a 5-point Likert scale, along with information regarding the non-instructional factors. The non-instructional factors are listed below in Table 4 along with the possible values for each of them.

Factor	**Values**
Semester	Fall, Spring, Summer
Time of day	Morning (starting before noon) Afternoon (starting before 4:30 PM) Early Evening (starting before 7:00 PM) Evening
Course Type and Level	Graduate non-core (GN) Graduate core (GC) Undergraduate non-core (UN) Undergraduate core (UC)
Instructor Gender	Female, Male
Instructor Rank	Tenured Non-tenure Track (NTT) Part time instructor (PTI) Graduate teaching assistant (GTA) Tenure Track (TT)
Class Location	Aderhold Brookhaven Alpharetta Classroom South General Classroom Building Sparks Hall
Class size	Numeric variable with the number enrolled.

Table 4: Non-instructional factors used in the study

Dummy variables were created to indicate various subgroups for time of day, location, rank, gender, course type and course level, and a regression analysis performed with the OIE score as the dependent variable, and the dummies as well as the class size as the independent variables.

The current norming process at our college involves using four segments initially proposed by Brightman et al. (1993) - undergraduate core, undergraduate non-core, graduate core and graduate non-core. The impact of various non-instructional factors was therefore analyzed individually, within each of the four segments. Average scores for OIE for each non-instructional factor within all four segments were compared using 2-sample t-tests and ANOVAs. The variances in the subgroups were not significantly different, making the use of t-tests and ANOVA appropriate. Where ANOVAs were significant, Tukey's two-way comparisons helped to determine specific differences among subgroups.

RESULTS

In order to examine the impact of all the non-teaching factors taken together on the overall rating of instruction, a regression was performed on the entire dataset. OIE score was used as the dependent variable, and dummy variables were created for the categorical independent variables to represent the semester, time of day, location, course level and course type, instructor rank, instructor gender, and class size. Table 5 shows the final model with the significant variables.

	Adjusted R Square	0.0390964			
	Standard Error	0.5276773			
	Observations	5996			
		Coefficients	*Standard Error*	*t Stat*	*P-value*
	Intercept	4.3709	0.0253	172.4805	0.0000
	Spring	0.0479	0.0155	3.0925	0.0020
	Summer	0.1230	0.0184	6.6978	0.0000
	Morning	-0.0568	0.0202	-2.8133	0.0049
	Afternoon	-0.1040	0.0210	-4.9591	0.0000
	Early Evening	-0.0969	0.0176	-5.4925	0.0000
	UC	-0.0478	0.0182	-2.6305	0.0085
	GC	-0.0900	0.0208	-4.3240	0.0000
	Tenured	0.0433	0.0228	1.9046	0.0569
	NTT	0.0752	0.0223	3.3723	0.0008
	PTI	-0.0652	0.0254	-2.5666	0.0103
	GTA	-0.1268	0.0317	-3.9979	0.0001
	Numb Enroll	-0.0018	0.0004	-4.2404	0.0000

Table 5: Regression of Q34 on non-instructional factors. Highlighting is to show groups of dummies for a given variable together.

As seen above, overall ratings for summer and spring are significantly higher than for fall, summer ratings being the highest. Similarly time of day seems to matter, with each of the three times shown scoring less than the evening classes, with afternoon classes scoring the least. Core classes in general score lower than non-core, with graduate core scoring the least. Differences in faculty rank were also significant, with non-tenure track faculty scoring the highest and graduate teaching assistants the lowest.

Given the significance of all these factors in the presence of the others, we examine each non-instructional factor separately, as has been done by various researchers.

Course Type and Level

Tables 6 shows the results of a 2-sample t-test for the mean OIE scores (Likert scale, 1=low, 5=high) for core and non-core classes.

Course Type	
Core classes	4.239 n=2490
Non-Core classes	4.320 n= 3334

$p < 0.001$

Table 6: OIE ratings by type (Core vs NC) overall

Tables 7 shows the results of a 2-sample t-test for the mean OIE scores (Likert scale, 1=low, 5=high) for graduate and undergraduate classes.

Course Level	
Graduate classes	4.315 n= 2165
Undergraduate classes	4.268 n=3659

$p < 0.01$

Table 7: OIE ratings by 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, consistent with expectations. Based on the above findings as well as Brightman (1993) results, four segments were created based on the combination of course level and the course type dimensions, rather than looking at each dimension independently. The results are shown in Table 8 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 8: OIE ratings by segment - course level and type combined

Looking at the rows in the table, the ratings are not significantly different for undergraduate and graduate core classes. Among non-core classes, however, ratings for graduate classes are significantly higher than for undergraduate classes. Looking at the columns in the table, ratings for non-core classes are higher than core classes in both the undergraduate and graduate segments. These findings are a little different from those in the regression analysis, which controls for all other factors.

Instructor Gender and Rank

Table 9 below summarizes our findings regarding instructor gender within each of the four segments

	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 9: OIE Ratings by Instructor Gender by segment.

For the core segment, no significant differences were found between male and female instructors. For the non-core segment, the ratings for female instructors were higher than for male instructors among undergraduate students, while the reverse was true among graduate students. There was no difference between the male and female instructor ratings when all four segments were combined.

Table 10 below summarizes the results of OIE ratings by faculty rank.

	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 10: OIE Ratings by Faculty Rank within each Segment

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).

Semester, Time and Class Size

Overall ratings in the regression were found to be significantly higher during summer compared to spring, and likewise significantly higher for spring compared to fall. Examining the impact of semester within the four segments, we found the following results (Table 11):

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

Table 11: OIE Ratings by semester for each of the four segments

Among undergraduate core classes, summer ratings were significantly higher than for spring and fall. There was, however, no significant difference in ratings for core graduate classes, perhaps due to the lower sample size in that category. Among undergraduate non-core classes, summer ratings were significantly higher than for spring, which were significantly higher than for fall. For graduate non-core classes, summer ratings were significantly higher than for fall, but ratings for spring were not significantly different from either fall or summer.

To test for differences in ratings for sections taught at various times during the day, the day was divided into four time segments. 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 12 below.

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

Table 12: OIE Ratings by Time of day by segment

The results are mixed. Undergraduate core classes show no difference overall, while undergrad non-core do better in the morning and early evenings. 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. There was no difference in overall ratings between the four times of day when all four segments were combined.

Finally, a scatter plot of OIE ratings vs. class size is shown below in Figure 1.

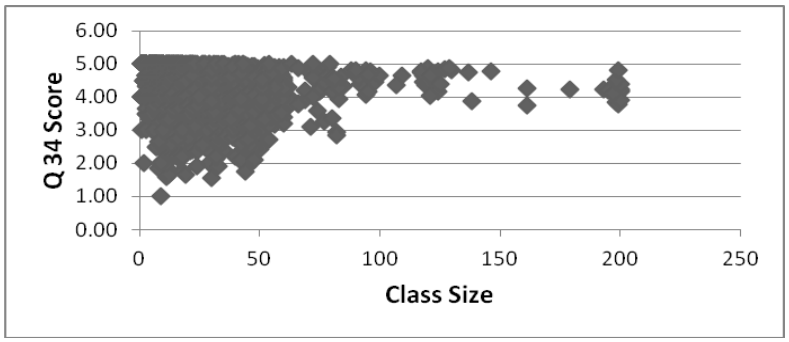

Figure 1: OIE Rating by Class Size

It is difficult to discern a relationship between the two variables from the plot, given the high density of points. The only visible pattern seems to be a slightly downward trend among the very large class sizes (over 100).

The average class size was 28.36. We tested for differences in ratings between class sizes of 30 and below with class sizes over 30. Table 13 below shows the results.

	Class size <=30	Class Size >30
Mean	4.34	4.24
Standard Deviation	0.5515	0.5123
Sample Size (number of sections)	3596	2400

$p < 0.001$

Table 13: OIE Ratings and Class Size[1]

The overall ratings for the smaller class sizes were significantly higher than for the larger ones.

DISCUSSION

Instructor ratings are significantly different for course related factors like the course level and type. Ratings are higher for non-core classes compared to core classes. This is consistent with our expectations based on the literature. It seems to be fairly well established that initial liking for a course does in fact affect the ratings of an instructor. Graduate classes overall get better ratings than undergraduate classes. Graduate students are generally expected to be better prepared and have a greater liking for the subject than undergraduates. Among core classes, there is no difference in ratings for undergraduate and graduate classes. However, among non-core classes, there is a difference between the two.

[1]We also compared class size 20 and under with class size 21-39 and class size 40+ with an ANOVA. The results were uniformly in the same direction, with higher overall ratings for smaller class sizes.

Among core classes, there is no significant difference in ratings between male and female instructors. However, we see an interesting effect in the non-core classes. Undergraduate students rated female instructors higher than male instructors, while graduate students rated male instructors higher than female instructors. Younger students may prefer the nurturing characteristics attributed to female instructors. Similarly, the older graduate students perhaps prefer the perceived stereotypical qualities among male instructors of being forceful and goal driven.

Instructor rank or status also has an impact on overall ratings. In all four segments, non-tenure track (NTT) instructors consistently show higher ratings than untenured tenure track (TT) faculty. However, tenured faculty performed very well, especially in graduate classes. Among undergraduate classes, part time instructors (PTIs) have better ratings than untenured TT faculty. In our opinion, this finding is consistent with the incentive structure in place for faculty at research institutions. NTT faculty is primarily evaluated on teaching effectiveness, while TT faculty is evaluated primarily on research, with lower emphasis on teaching. However, when they do get tenure, the emphasis on research is reduced, giving them time to focus on teaching.

The influence of administrative factors like semester, time of day and location (classroom quality) on overall ratings of instructors was mixed. Summer semester ratings are consistently higher than the ratings for spring or fall, with being graduate core classes being the only exception. Summer classes on average have around 20-25 students, while fall and spring classes have 30+ students on average. The regression analysis shows the effect of the semester to be

significant even after controlling for the class size effect. An explanation for better summer ratings may be that students take fewer classes during summer, allowing greater focus on those classes. Further, frequent meetings during summer may build a better rapport with the instructor and better retention of material.

As for time of day, the regression shows a progression of rating differences, with instructors being rated the highest for evening classes, followed by morning, early evening, and afternoon classes respectively. When the effect of timing was examined by itself for each of the four segments, we find some differences. Within the graduate core, morning classes receive a higher rating than evening, and not many classes are offered in the afternoon. Also, many of these morning courses are offered on Saturdays, when the graduate students are relatively free from work related pressures. Within the undergraduate core, morning and early evening classes scored higher than afternoon classes, consistent with our expectation based on tiredness/sleepiness after lunch. Finally, in the graduate non-core, evening classes score higher than early evening (there are very few classes taught in the morning or afternoon). 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. None of the classroom location variables came in significant in the regression. In other words, location (and by proxy, classroom quality) did not affect OIE ratings.

Class size effect on OIE ratings is consistent with recent literature. Smaller class sizes have significantly higher ratings than larger ones. We first tested class sizes under 30 against 30+, since it was close to the overall average class size of a little over 28. To see if there was a hint of

a U-shaped relationship as indicated by Wood et al. (1974), three groupings of class size - less than 20, 21 to 40 and 40+ were also tested. The results were unidirectional, with larger classes getting lower ratings on average.

CONCLUSION

As Brightman (2005) points out, in order to effectively use SEIs for assessment, the instrument must first be valid. The validity of the instrument used at the College of Business of this large public university was established by Brightman et al. (1989) and the instrument was revalidated in recent times by Nargundkar and Shrikhande (2012). Further, the results of the SEIs should be appropriately normed for fair feedback to faculty. In other words, the impact of non-instructional factors on overall ratings of instruction must be controlled for in evaluating faculty. Non-instructional factors are by definition not relevant to one's teaching ability or effectiveness, and are beyond the instructor's control. However, these factors have the ability to bias an instructor's effectiveness ratings, as shown in this paper. This has a major implication for administrators evaluating faculty.

Based on our findings, administrators should look at various non-instructional factors when assessing faculty performance through student evaluations. At our business school, the four segments currently used for norming (undergraduate core/non-core, graduate core/non-core) by administrators are appropriate, given the results of this study. However, this study suggests that they are insufficient, and that several additional factors, namely, semester, time of day, instructor gender and rank and class size also need to be considered. Based on our regression model, an instructor with an average score of 4.37 that happens to hit upon an adverse combination of these

factors can in the worst case end up with a score of 4.05, while an instructor that hits upon the best combination of these factors can end up with a score of 4.57. In other words, two instructors with identical teaching effectiveness could get overall student ratings that differ by as much as 0.52 on a scale of 1 to 5. Given that most SEI ratings vary between 3.0 and 5.0 (a range of 2.0), a difference of 0.52 due to extraneous factors can be drastic. This implies that an administrator's perception of an instructor's effectiveness has the potential to be distorted to a significant degree by non-instructional factors beyond the instructor's control.

For other colleges, the implication of our study is that norming is essential, and administrators at each college must identify the non-instructional factors most relevant to norming in their institutional setting. Such a study is worth doing at every college that uses SEIs to evaluate faculty. The non-instructional factors we identified as significantly impacting student ratings of instruction may be specific to our institution alone.

Recent research (Benton and Cashin, 2012) suggested that it is a misconception to attribute poor overall ratings to such non-instructional factors. Our results suggest that while non-instructional factors cannot entirely explain poor (or good) ratings, they do have the potential to bias the ratings sufficiently to matter in administrative decisions. Peterson et al. (2008) in their study of a single department within a business school suggest the possibility that instructors may try to game the system by using non-instructional factors to improve their ratings without necessarily improving teaching effectiveness. Appropriate norming procedures can eliminate this problem.

While our study suggests ways to mitigate the distortions caused by non-instructional factors on teaching effectiveness ratings, student evaluations are by no means the only measure of teaching effectiveness and student learning. Many researchers provide ways of guarding against potential bias in student evaluations of instruction (Baldwin and Blattner, 2003). Using alternative approaches such as portfolios, peer feedback sessions, and informal student surveys in addition to SEIs can further help to combat or circumvent these potential biases. Michael Scriven (2011) suggests three models for teacher evaluation in increasing order of desirability. First, a self-assessment by faculty members; second, student evaluation of instructors reported to administrators (the method most commonly adopted); third, an external examiner evaluating student achievement and thereby inferring the efficacy of the teacher.

Overall, the debate in the literature tends to either extol the virtues of SEIs or denigrate them as useless. Our research shows that SEIs can be useful instruments as long as they are validated, and the biases that affect them are accounted for in the evaluation process.

REFERENCES

Aigner, Dennis J. and Thum, Frederick D. (1986) On Student Evaluation of Teaching Ability. *Journal of Economic Education*, Fall 1986, pp. 243-265.

Aleamoni, L. M. (1981) Student Ratings of Instruction. In J. Millman (Ed.*) Handbook of Teacher Evaluation* (pp. 110-145). Beverly Hills, CA: Sage.

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

Baldwin, T. and Blattner, N. (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., and 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, K. and Kuhn, P. (2008) Where class size really matters: Class size and student ratings of instructor effectiveness. *Economics of Education Review*, 27, 253-265.

Benton, S. L. and Cashin, W. E. (2012) Student Ratings of Teaching: A Summary of Research and Literature. IDEA Paper # 50, IDEA Center, Kansas State University.

Benton, S. L., Douchon, D., Pallett, W. H. (2013) Validity of student self-reported ratings of learning, *Assessment and Evaluation in Higher Education*, Vol. 38, No. 4, p. 377.

Bergquist, Timothy M.; Maggs, Anne. (2011) A Bookstore for Bailey: A Novel Approach to Teaching a Small-Business Management Course. *Decision Sciences Journal of Innovative Education.* May2011, Vol. 9 Issue 2, p269-274.

Blackburn, R. T., and 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 teachingevaluations 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, pp. 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.

Centra, J. A. and Gaubatz N. B. (2000) Is there a Gender Bias in Student Evaluations of Teaching? *Journal of Higher Education,* 70, pp.17-33

Centra, J. A. (2009) *Differences in Responses to the Student Instructional Report: Is it Bias?* Princeton, NJ: Educational Testing Service.

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

Davies, M., Hirshberg, J., Lye, J., Johnson, C. & McDonald, I. (2007) Systematic influences on teaching evaluations: the case for caution. *Australian Economic Papers*, Vol. 46 Issue 1, pp.18-38.

Davis, B. G. (2009) *Tools for Teaching*, 2nd Ed. San Francisco, Jossey Bass.

Del Boca, F. K., and Ashmore, R. D. (1980) Sex stereotypes and implicit personality theory.II.A trait-inference approach to the assessment of sex stereotypes. *Sex Roles*, Volume 6, Number 4, 519-535.

Dobbins, G. H., Cardy, R. L., and 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.

Feldman (1978) Course Characteristics and College Students' Ratings of their Teachers: What we know and what we don't. *Research in Higher Education*, 9, pp. 199-242.

Feldman, K. A. (1984). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher 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.

Feldman, K. A. (2007) Identifying Exemplary Teachers and Teaching: Evidence from Student Ratings. In R. P. Perry & J.C. Smart (Eds.), *The Scholarship of Teaching and Learning in Higher Education: An Evidence Based Perspective* (pp. 93-129) Dordrecht, The Netherlands, Springer.

Greenwald, Anthony G. (1997) Validity concerns and usefulness of student ratings of instruction. *American Psychologist*, Vol 52(11), Nov 1997, 1182-1186.

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.

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

Isley, Paul and Singh, Harinder (2005) Do higher grades lead to favorable student evaluations? *Journal of Economic Education*, 36, pp. 29–42.

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

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

Lackritz, James R. (2004) Exploring burnout among university faculty: incidence, performance, and demographic issues. *Teaching and Teacher Education*, Volume 20, Issue 7, October 2004, Pages 713–729.

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., Overall, J. U., and Kesler, S. B.(1979) Validity of Student Evaluations of Instructional Effectiveness: A Comparison of Faculty Self-Evaluations and Evaluation by their Students. *Journal of Educational Psychology*, 71, pp. 149-160.

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.

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 Facultyand 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.

Pounder, J. S. (2007). Is student evaluation of teaching worthwhile? *Quality Assurance in Education, 15*(2), 178-191.

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.

Richardson, John, T. E., (2005) Instruments for obtaining student feedback: A review of the literature. *Assessment & Evaluation in Higher Education*, Vol. 30, No. 4, August 2005, pp. 387-415.

Seldin, Peter (1993) *Successful use of teaching portfolios*, Anker Pub Co (March 1993).

Scriven, M. (2011). Evaluation Bias and its Control. *Journal of Multi Disciplinary Evaluation*, *7*(15), 79-98.

Snider, Brent R.; Eliasson, Janice B. (2013) Beat the Instructor: An Introductory Forecasting Game. *Decision Sciences Journal of Innovative Education*. Vol. 11 Issue 2, p147-157.

Sprague, Joey and 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.

Wood, K., Linsky, A. S. and Straus, M.A. (1974) Class Size and Student Evaluations of Faculty. *The Journal of Higher Education*, 45, 7, October 1974.

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