Higher Education Studies; Vol. 5, No. 6; 2015
ISSN 1925-4741 E-ISSN 1925-475X
Published by Canadian Center of Science and Education

Response Rate and Teaching Effectiveness in Institutional Student Evaluation of Teaching: A Multiple Linear Regression Study

Faisal Al-Maamari[1]

[1] The Language Centre, Sultan Qaboos University, Muscat, Oman

Correspondence: Faisal S. Al-Maamari, The Language Centre, Sultan Qaboos University, Muscat, Oman. P. O Box 43, PC123. Tel: 968-9936-4299. E-mail: faisalf@squ.edu.om

Received: September 10, 2015 Accepted: September 28, 2015 Online Published: November 2, 2015

doi:10.5539/hes.v5n6p9 URL: http://dx.doi.org/10.5539/hes.v5n6p9

Abstract

It is important to consider the question of whether teacher-, course-, and student-related factors affect student ratings of instructors in Student Evaluation of Teaching (SET) in English Language Teaching (ELT). This paper reports on a statistical analysis of SET in two large EFL programmes at a university setting in the Sultanate of Oman. I carried out a multiple regression analysis to address the research questions of whether instructor sex, class size, course type and percent participation would affect teaching effectiveness scores, and whether or not response rate can be predicted by instructor sex, class size and course type. The study utilizes a dataset of over 2000 student ratings obtained from an SET survey covering the period from Fall 2011 through to Spring 2014in these two programmes. Results indicated that the modeled predictors showed extremely low bias towards both teaching quality scores and response rate. Although the effect sizes of these results are extremely small, they are still significant due to the large sample size (comprising over 2000). The findings also suggest that contrary to common parlance in some quarters claiming students' unreliable ratings, this analysis has shown that students can judge teaching effectiveness and do not allow other teacher-, course- and student-related factors to bias their responses. The study's significance stems from the fact that it adds to instructional evaluation in ELT, a field characterized by a clear lack of research on SET.

Keywords: course and teaching survey, higher education, Oman, student evaluation of teaching, SET, response rate

1. Introduction

Student Evaluation of Teaching (SET), wherein upon completing a course of study students undertake an electronic or pencil and paper questionnaire to rate the course and their instructor, has achieved centrality and popularity in higher education (Marsh, 2007). SET has been intensively researched, making it the most area in performance management which "evokes emotional debate" (Macfadyen et al., 2015). Although SET results have historically been put to use formatively in order to offer a developmental perspective (i.e., diagnose teachers' strengths and weaknesses from the perspective of students) in order to improve teaching and learning, they have more recently come to use for quality assurance and evaluative purposes (i.e., in decisions pertaining to salary raise, promotion, tenure, and (re-)appointment).

Abundant research has been undertaken to resolve both conceptual and statistical issues with regards to SET (for major reviews in the field see Marsh, 2007; Spooren, Borckx, & Mortelmans, 2013; Sproule, 2000). Amongst the issues investigated is the multidimensionality of the construct of teaching effectiveness, the validity and reliability of the instrumentarium (i.e., the different forms of SET whether they are institutional or online-based) and the respondent (i.e., in this case, the course students), and the uses to which the SET results are put. An increasingly important area for SET investigation that fits into the validity and reliability of SET is bias in the students' ratings of their course and instructor. The bias question is an important one, and is the focus of this paper.

Conceptually, the manner bias has been construed is not straightforward. Factors which may be biasing in SET can include the instructor's gender, the instructor's language background, the student's initial interest level, the student's prior knowledge about the subject area, course difficulty, course level to name but a few. More recently however, research has reached a clearer conceptual basis on the factors which may confound students' ratings of their instructor, and SET researchers have dubbed those influencing and confounding factors. In some cases, the

distinction made between factors which qualify to bias and those which do not is ultimately a classic chicken egg one. For example, should a factor such as "student attendance" be considered a biasing or a confounding variable? Proponents not saying of the bias question reason the direction of association from teaching effectiveness to students attending class, thereby arguing that a good instructor affects his/her students in such a way that they do not want to miss lectures/classes. Opponents see the directional effect of student attendance on SET, thereby suggesting that low class attendance rate is associated with teaching ineffectiveness or is preconceived in the student's attitude. Therefore, the mainstream conclusion of bias research into SET is that "the same influencing factor may be a confounder or a valid contributor, depending on the underlying construct of 'high quality teaching'" (Schiekirka & Raupach, 2015).

Therefore, there is now agreement that gender (pertaining to instructor and student), the course discipline, halo effects such as personal qualities (e.g., charisma), instructor's race, and instructor's language background can be considered biasing factors. These factors contribute to SET scores when they should not, as they are not indices of teaching quality or effectiveness, and are thus extraneous attributes, irrelevant to the construct.

It is so important to study biasing factors, inasmuch as they do not make the SET score or scores a true measure of teaching quality and effectiveness, which renders the score and the concomitant interpretation and use invalid. Therefore, one of the goals of the study is the determination of the effect of level of (non)participation on two score aggregates obtained from a Course and Teaching Survey (CTS) which is termly conducted in the author's institution through the quantitative exploration of a dataset spanning eight academic semesters from Spring 2011 through to Fall 2014.

It is hoped that the study will add to the theoretical literature of bias factors into SET in English Language Teaching settings specifically, which previously has witnessed a striking lack of SET studies. Institutionally, the study is hoped to further provide statistical evidence for how "faculty members should be evaluated on a comparable basis only in those areas they can affect" or "by a methodology that corrects for those influences beyond the faculty member's control" (Mason, Stegall, & Fabritius, 1995).

2. Literature Review

The replacement of pencil and paper Student Evaluation of Teachers (SET) instruments with electronic ones have led to the increased popularity of SET, but also to the emergence of a major problem, namely the issue of low response rate. This literature review will consider the advantages to using electronic SET, the question of low response rate, its causes and concomitant effects. Next, it will consider the ongoing debate of whether response rate is a bias that leads to error or an argument that enhances the validity of SET. The section will justify the focus as to why it is important to study low response in SET especially in an English for Academic Purpose context, which is clearly characterized by a remarkable lack of focus on SET studies despite the pivotal role EAP programmes play in the higher education arena.

In recent years, the electronic evaluation of teachers appears to have substituted the classic paper-and-pencil questionnaire as the most common institutional means of gathering SET worldwide (Arnold, 2009; Nulty, 2008). This has provided the territory for SET to grow exponentially, and gradually transcend institutional borders to the World Wide Web through student rating instrumentarium sites such as RateMyProfessors.com, ProfessorPerformance.com, and Reviewum.com (Otto, Sanford, & Ross, 2008). The primary reasons given for shifting to electronic SET relate to benefits to students, instructors and institutions, and include (a) greater accessibility to students, (b) student anonymity (e.g., decreased risk of recognition due to handwriting), (c) better/more written comments by students, (d) decreased vulnerability to instructor influence, (e) little disruption of class time, (f) more accurate analysis of the data, (g) efficient (accurate & quick) feedback, (h) lower costs, and (i) reduced time demands for administrators (Bothell & Henderson, 2003).

This favourable movement to electronic mode has brought with it an unfavourable outcome: low response rate. In their major report on the validity of SET, Spooren, Brockx and Mortelmans (2013) warn against relying on internet-based (but also electronic) SET instruments as the only source of gathering data about the quality of instruction and courses as they are prone to self-selection bias because of the low response rate (i.e., sample respondents may not be representative of the whole SET population). Pre-electronic modes required the instructor of the course or a university representative to administer SET to students, and so self-selection was not an issue. It is well-known that "a student's decision to complete a SET is not a random process" (Macfadyen, 2015). This is especially so in electronic SET wherein a constellation of course-, teacher- and student-specific factors coalesce not only to influence the student's decision to partake in the SET process, but also to evaluate instruction. The factors could be related to "student characteristics, teaching (in terms of structure, process and content), satisfaction with examinations and the evaluation procedure itself" (Schiekirka & Raupach, 2015).

Various explanations could be given for the increasingly low response rates in SET. In their review of SET research, Spooren, Brockx and Mortelmans (2013) identify five concerns teachers have about the validity of teacher ratings such as the student's uninformed perspectives about teaching quality, the psychometric properties of the instruments, the impersonal nature of the SET administration, teachers' unfamiliarity with SET research and the complex nature of SET scores' interpretation. Nair and Adams (2009) list student apathy, technical issues, lack of engagement of students, and the little action taken because of students' feedback as major causes of low response. If students' expectations are not met and they do not see that their feedback leads to meaningful contribution, their motivation to participate in subsequent evaluations is negatively affected. The situation is not improved by the fact that administrators prefer aggregated scores of student satisfaction, often failing to consider both basic statistical and methodological matters (e.g., response rate, score distribution, sample size), and over-rely on global scores to make high-stakes personnel decisions (Gray & Bergmann, 2003; Menges, 2000). Finally, others fear that SET results obtained through electronic means are easier to trace and can be accessed by almost everyone (Gamliel & Davidovitz, 2005). These reasons, individual or combined, contribute to students' refraining from participation in SET and some instructors' disbelief in SET.

That response rates in SET are generally low (fluctuating between 30% and 50%), especially in the case of online course evaluations, is well-established in the literature (Arnold, 2009; Dommeyer, Baum, Hanna, & Chapman, 2004). Moreover, although electronic surveys afford numerous advantages as highlighted above, their greatest caveatremains the low student participation. Dommeyer et al. (2004) reported average response rates of 70% for in-class teacher surveys and 29% for electronic surveys. Wennerstrom and Heiser (1992) reported response ratesas high as 62% in the academic ESL programme and 69% in the intensive programmein paper and pencil SETs.

Less established, however, is the effect of low response rate on SET scores. Macfadyen (2015) reports that "declining SET responses rates … continue to fuel academic mistrust and cynicism" thus "allowing critics to call into question the validity of SETs". Beran and Violato (2005) found that various students' characteristics explained 7% of the total variance in SET scores. Smith et al. (2007) noted statistically significant effects of sex of students and sex of instructors on SET scores, but these predictors only accounted for 1% of the explained variance in SET. Further, based on their review of research, Marsh (2007) found that 16 background characteristics explained about 13% of the variance in SEEQ (Student Evaluation of Education Quality) survey. These findings suggest that SET outcomes depend primarily upon teaching behavior (Barth, 2008; Greimel-Fuhrmann & Geyer, 2003). The magnitude at which the response rate interacts with these factors is open to research.

As a response to the low level of participation, institutions have adopted different techniques to encourage students to complete SET instruments. Johnson (2003) suggested several strategies for increasing electronic SET response rates, including encouragement by the faculty (i.e., if instructors show genuine interest in SET, students will be more inclined to participate), increasing the intrinsic motivation of students to participate (e.g., by highlighting their important role as raters), providing access to the electronic evaluation system, and giving clear instructions concerning participation in the SET process. Further, response rate may be enhanced if students see the connection between the feedback they volunteer in SET forms and improvements to teaching, learning and the curriculum (Spooren, Brockx, & Mortelmans, 2013). Therefore, institutions need to be transparent about how much the feedback gained from SETs is used to improve the instructional and learning experiences for the students.

It is important to research low response rates for several reasons. Firstly, current SET research points to the paucity of and "evident lack of literature that has focused on factors of nonresponse in web-based SETs" (Adams & Umbach, 2012). Secondly, studying low participation levels allows us to quantify the amount of error they contribute to the SET score (problems with external validity) and enable us to understand the reasons for the declining rates of completion rates and consequently be more equipped to deal with them. Thirdly, studying low response rates enables us to increase the quality of the data, strengthen any interpretation we make of them and recommend legitimate use (outcome validity) of the results (Adams & Umbach, 2012). Finally, for evaluation data to achieve value, response rates need to be sufficiently high so that they are representative of the student cohort (Bennett & Nair, 2010). In these same terms, Schiekirka and Raupach (2015) argue that as participation in evaluation activities is not mandatory at all universities, self-selection of students providing course ratings might produce biased samples. As a consequence, the reliability and validity of course ratings would depend on response rates.

In studies conducted on SETs, student ratings have revealed bias. For example, female students are more likely to respond than males (Avery et al., 2006; Porter & Umbach, 2006). High performance and achievement (as

measured by grade, cumulative GPA, and SAT score) showed positive association with the likelihood of response (Avery et al., 2006; Porter & Umbach 2006). Adams and Umbach (2012) found that "Not only are students with higher grades typically awarding higher ratings, but they are also the ones who are more likely to respond". Schiekirka and Raupach (2015) found effects of student characteristics and student satisfaction with exams on completion rates in the sense that "female and more motivated students, high-achievers and those who are more satisfied with exams tend to provide more positive course ratings" and therefore they concluded that "any selection procedure favouring these groups might entail inflated ratings".

As a potential cause of the low response rate, Adams and Umbach (2012) hypothesize that it is possible that students perceive non-credit and pass/fail courses as less important and less deserving of their attention and will thus yield low response rates. In ESL/EFL contexts, there is a clear lack of research on SET (Pennington & Young, 1989). Wennerstrom and Heiser (1992) conducted an SET study investigating student bias in two ESL programmes at the University of Washington. They found that systematic bias existed due to learner's ethnic background, level of English language proficiency, course content and attitude towards the course. Chinese, Latin American, Indonesian and Arabic speaking students gave consistently higher ratings compared to their Japanese colleagues in the intensive programme. No significant relationships were found in the academic programme.

Because of the differences between the EAP curriculum and other disciplines and the varying reasons EAP students have for enrolling in EAP programmes in tertiary education, the hypothesis pertaining to the influence of degree (credit-bearing) and non-credit (pass/fail) can be naturally tested, without having to contrive any experimental intervention through the excavation and exploration of an SET dataset comprised of over 2000 entries. The study will also explore the association levels between participation levels and instructor sex, course type, and class size on the one hand and these factors and teaching effectiveness on the other.

3. Method

3.1 Research Questions and Related Hypotheses

The current study seeks to provide further insight into the factors that influence student low response rate on SETs in two EFL programmes, a foundation (intensive) English language programme of six levels from beginning to advanced, and a credit-bearing English for Academic Purposes programme (EAP). Few studies have explicitly focused on these EAP programmes, and so SET investigations of this kind are not heard of in the literature in spite of the unique position EAP programmes of study have in institutions of higher education. Further, little exists on researching the factors of instructor gender, course type and class size (i.e., students' enrollment in a section of a particular course) to predict teaching effectiveness. Therefore, this study aims to address this gap. Based on research conducted elsewhere as reviewed above, this study posits that it is more likely that the response rate in non-degree programmes is lower than the percent level of participation in degree programmes.

3.2 Variables

The study is premised on the statistical manipulation of six variables. These are shown in Table 1 below.

Table 1. The six-variable research design

Variable name	Variable type	Measurement type
Instructor gender	Predictor	Nominal
Course type	Predictor	Nominal
Class size	Predictor	Interval
Percent participation	Predictor/ Outcome	Interval
Aggregate teaching mean	Outcome	Interval
Instructor mean	Outcome	Interval

Three are independent variables, and these are class size (scale type variable which ranged from 6 to 34), course type (binary variable where 0 indicated non-degree, non-credit, non-grade courses and 1 indicated courses that are part of the undergraduate degree, are for credit and are normally graded) and instructor gender (binary variable where 0 indicated Female and 1 indicated Male). Two are dependent variables: they are total teaching

aggregate score mean (out of total of 4 points) and a 4-point instructor global score, or what is known as SMIQ (the Single Most Important Question) according to Sproule (2000). The sixth variable is "percent participation levels", which is an independent variable in the prediction of teaching effectiveness, and it is itself also a dependent variable to be predicted by instructor gender, course type and class size. In the data file, percent participation is a computed variable created by multiplication of the number of respondents into the total number of enrollment, both of which are given in the original CTS data file.

3.3 Subjects

The data file utilized a total of 6 variables to answer the research questions and test the posed hypothesis. The descriptive statistics for the full sample, including missing data, are given in Table 2.

Table 2. Descriptive statistics for the six-variable design

	N	Count	Min.	Max.	M	SD
Instructor Gender	2061	M (845); F (1216)	0	1	.41	.492
Course Type	2095	FP (1266); CELP (829)	0	1	.40	.489
Class Size	2095		6	34	18.88	3.43
Percent Participation	2095		30	100	62.41	21.08
Teaching Mean	2095		1.63	4.00	3.27	.317
State. 15 Mean	2095		1.50	4.00	3.40	.387
Valid N (listwise)	2061					

From the table, the male teachers represent 40.33% and the female teachers comprise 58.04% with 1.63% representing data where instructor gender could not be obtained. Foundation programmes make up 60.4% and credit programmes make up 39.6%. Therefore, the analysis will not include the 34 cases with the missing gender data and will therefore be based on the 2061 valid data.

3.4 Instrument and Institutional Baseline Response Data

The teacher evaluation survey used in the researcher's institution is called the Course and Teaching Survey (CTS, henceforth) which consists of 13 (Note 1) close-ended statements that ask students to rate both the course and the instructor on a variety of measures using a five-point semi-Likert Scale (1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree, 5 = Not Applicable). The survey has three statements designed to measure course characteristics (course assessments, course objectives, written instructional materials), and eight statements designed to measure instruction (feedback, teaching aids, teaching approach, student assistance, instructor availability, explanation of material, curriculum coverage, punctuality, overall learning). There is one global question (SMIQ) to evaluate the instructor's overall teaching performance (i.e., "15. Overall, this instructor is a good teacher."). In addition to these discrete statements, the CTS comprises three open-ended questions whose aim is to elicit from students their comments on the course and the instructor and to give them the opportunity to suggest improvements. However, those invoke qualitative type comments, and are therefore not included in this study.

The procedure followed in the administration of the CTS typically follows the same procedure outlined by Sproule (2000) for administering institutional surveys, and it involves the following basic steps. Centrally, three weeks before the conclusion of the semester, the institution sends an email to students and instructors, which directs students to the page where the CTSs reside. At this site, the students enter their university credentials (username and password) and are subsequently taken to the CTSs for all the courses in which they are enrolled for the semester. Upon the selection of the appropriate CTS, the student then submits the completed form, which is stored centrally in the University's office. Throughout the three week evaluation period, the university sends e-reminders as is required. Instructors are also reminded by heads of departments to encourage their students to complete the surveys. Analysis is done, and two weeks after the conclusion of the final examination period the CTS analysis is promptly released for each instructor; aggregate scores (i.e., the overall teaching mean and the instructor mean) are also sent to heads of departments.

It is important to note that in accordance with the institution's commitment to students on data privacy, access to student data such as those that could identify the students and their comments to the instructors is protected and is prohibited. The implications of this for this study are that any student data supplementing the original CTS file are not readily available to the researcher. However, the availability of other records enabled me to acquire the additional data through the juxtaposition of another database which had entries on instructor gender and course type. Class size and response rates are provided in the generic CTS data file. The University sets a 30% response rate, below which ratings are considered invalid by the institution, and are eliminated from the original file. Therefore, CTS scores which are reported for the main study are only those that satisfy this condition of 30%. The final dataset file compiled by the researcher contained numerous other student, teacher and course variables such as nativity and ethnicity of the instructor, the course academic (proficiency) level and medium of instruction of the student's chosen specialization, but these are not included as part of this analysis. A total of six variables as previously described (Tables 1 & 2) are extracted for analysis in this paper. Therefore, to summarize, the aims of this study are to identify:

1) The association between instructor gender, course type and class size and percentage of students (not)-completing the surveys.

2) The association between instructor gender, course type and class size and both the aggregate teaching mean and the good instructor statement mean as measures of teaching effectiveness.

3) The association between percent participation and both the aggregate teaching and instructor means.

Figure 1 is extracted from the dataset in order to illustrate the *status quo* with regards to CTS completion over a four-year period extending from Spring 2011 through to Fall 2014.

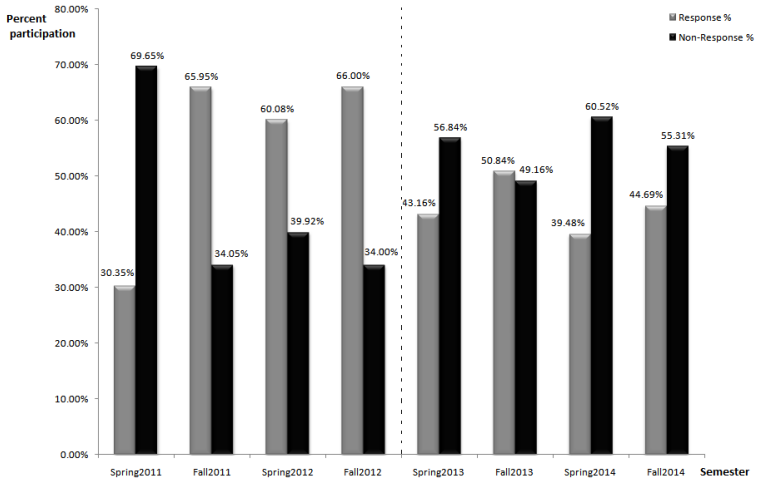

Figure 1. Response and non-response percentage in CTS over four years

As can be seen, the broken line divides the graph into two four-semester groups. The right hand division is a more accurate calculation of (non)-response percentage because it further incorporates the courses, which do not satisfy the 30% minimum cut-off point for the validity of the CTS based on the institution, as was explained above. At the time, the invalid data for the left hand division could not be obtained. In general, the graph shows that there is a high non-response rate ranging from 34% to 69.65%. Further, looking at the response rate from Spring 2013 through to Fall 2014, it is clear that the non-response rate, though varying, is higher than the response rates for these two years. The question that remains to be asked is to what extent the non-response percentage introduces bias which affects both the aggregate teaching mean and the good instructor statement mean, and to what extent percent participation could be predicted from instructor gender, course type and class size.

3.5 Statistical Analysis

Before conducting the main analyses related to the research questions, I conducted several preliminary analyses to determine if there are any differences between instructor gender, course type, teaching and instructor means and percent participation. I next ran simple correlations amongst the variables, followed by multiple regression

analysis to test for bias in participation levels and teaching effectiveness due to the aforementioned variables. MLR methodology allows for the inclusion of various variables whilst holding others constant, and so it is an appropriate methodology for this study.

A linear models analysis confirmed that the course type effect was significant for all four dependent measures: $F(2,094) = 100.74$, $p < 0.001$ for the class size; $F(2,094) = 269.27$, $p < 0.001$ for percent participation; $F(2,094) = 139.71$, $p < 0.001$ for teaching mean; and $F(2,094) = 227.5$, $p < 0.041$ for statement 15 mean. Similarly, although less so here for gender differences, there were significant differences in percent participation, $F(2,060) = 4.2$, $p = 0.043$, teaching mean, $F(2,060) = 4.3$, $p = 0.039$, and statement 15 mean, $F(2,060) = 4.3$, $p = 0.039$. However, class size, $F(2,060) = 3.2$, $p = 0.074$, showed no significant differences. The p-value difference suggests that the effect is greater for course type than it is for instructor gender. Further, to test for the normality of the dependent variables, the Q-Q plots for class size, percent participation and teaching and statement 15 means showed that the observed values fitted approximately the expected line. Checking the assumption of linearity was not done through skewness and kurtosis measures, Kolmogorov-Smirnov and Shapiro-Wilk tests, and/or a visual inspection of histograms, as these tests are not sensitive to large sampled data (Tabachnick & Fidell, 2007) such as these used in the study. Therefore on the basis of the plots, I conclude that the variables are approximately normally distributed and thus seemed to display favourable characteristics for subsequent analysis.

Further, a Pearson correlation coefficient matrix revealed that the three independent variables for the regression analysis (instructor gender, course type and class size) had positive significant correlations with percent participation, teaching mean, and statement 15 mean, the dependent variables. Additionally, multi-collinearity statistics have shown that the tolerance values are larger than 0.95 for all variables; therefore, the three independent variables (instructor gender, course type and class size) are independent of one another.

4. Results

4.1 Predicting Percent Participation from Instructor Gender and Course Type

The first question relates to identifying the relationship between percent participation and both instructor gender and course type, the latter of which pertains to whether or not a course belongs to a degree or non-degree programmes. Therefore, I regressed percent participation against the two predictor variables using a forced entry regression analysis which forces all specified variables into the equation. Table 3 shows the output for this regression.

Table 3. MLR model summary of participation as an outcome variable and instructor gender and course type as predictors

Model	R	R^2	Adjusted R^2	SE of the Estimate	R^2 Change	F change	Sig. F Change
1	.045[a]	.002	.002	21.08	.002	4.11	0.043
2	.230[b]	.053	.052	20.54	.051	110.2	0.000

					95% Confidence Interval for B	
Model	Beta	t	Sig.	SE of the Estimate	Lower Bound	Upper Bound
Course type	-.225	-10.50	0.000	21.08	-11.51	-7.88

a. Predictor: (Constant), Instructor Gender

b. Predictors: Instructor Gender, Course Type

Model 1 results show that instructor gender explains 0.2 % (0.002) of the variance in percent participation, and that the change in the F test from this predictor is statistically significant at the 0.05 level. Model 2 results show that course type is also a significant predictor (as determined by the "Sig. F Change" results). Course type alone explains 5.1% of the variance in percent participation. All in all, whilst instructor gender contributes little to percent participation, course type is a better predictor of participation levels. Although the effect size is small, the F change value is significant at 0.001 level. Because instructor gender is not a strong predictor, no further analysis was undertaken to examine the ANOVA and the coefficients. However, examining the coefficients table for course type, it can be seen that the percent participation in credit-bearing programmes decreases by

approximately a quarter point (9.5%) as opposed to participation in foundation, non-credit bearing programmes. The beta coefficient is significant at the 0.001 level.

4.2 Predicting Teaching Qualityfrom Instructor Gender and Course Type

For the second research question, I regressed teaching effectiveness as operationalized by the teaching mean against instructor gender, course type, class size, percent participation and statement 15 mean, the last variable of which refers to the single most important question relating to the students' assessment of their instructor (See Table 4).

Table 4. MLR model summary of teaching mean as outcome variable and instructor gender, course type, class size, percent participation and Statement 15 mean as predictors

Model	R	R^2	Adjusted R^2	Std. Error of the Estimate	R^2 Change	F Change	df1	df2	Sig. F Change
1	.045[a]	0.002	0.002	0.32	0.002	4.255	1	2059	0.039
2	.114[b]	0.013	0.012	0.32	0.011	22.876	1	2058	0.00
3	.114[c]	0.013	0.012	0.32	0.00	0.049	1	2057	0.825
4	.125[d]	0.016	0.014	0.31	0.003	5.307	1	2056	0.021
5	.917[e]	0.84	0.84	0.13	0.825	10610.76	1	2055	0.00

a. Predictor: (Constant), Instructor Gender

b. Predictors: (Constant), Instructor Gender, Course Type

c. Predictors: (Constant), Instructor Gender, Course Type, Class Size

d. Predictors: (Constant), Instructor Gender, Course Type, Class Size, Percent Participation

e. Predictors: (Constant), Instructor Gender, Course Type, Class Size, Percent Participation, Statement 15 Mean

It is clear that the first four variables did not account for more than 1.4% (0.014) of the variance in teaching mean, 1.1% of which is explained by course type alone. It is important to note that the addition of class size to Model 3 did not contribute any predictive power to explain the variance in teaching quality. Once again as in the case of percent participation, as a regressor course type appears to be a better predictor of teaching mean.

From Model 5, it is clear that instructor mean (Statement 15) explains by itself 82.5% of the variance in teaching mean. The correlation between Teaching Mean and Statement 15 mean is at R = 0.92, which is a highly significant positive correlation between the two. The latter finding is significant in the sense that teaching mean, which is measured in the CTS by 12 statements in the Course and Teaching Survey (CTS), can be predicted accurately by a single statement. The ANOVA table (Table 5) also shows that the model is significant at $p = 0.001$ level, $F(5, 2055) = 2162.3$, $p < 0.001$.

Table 5. ANOVA results of five variables predicting teaching mean

Model		Sum of Squares	df	Mean Square	F	Sig.
5	Regression	174.06	5	34.81	2162.3	.000[a]
	Residual	33.08	2055	0.016		
	Total	207.14	2060			

a. Predictors: (Constant), Instructor Gender, Course Type, Class Size, Percent Participation, Statement 15 Mean

Examining the coefficients table, it can be seen that if statement 15 mean goes by one point, it follows that teaching mean goes by 0.75 point. Again, the 95% confidence interval for Beta shows that this effect size (B = 0.91) is statistically significant as the lower and the upper bounds do not cross zero.

Table 4. Coefficients, t-tests and confidence intervals of the five models predicting teaching quality

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	3.27	0.009		359.29	0.00	3.25	3.28
	Instructor Gender	0.03	0.014	0.05	2.06	0.039	0.00	0.06
2	(Constant)	3.24	0.011		305.36	0.00	3.22	3.26
	Instructor Gender	0.03	0.014	0.04	1.96	0.05	0.00	0.06
	Course Type	0.07	0.014	0.11	4.78	0.00	0.04	0.10
3	(Constant)	3.25	0.042		78.24	0.00	3.17	3.33
	Instructor Gender	0.03	0.014	0.04	1.97	0.049	0.00	0.06
	Course Type	0.07	0.015	0.10	4.62	0.00	0.04	0.10
	Class Size	0.00	0.002	-0.01	-0.22	0.825	-0.01	0.00
4	(Constant)	3.19	0.049		65.49	0.00	3.09	3.28
	Instructor Gender	0.03	0.014	0.04	1.84	0.066	0.00	0.05
	Course Type	0.08	0.015	0.12	5.04	0.00	0.05	0.11
	Class Size	0.00	0.002	0.00	-0.03	0.976	0.00	0.00
	Percent Participation	0.00	0.00	0.05	2.30	0.021	0.00	0.00
5	(Constant)	0.78	0.031		25.49	0.00	0.72	0.84
	Instructor Gender	0.00	0.006	0.00	0.50	0.615	-0.01	0.01
	Course Type	0.04	0.006	0.06	6.06	0.00	0.03	0.05
	Class Size	0.00	0.001	-0.02	-1.90	0.058	0.00	0.00
	Percent Participation	0.00	0.00	-0.03	-3.20	0.001	0.00	0.00
	State. 15 Mean	0.75	0.007	0.91	103.01	0.00	0.73	0.76

Examining the coefficients for model 5 to explain teaching effectiveness, course type with an effect size of 0.06 is the only significant predictor. Statistical significance is achieved at the $p = 0.001$ level. It was earlier stated that percent participation in non-foundation credit-bearing programmes decreased by 9.5% as opposed to participation in foundation, non-credit bearing programmes. Therefore, in order to compare the effect size of course type on percent participation and teaching effectiveness, the standardized Beta value of course type on teaching effectiveness (0.11) was compared to the Beta value of its effect on percent participation (-0.23). The result indicates that the effect size of course type on percent participation is at least two times greater than it is on teaching effectiveness, though this size effect is on the negative side.

5. Discussion and Conclusions

Using a Multiple Linear Regression (MLR) research design, the primary goal of this research was to assess whether or not any relationship existed in SET completion rates and teaching effectiveness as a function of instructor gender, class size and course type. The general conclusion of the statistical analysis showed that there were extremely small effect sizes of instructor gender and course type on both percent participation and teaching effectiveness. However, these effects were still significant due to the large sample size (over 2000 instructor

ratings). This is in agreement with the extant literature where significant results with small effect sizes were reported (Smith et al., 2007).

The first research question addressed the variables which could predict percent participation (the percentage of students completing the SET survey). The results indicated that instructor gender explained 0.2% of the variance in percent participation. Course type (i.e., whether a certain course was for credit or for non-credit) explained 5.1% of the variance in percent participation with the result that participation rates in non-foundation credit-bearing programmes are 9.5% lower than they are in foundation, non-credit bearing programmes. This finding contradicts Adams and Umbach (2012) who hypothesized non-credit and pass/fail courses to yield low response rates because of students' perceptions that they are less important and less deserving of their attention. Further, the study by Macfadyen et al. (2015) found that students enrolled in lecture based courses were more likely to complete SET surveys than any other course type (e.g., individual study, experiential course, small group course). An important implication here is that based on the nature and the context of the investigated programmes, researchers need to be cognizant of the precise definition of variables, as the case for course type, has shown here (see also the example in the study by Wennerstrom and Heiser (1992) below).

The second research question investigated the association between several student-, instructor- and course-related factors and teaching effectiveness. The results indicated that instructor gender, course type, class size, and percent participation accounted for 1.4% of the variance in teaching mean, with the greatest variance (1.1%) explained by course type alone. In the literature, Smith et al. (2007) found that female instructors received higher ratings from both male and female students. For this study, male instructors received slightly higher ratings than female instructors, but as was indicated the effect size is extremely small. A comparable examination of the effect size of course type on percent participation and teaching effectiveness revealed that course type had a greater explanatory power of percent participation (-0.23) than it had oft eaching effectiveness (0.11). Wennerstrom and Heiser (1992) found that course type introduced negative bias to instructor ratings in the two ESL programmes they investigated. For example, reading and writing were rated lower than grammar in the intensive programme, whereas in the required academic programme, listening and speaking were rated lower than reading and writing. However, we must exercise caution in the interpretation of the findings of the current study, since as will be explored below, the teaching mean measure is an aggregate score of both instructor- and course-related factors and the determination of bias is not a straightforward process.

When instructor mean (Statement 15) was entered into the teaching effectiveness regression equation, it was found that holding other variables constant it explained by itself 82.5% of the variance in teaching mean with the result that an increase of one score in instructor mean (a 4-point scale) led to a 0.75 point corresponding increase in teaching mean. This finding is significant in the sense that teaching mean, which is measured by 12 statements in the Course and Teaching Survey (CTS), can be predicted accurately by a single statement.

This study has more theoretical as opposed to practical significance. It is one of the very few studies that investigated SET, and particularly bias. Apart from the study by Wennerstrom and Heiser (1992) which focused on instructional evaluation in ESL programmes and found evidence for bias in students' ratings of their instructors, no other study was conducted to investigate the validity and reliability of SET. The few studies that existed in the field of ELT either reviewed SET as an instructional evaluation tool (Pennington & Young, 1989) or explored teachers' perspectives of the SET process (Burden, 2008). Because of this rarity of research and because of the demonstrated effect sizes on participation levels and teaching effectiveness, it is of the essence that studies of this kind which examine the validity of SET be carried out in ELT in general and in English for Academic Purposes programmes in particular.

There are two limitations to this study. Firstly, in the absence of any research on the Course and Teaching Survey, the validity of the CTS as a measure of teaching effectiveness could not be ascertained. An alternative research design could have examined CTS for multi-dimensionality through the employment of factor or cluster analysis. In its place, the study relied on the existing bi-dimensionality of the CTS in terms of teaching effectiveness and the good instructor statement. This leads to the second limitation whereby the study relied on two aggregate scores/subgroupings (the teaching mean and the good instructor statement mean). Though looking at the CTS one can still identify two subgroupings, these have different number representation. In other words, instructional methodology is assessed by eight statements and course is assessed by three statements. This differing sub-grouping could produce different ratings. The problem with using the given sub-groupings for this study is that scores pertaining to course and those pertaining to instructor are amalgamated into one aggregate teaching mean score. The separation of these scores into teaching methodology and course is especially important for these EFL programmes, where the curriculum is controlled centrally (the results of which is that one expects little variation here and thus homogenous evaluation by the students) and where potential variation may transpire

in the teaching methodology of the instructors.

In conclusion, the question of bias is not easy to answer. The conclusion that students in for-credit non-foundation programmes participate less in the completion of SET cannot be conclusively said to introduce bias to student ratings. This could well be explained by the course structure where a single instructor teaches a section in the for-credit programmes, whilst two to three instructors team teach the non-for-credit programmes. Equally, the conclusion that students enrolled in for-credit non-foundation programmes award higher scores to the instructors teaching in those programmes may not indicate a bias in and of itself. It could be the case that the different ratings reflect actual differences in course content, and greater student and teacher autonomy in the for-credit programmes. Also, the high means in both measures of teaching quality and instructor and thus the slight positive skewness in favour of the instructors could either be interpreted to mean that bias is at play or that the instructors teaching on those programmes are on average more experienced instructors. As suggested by Feldman (1992), precise determination of bias can be done through replication of the earlier lab and experimental studies where teaching behaviors and confounding variables can be easily manipulated, controlled for and measured.

References

Adams, M. J. D., & Umbach, P. D. (2012). Non-response and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments. *Research in Higher Education*, *53*(5), 576-591. http://dx.doi.org/10.1007/s11162-011-9240-5

Arnold, I. J. M. (2009). Do examinations influence student evaluations? *International Journal of Educational Research*, *48*, 215-224. http://dx.doi.org/10.1016/j.ijer.2009.10.001

Avery, R. J., Bryant, W. K., Mathios, A., Kang, H., & Bell, D. (2006). Electronic SETs: Does an online delivery system influence student evaluations? *Journal of Economic Education*, *37*, 21-37. http://dx.doi.org/10.3200/JECE.37.1.21-37

Barth, M. M. (2008). Deciphering student evaluations of teaching: A factor analysis approach. *Journal of Education for Business*, *84*, 40-46.

Bennett, L., & Nair, C. (2010). A recipe for effective participation rates for web-based surveys. *Assessment & Evaluation in Higher Education*, *35*(4), 357-365. http://dx.doi.org/10.1080/02602930802687752

Bothell, T. W., & Henderson, T. (2003). Do online ratings of instruction make sense? *New Directions for Teaching and Learning*, *96*, 69-79. http://dx.doi.org/10.1002/tl.124

Burden, P. (2008). ELT teacher views on the appropriateness for teacher development of end of semester student evaluation of teaching in a Japanese context. *System*, *36*, 478-491. http://dx.doi.org/10.1016/j.system.2007.11.008

Dommeyer, C. J., Baum, P., Hanna, R. W., & Chapman, K. S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: Their effects on response rates and evaluations. *Assessment & Evaluation in Higher Education*, *29*, 611-623. http://dx.doi.org/10.1080/02602930410001689171

Feldman, K. A. (1992). College-students views of male and female college-teachers (Part I). Evidence from the social laboratory and experiments. *Research in Higher Education*, *33*, 317-375. http://dx.doi.org/10.1007/BF00992265

Gamliel, E., & Davidovitz, L. (2005). Online versus traditional teaching evaluations: Mode can matter. *Assessment & Evaluation in Higher Education*, *30*, 581-592. http://dx.doi.org/10.1080/02602930500260647

Gray, M., & Bergmann, B. R. (2003). Student teaching evaluations: Inaccurate, demeaning, misused. *Academe*, *89*, 44-46. http://dx.doi.org/10.2307/40253388

Greimel-Fuhrmann, B., & Geyer, A. (2003). Students' evaluation of teachers and instructional quality—Analysis of relevant factors based on empirical research. *Assessment & Evaluation in Higher Education*, *28*, 229-238. http://dx.doi.org/10.1080/0260293032000059595

Johnson, T. D. (2003). Online student ratings: Will students respond? In D. L. Sorenson, & T. D. Johnson (Eds.), *Online student ratings of instruction: New directions for teaching and learning* (Vol. 96, pp. 49-59). Hoboken, NJ: Wiley.

Macfadyen, L. P., Dawson, S., Prest, S., & Gašević, D. (2015). Whose feedback? A multilevel analysis of student completion of end-of-term teaching evaluations. *Assessment & Evaluation in Higher Education*, 1-19.

http://dx.doi.org/10.1080/02602938.2015.1044421

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry, & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319-383). New York: Springer. http://dx.doi.org/10.1007/1-4020-5742-3_9

Mason, P., Steagall, J., & Fabritius, M. (1995). Student evaluations of faculty: A new procedure for using aggregate measures of performance. *Economics of Education Review*, *12*(4), 403-416. http://dx.doi.org/10.1016/0272-7757(95)00016-D

Menges, R. J. (2000). Shortcomings of research on evaluating and improving teaching in higher education. *New Directions for Teaching and Learning*, *83*, 50-11. http://dx.doi.org/10.1002/tl.8301

Nair, C. S., & Adams, P. (2009). Survey Platform: A factor influencing online survey delivery and response rate. *Quality in Higher Education*, *15*(3), 291-296. http://dx.doi.org/10.1080/13538320903399091

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment & Evaluation in Higher Education*, *33*, 301-314. http://dx.doi.org/10.1080/02602930701293231

Otto, J., Sanford, D. A., & Ross, D. N. (2008). Does RateMyProfessor.com really rate my professor? *Assessment & Evaluation in Higher Education*, *33*, 355-368. http://dx.doi.org/10.1080/02602930701293405

Pennington, M. C., & Young, A. L. (1989). Approaches to Faculty Evaluation for ESL. *TESOL Quarterly*, *23*(4), 619-646. http://dx.doi.org/10.2307/3587535

Porter, S. R., & Umbach, P. D. (2006). Student survey response rates across institutions: Why do they vary? *Research in Higher Education*, *47*(2), 229-247. http://dx.doi.org/10.1007/s11162-005-8887-1

Smith, S. W., Yoo, J. H., Farr, A. F., Salmon, C. T., & Miller, V. D. (2007). The Influence of Student Sex and Instructor Sex on Student Ratings of Instructors: Results from a College of Communication. *Women's Studies in Communication*, *30*(1), 64-77. http://dx.doi.org/10.1080/07491409.2007.10162505

Spooren, P., Mortelmans, D., & Denekens, J. (2007). Student evaluation of teaching quality in higher education: Development of an instrument based on 10 Likert-scales. *Assessment & Evaluation in Higher Education Journal*, *32*(6), 667-679. http://dx.doi.org/10.1080/02602930601117191

Sproule, R. (2000). Student evaluation of teaching: A methodological critique of conventional practices. *Education Policy Analysis Archives*, *8*(50), 1-23.

Tabachink, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson Education.

Wennerstrom, A. K., & Heiser, P. (1992). ESL bias in instructional evaluation. *TESOL Quarterly*, *26*(2), 271-288. http://dx.doi.org/10.2307/3587006

Note

Note 1. Usually excluded from these statements and institutional analysis for these EFL programmes is statement # 13 which is about the value of the laboratory sessions.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).