Student Evaluation of Teaching in Higher Education: Evidence from Hong Kong

Jang C. Jin[1]

[1]Department of Economics, George Mason University Korea, Songdo, Incheon, Korea

Correspondence: Jang C. Jin, Department of Economics, George Mason University Korea, Songdo, Incheon, Korea. E-mail: jjin7@gmu.edu

Received: July 23, 2019 Accepted: August 20, 2019 Online Published: August 21, 2019

doi:10.5430/ijhe.v8n5p95 URL: https://doi.org/10.5430/ijhe.v8n5p95

Abstract

This paper examines empirically the determinants of student evaluation of teaching (SET). Empirical models were specified and estimated using the SET data collected in Hong Kong over six academic years. A key finding is that three different origins of students had a differentiated impact on teaching evaluation. In particular, students from mainland China appreciated and rated teaching favorably, and hence the more mainland talents in the class, the higher the class-average SET scores. However, local Hong Kong students valued teaching and learning effectiveness unfavorably. Exchange students from abroad also dropped the class-average SET scores, as well as class-average student performance. The results suggest that raw SET scores should be used with care if classes are unbalanced with a large group of atypical students who work less but blame instructors for everything.

Keywords: student evaluation of teaching (SET), student performance, origin of students, regression analysis, reverse causality

JEL codes: A2, I21

1. Introduction

The student evaluation of teaching (SET) has long been used for tenure review and promotion (e.g. Becker and Watts, 1999; Becker, Bosshardt, and Watts, 2012; Zhang and Liu, 2018). Although SET is widely used in higher education, some problematic issues have remained unresolved. More specifically, SET is largely determined by students' performance on exams, and hence junior members such as instructors and tenure-track professors are more lenient in their marking to ensure higher SET scores from students; once professors gain tenured positions, they normally cease such a 'grade inflation' (Krautmann and Sander, 1999; Johnson, 2003; Weinberg, Hashimoto & Fleisher, 2009, among others). Another drawback of SET is that students' rating of a teacher is inversely related with his/her enthusiastic efforts: the more seriously he/she teaches, the less favorably being rated (Carrell and West, 2010; Braga, Paccagnella, and Pellizzari, 2014). These two shortcomings reflect our general perception that the feedback from students is potentially 'demoralized' to academia without improving teaching and learning effectiveness (Marsh, 2007).

Furthermore, SET can also be biased against teaching and learning effectiveness, since class-average SET scores may vary depending on the origin of students. For example, if a group of peculiar students are hostile to a certain characteristic of a 'fine' scholar, (Note 1) they may dislike her/his way of teaching and thus the class average of SET scores would be lowered regardless of how the class has been taught. In this case, learning effectiveness may not play a role. The fall of SET scores would be proportional to the number of antagonistic students in the class. Conversely, if more promising students are a majority in the class, quality teaching is expected to be greatly appreciated. Serious students will most likely rate the instructor's enthusiastic efforts favorably.

The two contrasting effects can be identified if distinct origins of students coexist on the same campus. One suitable case is a higher education in Hong Kong where three idiosyncratic origins of students are available: local Hong Kong students, students from Mainland China, and exchange students from overseas. Although such different origins of students play an important role in teaching evaluation, no studies until today have focused on this particular relationship between the origin of students and teaching evaluation. (Note 2)

We thus classify students by their origin and use a textbook-style estimation method to examine their potential role in teaching evaluation. First, poor performers on midterm and final exams are found to lower the class average of SET scores substantially, and thus the bottom 10% is disregarded for further estimation. The adjusted SET scores that delete the bottom 10% appear to be positively and significantly affected by student performance in the exam. Second, the origin of students also matters. For example, students from Mainland China appreciate teaching and improve the class average of SET scores. In contrast, local Hong Kong students lower the class-average SET scores: the more Hong Kong students in the class, the lower the class-average SET scores. The worst case is the role of exchange students who have reduced class performance in the exam, as well as the class-average SET scores.

Section 2 discusses student characteristics in Hong Kong. Section 3 describes SET data collected in Hong Kong and other relevant data for model variables. Section 4 specifies an empirical model and explains the variables used. Section 5 discusses basic results. Section 6 examines a potential endogeneity problem of the model. Section 7 further investigates the impact on learning effectiveness of three origins of students. Section 8 concludes with some policy implications.

2. Student Characteristics in Hong Kong

There are largely three groups of students in Hong Kong: one is the group of students from Mainland China (11.9%); another group is local Hong Kong students (84.5%), a majority on campus; and the third group is the exchange students (3.6%) from all over the world (University Grants Committee of Hong Kong SAR, 2014/2015).

2.1 Mainland Chinese Students

Many students from Mainland China receive a full scholarship from the Hong Kong government or universities. The students on scholarships were selected competitively among high school graduates in China who had been admitted to top-tier universities in China, such as Peking or Tsinghua Universities. Most of the granted students ranked very high in the National Higher Education Entrance Exam (NHEEE) in China. (Note 3)

First of all, the study attitude of mainland Chinese students is highly regarded. When they first arrive in Hong Kong as freshmen, their spoken English is, on average, not as good as local Hong Kong students, so that mainland students are expected to have some difficulties in the class. But one important aspect is that they know how to study. For example, mainland students skim through textbooks or class materials before they attend classes. This reading in advance fills the gap, which may arise due to the lack of English comprehension in classes. Amazingly, they also compete with each other to sit in front rows because they do not want to miss even a small discussion in the class. In addition, unlike other Asian students, mainland Chinese students do not hesitate to participate in classroom discussions. Their questions often make class discussions enriched. After classes, they again read their textbooks and relevant materials more carefully and repeatedly until they fully digest the concepts.

Although mainland Chinese students' first midterm scores are, on average, a bit lower than local Hong Kong students, their performances on the second midterm and final exams are even better than average Hong Kong students. After all, most A's and B's are taken by mainlanders. However, some other mainland students, who are admitted to Hong Kong universities without scholarships, perform poorly. About half of them are in this category.

2.2 Local Hong Kong Students

Compared to mainland talents, Hong Kong local students are, in general, good at spoken English, as well as written English, since it is common to use English as a medium for instruction at the senior secondary level. (Note 4) However, they seldom participate in classroom discussions at the university level unless a certain type of incentive is given. This may be attributed to its culture in Hong Kong.

Another difference would be that, unlike mainland students, most Hong Kong students who are admitted to economics or business schools already learned economics in high school. (Note 5) While good performers in the A-level economics at the time normally understand the university-level Principles of Economics much better, some other students who got lower grades in the A-level economics show difficulties at the university level, perhaps due to a misunderstanding of basic concepts in high school economics.

Generally speaking, Hong Kong students seem to be less serious than mainland students at the university-level economics because they already learned basic economics in high school. Some other Hong Kong students just sit in the class to check whether or not university professors correctly teach the same topic of economics as in high school. If the contents and materials slightly differ from high school economics, they easily get lost. Especially, poor performers in high school economics are more confused with the ways of teaching at the university level. These students normally get C's and D's at the university level. (Note 6)

2.3 Exchange Students from Abroad

Exchange students are divided into two groups: Exchange students from the U.S. and European countries, and the other group from neighboring Asian countries. In general, exchange students do not seem to be very serious about studying while in Hong Kong. It is largely ascribed to their mindset. For example, many exchange students are interested in exploring nearby cities in Mainland China and other East Asian countries as much as possible even in the middle of the semester, often leading to an absence in classes. Although fall or spring season during a semester is a perfect time to make a trip since it is an off-peak season and hence it takes lower explicit costs of traveling, many exchange students overlook one thing: the opportunity costs of skipping classes would be even higher than what they thought.

Another misjudgment by exchange students, especially from western countries, is their low expectation of learning effectiveness while in Hong Kong. Such misunderstanding might be attributed to relatively mediocre facilities of Hong Kong universities, such as old-fashioned libraries, relatively small-scale bookstores, Chinese-style student unions, non-modernized noisy cafeterias, and so forth. In addition, Hong Kong's educational history is shorter. The oldest one is the University of Hong Kong (HKU, founded in 1911); relatively new schools are Chinese University of Hong Kong (CUHK, founded in 1963) and Hong Kong University of Science and Technology (HKUST, founded in 1991); and five smaller universities in Hong Kong.

Furthermore, that the economics courses are taught in English by non-native English speakers may not be much appreciated by exchange students. (Note 7) In fact, almost all faculties in Hong Kong universities are educated and trained at graduate schools of the United States and European countries, and they are generally active in both teaching and research in English. Jin and Yau (1999) and Jin and Hong (2008) found that their research productivity of top-three universities is competitive with major state universities in the U.S. In addition, non-native faculties, in general, tend to teach a bit higher level even in the principles courses, and they often grade more rigorously. Their grading rigor might be another important component in determining the overall rating of teaching negatively. Bosshardt and Watts (2001) and Weinberg et al. (2009) observed similar cases in the U.S., especially for economics faculties whose first language was not English.

Finally, exchange students often disturb a positive learning environment in Hong Kong. For example, some exchange students ask for favors such as extra work for better grades. Other students also make up stories to obtain any sort of advantages or sympathies. They often show unreasonable excuses to take make-up exams, and they even attempt to threaten faculties for better grades. Such dishonesty should be severely punished, using strict rules and regulations from the beginning of the semester.

3. The SET Data

This paper uses the author's own SET scores over the period of 2008-2013. Over the sample period of six consecutive years, the author taught 20 sections of Principles of Microeconomics in English at Business School, Chinese University of Hong Kong (CUHK), Hong Kong SAR in China. (Note 8) One important merit of using one faculty's SET scores over time is that many variations among different faculties are controlled in nature. Watts and Bosshardt (1991), for example, found that 'instructor effects' in cross-sectional data were significantly different in student learning. These instructor effects can be avoided by using one faculty's SET scores. For example, teaching quality and teaching materials would be almost constant over time unless new economic theories were developed over the sample period used. Only changes were students.

On average, the class size was 46.5 students, and the total number of students who took the author's principles classes in English were 929 students over the sample period (see Table 1). To find the percentage of 3 origins of students in each class, student names were used to identify their origin. For example, the family name 'Chen' is a mandarin pronunciation, which means that the student is obviously from Mainland China or Taiwan; the same family name with a Cantonese pronunciation 'Chan' means that he/she must be a local Hong Kong student; and exchange students from overseas are recognized by their foreign names. Mainland students were the majority (58.3%) in the class since their Cantonese, the Hong Kong local language, was not good enough to take other classes in Cantonese; local Hong Kong students made up of about one third (30.9%); and the rest of them were exchange students from abroad (10.8%).

Table 1. Summary Statistics

	SET	SET_adj	Test1(40)	Exam(100)	Exam_std	Grade	Mainland(%)	HK(%)	Exch(%)	CDFs(%)	Class Size
Mean	4.93	5.38	30.72	77.41	77.41	2.87	58.33	30.92	10.76	25.67	46.45
Standard Error	0.12	0.13	0.49	0.74	0.70	0.06	4.75	3.24	3.26	2.94	3.02
Median	5.06	5.56	31.06	78.19	77.11	2.88	59.25	28.95	1.90	21.35	47.50
Standard Deviation	0.56	0.59	2.21	3.30	3.11	0.26	21.26	14.48	14.58	13.15	13.51
Sample Variance	0.31	0.35	4.88	10.89	9.67	0.07	452.11	209.63	212.53	173.03	182.58
Kurtosis	3.84	3.39	2.23	-0.23	-0.30	-0.01	-0.32	0.06	0.32	0.08	1.66
Skewness	-1.75	-1.60	-1.27	-0.82	-0.12	-0.73	-0.56	0.49	1.21	0.89	-1.16
Minimum	3.19	3.57	24.50	70.87	71.27	2.29	9.10	5.70	0.00	8.70	11.00
Maximum	5.47	6.08	33.94	81.42	82.98	3.24	86.80	61.10	44.40	54.50	66.00
Sum											929.00

Note: SET stands for the student evaluation of teaching that uses a 6-point scale. SET_adj is an adjusted SET score after deleting the bottom 10%. Test 1 is the first midterm that has 40 multiple choices. Exam includes 3 tests (75% in total) and 5 quizzes (25%). Exam_std is a standardized Exam score to reduce the difficulty levels of exams over time. Grade is a semester letter grade that is converted to a 4-point scale. Mainland, HK, and Exch are, respectively, the proportions of students from Mainland China, Hong Kong local students, and exchange students from abroad. CDFs are the proportion of students who got Cs, Ds, and Fs.

Teaching evaluation was rigorously conducted in the classroom 1-2 weeks prior to final exams, so that the average response rate was kept relatively high (about 90 percent). To avoid any possibility of biased results, instructors were not present during the surveys. The questionnaires included eighteen questions, in which the last question inquired about "satisfaction with the teacher". A six-rating scale was used: 1 for 'strongly disagree' and 6 for 'strongly agree'. In fact, the author often experienced high variations on an individual basis. Some students gave a perfect score 6 while some others rated 1 in the evaluation.

To reduce the variation of such individual SET scores, this study uses the class average of SET scores for each course that fluctuates less than individual SET scores, and hence more robust estimation can be obtained (e.g. Weinberg et al., 2009). Table 1 shows that the author's average SET score over time was 4.93 (out of the 6-point scale); the highest one being 5.47 and the lowest being 3.19 (see Appendix A for historical data). (Note 9) However, it should be noted that the SET scores below 3.5 would be a rare case unless something unpleasant happened in the class. For example, if instructors lose temper in the class, teaching evaluation dramatically falls. Ill-prepared lectures also lead to relatively lower evaluations from students. Strict classroom rules may have a negative impact on student evaluations as well.

For the author's particular case, one exchange student disturbed the class repeatedly and hence a 'yellow card' was given. (Note 10) After that, several other exchange students did not pay attention and started disturbing the class on purpose. In the end, four more students got the yellow card. Strict classroom rules might restrict students' freedom to abuse academic integrity. Consequently, such disciplinary measures were found to be negatively related to SET scores. This reveals another dark side of SET in which learning effectiveness does not play a role.

In addition, adjusted SET scores that delete the bottom 10% are used here, since a certain number of less disciplined students always subsist. Typically, these students poorly perform in the exam but blame everything on instructors. As a means to an end, 'bombs' burst out in teaching evaluations. If this happens, SET scores are generally skewed downward. To correct the skewedness, the Business School of CUHK designed to delete the bottom 10% for departmental reviews. In general, the raw SET scores are corrected after the deletion of such outliers. But if the poorly performing group is unexpectedly large, deleting the bottom 10% does not suffice to correct the skewed SET scores. In our sample, SET scores were corrected upward by 0.45 points on average (see Table 1).

Table 1 also shows several measures of student performance on exams. The average exam score over time was 77.4%, with the average semester grade being B- (2.87 in the 4.0 scale). (Note 11) It is also interesting to find that the average performance in the first midterm was 76.8% (30.72 corrected answers out of 40 multiple choice questions), which was very close to the average exam score in total. This suggests that the first midterm most likely determines the overall semester grades. Its standard deviation further indicates that the difficulty levels were slightly different over time. (Note 12) To align the varied difficulty levels, each semester's exam score was standardized as:

$$Exam\_std_{ij} = (X_{ij} / \dot{X}_i) * \ddot{X}, \quad (1)$$

where X_{ij} = the average exam score in semester i and class j, $\dot{X}_i$ = the overall average in semester i, and $\ddot{X}$ = the overall average for the entire sample period 2008-2013. In other words, the class average was first normalized by the semester average, and then standardized again by the overall average. Therefore, difficulty levels were normalized over the entire sample period.

Table 1 further shows that poor performers who got Cs, Ds, and Fs (CDFs) made up of 25.7% on average, and the proportion varied from 8.7% to 54.5% over time. When the proportion was relatively low, like 10%-20%, the SET scores were generally higher than 5. However, once the proportion of such poor performers rose to 30-50%, SET scores fell below 5. This negative correlation (r = -0.43) suggests that poor performers evaluated teaching unfavorably and hence the class-average SET scores declined further. If this happens, faculty members in general may react differently, depending on tenured or tenure-track faculties. For example, young, lenient, and fashionable tenure-track faculties, as well as instructors, will reduce the proportions of Cs, Ds, and Fs; whereas tenured faculties will remain strict in teaching and grading. (Note 13) However, the lenient grading makes it easier and warranted for most faculties to get rated better by students. Accordingly, the 'inflation' of GPA is difficult to avoid in higher education (e.g. Johnson, 2003).

Table 2. Correlation Coefficients

	SET	SET_adj	Test1	Exam_raw	Exam_std	Grade	Mainland	HK	Exch	CDFs	Class Size
SET	1.00										
SET_adj	0.98	1.00									
Test1(40)	0.26	0.37	1.00								
Exam-raw	0.32	0.39	0.67	1.00							
Exam-std	0.30	0.37	0.50	0.93	1.00						
Letter Grade	0.33	0.39	0.59	0.98	0.95	1.00					
Mainland(%)	0.55	0.61	0.67	0.66	0.55	0.58	1.00				
HK(%)	-0.35	-0.38	-0.28	-0.23	-0.25	-0.22	-0.73	1.00			
Exch(%)	-0.45	-0.50	-0.70	-0.74	-0.55	-0.62	-0.73	0.07	1.00		
CDFs(%)	-0.43	-0.46	-0.64	-0.92	-0.86	-0.95	-0.62	0.32	0.58	1.00	
Class Size	0.52	0.62	0.62	0.66	0.62	0.62	0.54	-0.17	-0.61	-0.63	1

Note: See Table 1 for the definition of variables.

Table 2 reports correlation coefficients of model variables to detect potential multicollinearity problems. First, correlation coefficients among independent variables are found to be relatively small, except for a few relations. In particular, three measures of student performance (Exam_raw, Exam_std, Letter Grade) are highly correlated with CDFs so that these pairs cannot be included in the same regression model. Two out of three origins of students (Mainland and Exch) are also highly correlated with Test1. For Letter Grade, both Mainland and Exchange students are highly correlated with, but not with Hong Kong students. As long as highly correlated pairs are not included in the same model, multicollinearity problems may not be serious. Other symptoms of the multicollinearity problem—distorted signs of parameter estimates, high R^2 but fewer significant t-values, or instability of standard error estimates—will also be detected in the sections below.

4. Empirical Model

To estimate the determinants of teaching evaluation in Hong Kong, our empirical model includes three explanatory variables: student performance (SP), origin of students (SO), and a dummy variable (DUM). The regression model is thus specified as:

$$SET_i = \beta_0 + \beta_1 SP_i + \beta_2 SO_i + \beta_3 DUM_i + \varepsilon_i. \quad (2)$$

As noted earlier, class-average SET scores are used here to reduce the variation of individual student SET scores. SP includes three measures of student performance: (a) raw exam scores for each semester, (b) standardized exam scores in which difficulty levels are normalized, and (c) overall semester letter grades that are based on a criterion of relative grading. Since the letter grades are more or less normalized in nature, the two measures in (b) and (c) are expected to have similar effects on SET. SO also consists of three origins of students: (a) students from Mainland

China, (b) local Hong Kong students, and (c) exchange students from abroad. The three origins of students are expected to have distinctive effects on teaching evaluation. In addition, DUM stands for a western dummy variable ('1' for exchange students from western countries and '0' otherwise). This western dummy is expected to isolate the effect of potential outliers.

In this case, residuals ε_i are best interpreted as unobservable measures that affect teaching evaluation, and thus the regression model may include more control variables such as: how instructors give a favor to students in the class, the frequency of teacher-student interactions on an individual basis, the amount of time spent for occasions other than teaching (e.g. research and administrative works), government policies on higher education, and even a variable for physical attractiveness of good-looking instructors. (Note 14) Rather than including all these variables, we control the effect of student performance (SP) which is known to be a key determinant of SET. We also include a dummy variable (DUM) to insulate the effect of an outlier on SET. After controlling these two variables, we focus more on the important aspects of the origin of students (SO) that may be another key determinant of students' rating of teaching. This is our major contribution to the literature.

5. Basic Results

Table 3 shows the effects of student performance on teaching evaluation. In Model (1), raw SET scores are allowed to be explained by raw exam scores. After isolating the negative effect of an outlier, the effect of exam scores on SET is found to be positive but insignificant at the conventional significance levels. The insignificant effect appears to be at odds with our general belief that the student performance largely determines SET scores (e.g. Weinberg et al., 2009). Model (2) replaces the raw exam scores by the standardized exam scores in which the difficulty levels were normalized. Again, the parameter estimate is found to be small and insignificant at the 5% significance level. Model (3) further shows that the effect of letter grades is insignificant.

Table 3. Student Performance and SET Scores

Dependent variable: Raw SET Scores

Independent	*Model (1)*	*Model (2)*	*Model (3)*	*Model (4)*
Variables				
Constant	1.990	1.075	3.351	5.429
	(2.062)	(2.092)	(0.939)	(0.167)
$DUM_{western}$	-1.752**	-1.809**	-1.769**	-1.762**
	(0.392)	(0.375)	(0.380)	(0.343)
Exam_raw	0.039			
	(0.026)			
Exam_std		0.050		
		(0.026)		
Letter Grade			0.580	
			(0.325)	
CDFs (%)				-0.016**
				(0.005)
Adj R-square	0.54	0.57	0.56	0.64
Stand Error	0.37	0.36	0.36	0.33

Note: For variable definitions, see Table 1. Standard errors are in parenthesis. * indicates significant at the 5% level, and ** at the 1% level.

One possible reason would be that a group of poor performers in the class may distort average SET scores downward. Model (4) thus replaces the student performance measures by the percentage of poor performers who got Cs, Ds and Fs (CDFs). As expected, the negative effect of CDFs appears to be significant even at the 1% significance level. This suggests that raw SET scores are biased downward because of the group of poor performers in the class, and thus the raw SET scores are limited to use for further estimation.

Rather than using raw SET scores, Table 4 thus employs adjusted SET scores as a dependent variable, which delete the bottom 10%. Standardized exam scores and letter grades are allowed to explain the adjusted SET scores in models (1) and (2), respectively. As noted earlier, these two measures are normalized and thus any possibility of varied difficulty levels that may arise over time is reduced. Unlike in Table 3, the size of the effects appears to be larger and statistically significant at the 5% level. The results are, in general, consistent with the findings in Weinberg et al. (2009), among others, in which teaching evaluations across faculty members were determined mainly by current course grades.

Table 4. The Origin of Students and SET Scores

Dependent variable: Adjusted SET scores (bottom 10% excluded)

Independent	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)
Variables					
Constant	0.271	3.359	4.774	6.046	5.534
	(2.165)	(0.988)	(0.258)	(0.184)	(0.122)
$DUM_{western}$	-1.880**	-1.830**	-1.551**	-2.003**	-1.677**
	(0.388)	(0.400)	(0.389)	(0.348)	(0.525)
Exam_std	0.067*				
	(0.027)				
Letter Grade		0.736*			
		(0.342)			
Mainland (%)			0.011**		
			(0.004)		
Hong Kong (%)				-0.018**	
				(0.005)	
Exchange (%)					-0.006
					(0.008)
Adj R-square	0.59	0.57	0.63	0.67	0.47
Stand Error	0.37	0.38	0.36	0.33	0.43

Note: For variable definitions, see Table 1. Standard errors are in parenthesis. * indicates significant at the 5% level, and ** at the 1% level.

Model (3) provides further evidence in which mainland Chinese students have a positive and significant impact on adjusted SET scores. This suggests that high ability students from Mainland China did see the benefits of being pushed in the class, and thus they appreciated such a teaching favorably although being harsh. The result is similar to the case of 'honors students' in Weinberg et al. (2009), in which students in honors sections rated teaching more favorably. Model (4) substitutes Hong Kong local students for Mainland talents and finds that the average SET scores significantly fall when more Hong Kong students are in the class. This negative effect is, in general, consistent with the results in Braga, Paccagnella, and Pellizzari (2014) in which students in Italy also disliked the teaching that required students to work harder. Carrell and West (2010) found similar results for U.S. Air Force Academy cadets as well. Model (5) replaces HK by exchange students and finds that the effect of exchange students on SET scores is negative, although insignificant. This marginal negative effect, together with a significant negative impact of the western dummy ($DUM_{western}$), suggests that exchange students, in general, tend to evaluate teaching unfavorably.

Overall, we confirm that the origin of students is an important determinant of teaching evaluation. Students' ratings are affected favorably or unfavorably depending on the proportions of such students in the class. If more mainland students are in the class, teaching evaluation normally improves. In contrast, the evaluation results are generally worsened if Hong Kong students are the majority in the class. However, exchange students behave differently: students from western countries responded unfavorably towards teaching evaluation, but students from neighboring Asian countries tended to evaluate teaching favorably. (Note 15, 16)

6. Reverse Causality

The standard regression results were, however, based on the assumption that regressors were uncorrelated with the error term ε_i, and hence parameter estimates were presumed to be unbiased and consistent. But in practice, the possibility of biased and inconsistent parameter estimation due to endogenous explanatory variables is not rare. In particular, student performance can be affected by changes in study hours, family income, IQ, and so forth. Therefore, OLS estimates can be biased and inconsistent unless there is no association between student performance measures and residuals.

One source of the endogeneity bias is reverse causality, in which student performance may depend on the results of teaching evaluation. In other words, if students evaluate teaching favorably, students' semester grades are hinted to be upgraded. In this case, causality may run from SET scores to Letter Grade. However, this causal direction is not possible unless instructors unlawfully give such a hint right before teaching evaluation. Alternatively, instructors who expect to receive bad evaluations may penalize their classes with the use of a harder curve in grading or harder final exams. In either case, causality may run backward from teaching evaluation to semester grades.

Table 5. Reverse Causality

Dependent variable: Letter Grade (class averages)

Independent Variables	*Model (1)*	*Model (2)*
Constant	0.741(-0.069)	0.536(-0.73)
Test 1	0.069**(-0.022)	0.060* (-0.024)
SET_adj		0.086(-0.09)
Adj R-square	0.3	0.3
Stand Error	0.21	0.21

Note: For variable definitions, see Table 1. Standard errors are in parenthesis. * indicates significant at the 5% level, and ** at the 1% level.

We address this possibility using a structural model in Table 5 such that overall semester grades are predicted by teaching evaluation. Test1 is included in the model since the first midterm most likely determines overall semester grades. This is similar to the notion that children's performance at an early age largely determines their lifelong achievements and successes (Tickell, 2011). Model (1) shows that the positive effect of Test1 on Letter Grade is significant at the 1% significance level. Model (2) shows that the causal effect of adjusted SET scores on Letter Grade, however, appears to be insignificant even at the 10% level. The insignificant effect remains intact across model specifications. In other words, higher SET scores do not guarantee better grades to students. The results suggest that fine students work harder and perform better, leading to a more appreciation towards teaching, but not

the other way around. Therefore, the reverse causality from teaching evaluation to student performance appears to be unconvincing here.

7. Learning Effectiveness

Table 6 further shows that learning effectiveness may vary depending upon the origin of students. Models (1) - (3) include three origins of students one-by-one to figure out which group of students learn better. (Note 17) Model (1) shows that the overall semester grades (Letter Grade) are significantly improved by the number of mainland Chinese students in the class. In contrast, Model (2) suggests that the average student performance is not much influenced by the proportion of Hong Kong students in the class. Although they had a certain level of backgrounds in high school economics, Hong Kong students did not seem to work hard as much as mainland talents, and thus the effect on their learning effectiveness appeared to be smaller. Model (3) shows the worst case that exchange students significantly lowered average exam scores. Regardless of nationality, exchange students from abroad did not have as much prior knowledge in economics as did Hong Kong students, and thus their average performance is even lower than Hong Kong students.

Table 6. Learning Effectiveness by 3 Origins of Students

Dependent variable: Letter Grade (class averages)

Independent Variables	*Model (1)*	*Model (2)*	*Model (3)*
Constant	2.349(0.501)	2.088(0.637)	2.752(0.538)
SET_adj	0.023(0.108)	0.154(0.106)	0.042(0.095)
Mainland (%)	0.006*(0.003)		
Hong Kong (%)		0.001 (0.004)	
Exchange (%)			0.001* (0.004)
Adj R-square	0.26	0.05	0.32
Stand Error	0.22	0.25	0.21

Note: For variable definitions, see Table 1. Standard errors are in parenthesis. * indicates significant at the 5% level, and ** at the 1% level.

However, it should be noted that the negative effect of Hong Kong students on learning effectiveness is seemingly at odds with the findings of Gill and Gratton-Lavoie (2011), among others, in which the effects of high school economics were found to be positive for Californian students. But for Hong Kong students, the negative effect found here was due to a direct comparison with mainland talents who raised the class average substantially. Because of the enhanced class averages, Hong Kong students' performance was relatively underestimated and thus it had a negative and smaller impact on average performance.

Overall, mainland Chinese students performed well, followed by Hong Kong students, and exchange students were the bottom. That the Hong Kong students performed better than exchange students was largely ascribed to an early education of economics in Hong Kong high schools. The result is, in general, consistent with the findings in the literature in which the importance of economics education in high school should be emphasized for the improvement of learning effectiveness at the tertiary level (e.g., Walstad, 1992; Salemi and Siegfried, 1999; Walstad and Rebeck, 2000; Walstad, 2001; Gill and Gratton-Lavoi, 2011).

8. Concluding Remarks

In this paper, we empirically examined two important determinants of teaching evaluation. First, we found that the student evaluation of teaching (SET) was determined largely by student performance. In particular, both measures of standardized exam scores and semester letter grades showed positive and significant effects on teaching evaluation. Students who got better grades were most likely satisfied with teaching. Nonetheless, the reverse causal impact from teaching evaluation to student performance appeared to be insignificant.

Second, the origin of students was found to have a segregated impact on teaching evaluation. For example, a group of students who worked hard to get a good grade normally appreciated and thus evaluated teaching favorably. Students from Mainland China were in this category. Another group of students was the one unfavorable to a certain characteristic of an instructor (e.g. strict classroom rules). Such peculiar students were averse to the

instructor's way of teaching and thus the class average of SET scores were undervalued no matter how the class was taught. Hong Kong local students were in this category. The third group represents the students who performed poorly in the class and tended to blame everything on instructors. This group of students typically deteriorated average SET scores and also reduced the class average of student performance. Exchange students were in this category. Therefore, SET results would be biased if classes are unbalanced with a large group of atypical students who work less but blame instructors for everything.

One policy implication is that SET scores should be used with care if SET results are used as a good standard for tenure reviews, promotions, and salary increments. Especially for junior faculties, a good teaching is normally a critical requirement, along with quality research, for tenure reviews. In this case, a sole use of students' rating may mislead tenure decisions. For example, if a 'fine' scholar teaches a group of irresponsible students who work less but solely blame teachers, the student evaluation of teaching is foreseen to be underrated. In this case, 'peer evaluation' is suggested to be an alternative. (Note18) Since teaching is appraised by senior faculties who have more experiences in teaching, peer evaluation has a merit to avoid the biased ratings of a group of irresponsible students. In this sense, peer evaluation of teaching complements the student rating of teaching, and hence the use of peer evaluation together with SET would be a more appropriate assessment of teaching effectiveness.

Another policy implication is that if no other alternatives are available, the class average of SET scores can be utilized, but it should be 'deflated' by the class average of student performance. (Note 19) Since the SET results have been found to be highly correlated with student performance, tenure-track assistant professors and instructors tend to be lenient in grading to get higher SET scores. Students are also elastic to an instructor's grading policy. Although both parties do not (and must not) unlawfully collude in the class, students are clever enough to get information through the internet how much lenient their instructor is and how much SETs are valuable to the instructor's future promotion. Students therefore tend to take more junior instructors who are normally easygoing in grading. Therefore, the so-called grade inflation prevalent on campus will gradually mitigate if SET scores are deflated by the exact amount of grading inflated, compared to a departmental average.

Another drawback would be that students often abuse SET. Some students mistakenly think that SET is a good instrument to negotiate with instructors. Especially in the final two weeks of a SET-conducting period, students generally become more demanding. Chronic absentees make unreasonable excuses to take make-up exams, while poor performers ask for unlawful requests. They also attempt a conspiracy to burst out a 'bomb' if their demands are not met. To avoid these hassles, some instructor's syllabus these days is more than 10 pages long that specifically describes strict rules and regulations from the beginning of the semester. Such disciplinary measures, however, tend to reduce SET scores since strict classroom rules are mistakenly thought among students to restrict their freedom to abuse academic integrity. This reveals another dark side of SET in which teaching and learning effectiveness does not play a role.

Although such glitches, SET also has positive sides. For example, bona fide instructors train students honestly and enthusiastically, but some unsound instructors illegitimately skip classes without a reason and teach several topics hither and thither without a sequence and without a logic. In this case, students are victim, but little ways are available for school administrators to detect such flaws. Classroom is sort of monopolized by an instructor, so that classroom teaching cannot be monitored by a third party. It is also because of academic freedom warranted on campus. However, SET prevents instructors from fallacious teaching. Students monitor an instructor in the classroom and evaluate teaching at the end of every semester. Therefore, SET is widely used as one good criterion for tenure and promotion reviews especially for junior members, and hence the current form of SET is unavoidable in academics until a new, better method of evaluation is developed. Until then, students' rating of teaching will be appropriate to use if the current form of SET is used together with peer evaluation to reduce any decision fallacy; SET scores also need to be deflated for cross-sectional comparisons; and disciplinary measures are suggested to include to prevent from students' abuse of SET as well.

We conclude with discussing two limitations of our study. First, our measure of teaching effectiveness employed the author's own SET scores over six academic years in Hong Kong, potentially limiting a generalization of our findings. Particularly, the author was teaching in Hong Kong as an expatriate, and the classes were taught in English. The language of instruction might have been important in teaching and learning effectiveness. A cultural difference might also have affected the SET scores. However, as the author was a tenured professor, whose teaching evaluation results were generally above par, the results found here would not necessarily reflect abnormal cases in Hong Kong.

Second, our sample included only twenty observations, and thus the validity of estimation results might be doubtful. More specifically, the basic assumption of a normal distribution requires a large number of observations for proper estimation. However, the descriptive statistics showed that the data are approximately normally distributed and thus unbiased estimation could be achieved. Once more information becomes available for the same course, the important role of the origin of students in determining teaching evaluation can be more adequately assessed. Meanwhile, the robustness of the results found here may not be dramatically mitigated even with the inclusion of new, larger observations.

References

Becker, W. E., Bosshardt, W. & Watts, M. (2012). How departments of economic sevaluate teaching. *Journal of Economic Education, 43*(3), 325-333. https://doi.org/10.1080/00220485.2012.686826

Becker, W. E. & Watts, M. (1999). How departments of economics evaluate teaching. *American Economic Review, 89*(2), 344-349. https://doi.org/10.1257/aer.89.2.344

Bosshardt, W. & Watts, M. (2001). Comparing student and instructor evaluations ofteaching. *Journal of Economic Education, 32*(1), 3-17.https://doi.org/10.1080/00220480109595166

Braga, M., Paccagnella, M. & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review, 41*(1), 71-88.https://doi.org/10.1016/j.econedurev.2014.04.002

Carrell, S.E. & West, J.E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy, 118*(3), 409-432.https://doi.org/10.1086/653808

Gill, A. M. & Gratton-Lavoie, C. (2011). Retention of high school economics knowledgeand the effect of the California State mandate. *Journal of Economic Education, 42*(4), 319-337.https://doi.org/10.1080/00220485.2011.606083

Hong Kong Census & Statistics Department. (2011). *Thematic Household Survey ReportNo. 46 - Hong Kong Students Studying Outside Hong Kong*. Hong Kong SAR, PRC. Available at http://www.statistics.gov.hk/pub/B11302462011XXXXB0100.pdf.

Inoue Y. & Clark, N. (2013). Secondary education in Hong Kong. *World Education News& Review*. January 1st, 1-11.

Jin, J. C. & Hong, J. H. (2008). East Asian rankings of economics departments. *Journal of Asian Economics, 19*(1), 74-82.https://doi.org/10.1016/j.asieco.2007.12.009

Jin, J. C. & Yau, L. (1999). Research productivity of the economics profession in East Asia. *Economic Inquiry, 37*(4), 706-710.https://doi.org/10.1111/j.1465-7295.1999.tb01458.x

Johnson, V. E. (2003). *Grade Inflation: A Crisis in College Education*. New York, NY: Springer-Verlag.

Krautmann, A. C. & Sander, W. (1999). Grades and student evaluations of teachers. *Economics of Education Review, 18*(1), 59-63.https://doi.org/10.1016/S0272-7757(98)00004-1

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,* 319-383, Dordrecht, The Netherlands: Springer.

Salemi, M. K. & Siegfried, J. J. (1999). The state of economics education. *American Economic Review, 89*(2), 355-361.https://doi.org/10.1257/aer.89.2.355

Tickell, D. C. (2011). *The Early years: Foundations for life, health and learning – An Independent Report on the Early Years Foundation Stage to Her Majesty's Government*. 1-105.

University Grants Committee of Hong Kong SAR. (2014/2015). *General Statistics of UGC-funded Institutions.* Hong Kong SAR, PRC. Available athttp://cdcf.ugc.edu.hk/cdcf/searchUniv.do?lang=EN.

Walstad, W. B. (1992). Economics instruction in high schools. *Journal of Economic Literature, 30*(4), 2019-2051.

Walstad, W. B. (2001). Economic education in U.S. high schools. *Journal of Economic Perspectives, 15*(3), 195-210.https://doi.org/10.1257/jep.15.3.195

Walstad, W. B. & Rebeck, K. (2000). The status of economics in the high school curriculum. *Journal of Economic Education, 31*(1), 95-101.https://doi.org/10.2307/1183345

Watts, M. & Bosshardt, W. (1991). How instructors make a difference: Panel data estimates from principles of

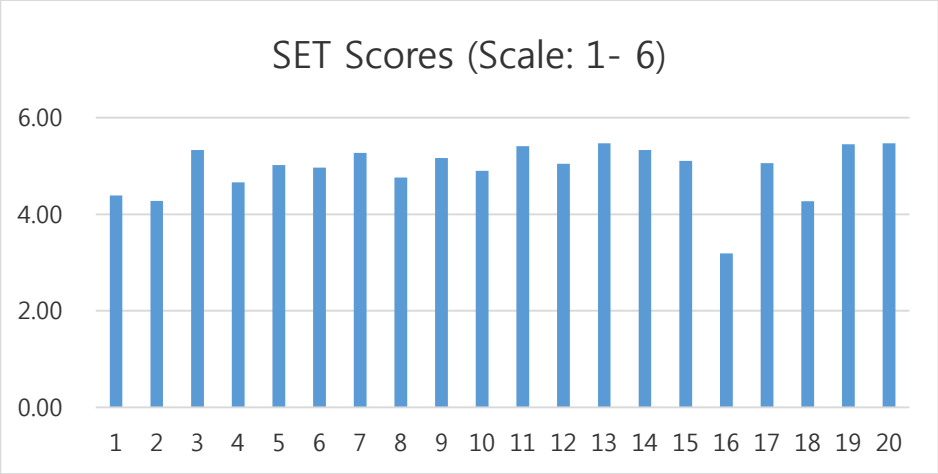
economics courses. *Review of Economics and Statistics, 73*(2), 336-340.

Watts, M. & Lynch, G. J. (1989). The principles courses revisited. *American Economic Review, 79*(2), 236-241.

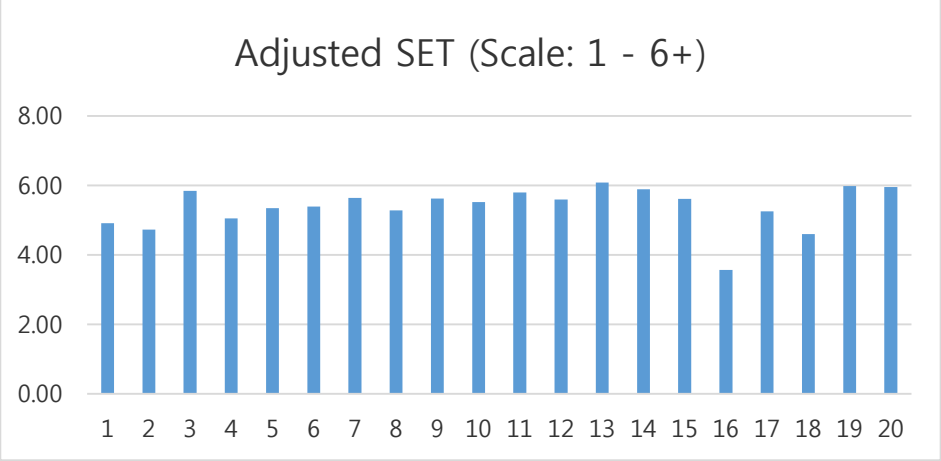
Weinberg, B.A., Hashimoto, M. & Fleisher, B.M. (2009). Evaluating teaching in higher education. *Journal of Economic Education, 40*(3), 227-261.https://doi.org/10.3200/JECE.40.3.227-261

Zhang, W. & Liu, B. (2018). The undergraduate teaching evaluation system in China: Progress, problems and suggestions. *Chinese Education & Society, 51*(4), 248-259.

Appendix A: Student Evaluation of Teaching (SET): Class-average Scores



Note: Values on the vertical axis are Raw SET scores.



Note: Values on the vertical axis are adjusted SET scores that delete the bottom 10%.

Notes

Note 1. Fine scholar is defined here as a bona fide faculty who has good research publications with her/his SET scores that are above average.

Note 2. Only Weinberg et al. (2009) briefly included 'honors sessions' as a special type of serious students and found a positive effect on teaching evaluations.

Note 3. Also known as National College Entrance Exam, it is similar to SAT in the U.S.A. A recent survey shows that mainland talents are almost indifferent in the choice of Peking, Tsinghua or Hong Kong Universities. The merit of Hong Kong universities would be quality faculties, as well as better job opportunities after graduation (Inoue and Clark, 2013)

Note 4. At the junior secondary level, schools have an option to adopt Chinese as the language of instruction for teaching, and up to 25% of classroom activities and research can be done in English (Inoue and Clark, 2013).

Note 5. Economics was an elective for Hong Kong Certificate of Education Examination (HKCEE) conducted at the end of Secondary 5. About one-third of HKCEE test takers continue to take Hong Kong Advanced Level Examination (HKALE) after two more years study in high school. However, the 7-year academic structure of secondary education was recently shortened to 6 years, and the two exams were replaced by one exam, the so-called Hong Kong Diploma of Secondary Education Examination (HKDSE), for the first time in the summer of 2012 (Inoue and Clark, 2013).

Note 6. About 25 percent of high school graduates are continuing their degree courses in Hong Kong; 6 percent go abroad for tertiary education; 7 percent enter the workforce; 10 percent are unemployed or unknown; and the rest of them pursue non-degree courses either locally or overseas (Hong Kong Census and Statistics Department, 2011).

Note 7. In fact, many economics professors in the United States are expatriates from all over the world. Some of them have a severe foreign accent.

Note 8. As a tenured professor, the author was an expatriate in Hong Kong.

Note 9. The original evaluation forms that include SET scores for all 20 classes are available upon request. The department average was around 4.6 points out of 6.0 over the sample period.

Note 10. Like a referee system in a soccer match, a yellow card is issued if students violate classroom rules such as no side talks, no mobile phones, and no other disturbances. One yellow card deducts 10% from overall semester grades, and two yellow cards are equivalent to a red card which means an F. This penalty system is specifically documented in a syllabus.

Note 11. There were 3 tests and 5 quizzes in total. The three tests were all multiple-choice questions, and five quizzes were one-page essay questions. Each test counted 25% of overall semester grades and hence 3 tests counted 75% in total, and five quizzes were another 25%.

Note 12. All exam questions were newly made each semester, so that the degree of technical hitches was difficult to avoid.

Note 13. Watts and Lynch (1989), for example, compared faculty instructors with teaching assistants (TA) and found that faculties were harder graders than TA, and their SET scores were not necessarily better than TA in principles courses.

Note 14. There are several other factors that could influence student responses to the survey. First, a teaching style of the instructor would be the one. The author followed the conventional way of teaching that uses a whiteboard most of the time. In this case, students had more time to write down and actually drew the graphs by themselves one by one, in which students understood better than just reading power-point slides. Second, the type of course taught would be another factor. In this case, the course taught was a lecture type rather than a group discussion or seminar type. This is also a common practice in economics. Third, Principles of Microeconomics was the first required economics course for business students, and hence all freshmen in business majors should have taken the course. Fourth, the author's course was taught in English rather than in Cantonese in Hong Kong, so that majority were the students who went to Hong Kong for the first time. Most of them were elite Mainland Chinese students who were selected by the Hong Kong government with full scholarships. These serious students knew how to appreciate the logical way of teaching although exams and the materials discussed in the class were more difficult than in other principles classes. This might be slightly different from the sample if SET data were selected from other principle

classes. However, all these factors are nearly constant over time since the SET data were obtained from one instructor for one particular course. The author thanks an anonymous reviewer who raised this issue.

Note 15. For example, a separate effect of exchange students from Korea only was positive and significant at the 5% level. The results are available upon request.

Note 16. It may have such an impression that the results found here cannot be generalized because the SET data were obtained from one faculty's teaching experiences. However, the faculty himself was a tenured professor as an expatriate in Hong Kong whose way of teaching was nearly constant over time and did not rely much on students' evaluation. It is thus not surprising to find that the basic results found here were consistent with the findings in the literature. Notice, however, that the SET data used here were taken from one particular course that was taught by one instructor over six years. This type of data, in fact, is unique and has a merit to avoid 'faculty variations' that have been potential problems in other SET studies in the literature. In other words, each faculty, in general, has different ways of teaching, and their difficulty levels in teaching also differs across subjects. However, the use of one faculty's historical data over six years will be rather a unique study in the literature since few faculties are willing to disclose their own SET scores to the public. The author thanks anonymous reviewers who raised this issue.

Note 17. Test 1 is not included here since it is highly correlated with mainland talents ($r = 0.67$), as well as with exchange students ($r = -0.70$). In other words, the more mainland Chinese students are in the class, the higher the class average of Test 1. For exchange students, the results are opposite.

Note 18. About half of economics departments in the U.S. conducted peer reviews for tenure and promotion decisions (Becker, Bosshardt, and Watts, 2012).

Note 19. This is similar to the measurement of real GDP for cross-country studies, in which nominal GDP is deflated by GDP deflator.