Economics of Education Review 49 (2015) 172–179

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Why do good performing students highly rate their instructors? Evidence from a natural experiment

Donghun Cho[a], Wonyoung Baek[b], Joonmo Cho[b,*]

[a] Department of Economics, Hallym University, 1 Hallimdeahak-gil, Chuncheon-si, Gangwon-do, 24252 Korea
[b] Department of Economics, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongno-gu, Seoul, 03063, Korea

ARTICLE INFO

Article history:
Received 22 August 2013
Revised 30 September 2015
Accepted 11 October 2015
Available online 19 October 2015

JEL classification:
I20
I21
I23

Keywords:
Teaching quality
Student evaluation
Student grade
Hedonic value of grade surplus
Psychological reward-penalty

ABSTRACT

This article analyzes the behavior of students in a college classroom with regard to their evaluation of teacher performance. As some students are randomly able to see their grades prior to the evaluation, the “natural” experiment provides a unique opportunity for testing the hypothesis as to whether there exists a possibility of a hedonic (implicit) exchange between the students’ grades and teaching evaluations. Students with good grades tend to highly rate the teaching quality of their instructors, in comparison with those who receive relatively poor grades. This study finds that students with better grades than their expected grades provide a psychological “gift” to their teachers by giving a higher teacher evaluation, whereas it is the opposite with those students receiving lower grades than their expectation. These empirical results demonstrate that a previous interpretation on the effect of student grades in an incumbent course with regard to the teaching quality may have to be somewhat discounted.

1. Introduction

As there has been an increased emphasis on the public accountability of universities, the role of the faculty in teaching and conducting research at a university is becoming more important. Teaching plays a major role in college education, and the student evaluation of teaching (hereinafter, SET) is a reference for improving the quality of instruction (Lee & Cho, 2014). In many universities, student evaluations are used as key materials for the academic promotion process as well as associating the number of course registrations with the students’ preference for faculty. As a result, not only the course content (e.g., clarity of instruction, adequacy of course materials and instruction methods), but also other factors such as an instruction itself (e.g., competency and enthusiasm of the faculty and grades assessed by student evaluations) are becoming significant.

In previous studies, there have been many efforts to identify the determinants of student evaluations at universities, such as characteristics of faculty, courses, students, etc. First, as for faculty characteristics, many studies tried to examine the effects of faculty age, gender, and position (Feldman, 1984; Fernández & Mateo, 1997; Marsh, 2007; Ting, 2000). However, at most, the effect of faculty characteristics is found to be very minimal, and varies across studies.

Second, there are also many studies on the effects of course characteristics on student evaluations. A study on the electivity of a course suggests that instructors teaching an elective course usually receive higher scores of student evaluations compared to the instructors teaching a required course (Marsh, Hau, Chung, & Siu, 1997). Among the fields of study, student evaluations are the highest for the faculty of college of arts and humanities, but the differences are not

[*] Corresponding author. Tel:.+82 10 6341-5450; fax: +82 2 744 5717, +82 33 256 3424.

E-mail addresses: hooncho@hallym.ac.kr (D. Cho), bwy86@hanmail.net (W. Baek), trustcho@skku.edu (J. Cho).

http://dx.doi.org/10.1016/j.econedurev.2015.10.001

large (Ory, 2001). Class size sometimes affects student evaluations. Feldman (1984) found that a very large or a very small class receives higher scores of student evaluations. Yet, on the contrary, a study by Bedard and Kuhn (2008) found that as class size becomes larger, student evaluations become lower.

Third, as for students' characteristics, some studies indicated that a student's grade also influences student evaluations, suggesting a statistically significant positive relationship between the students' grades in the current course and student evaluations (Arnold, 2009; Heckert, Latier, Ringwald, & Silvey, 2006; Spooren, 2010). Those studies interpreted this relationship as a reflection of the teaching effectiveness or student learning. Because students learn more and better from faculty who teach effectively than those who do not, they can get higher grades; thus, it naturally follows that a faculty member would obtain better student evaluations.

Unlike these studies, many economics studies are concerned about bias on student evaluations because of the students' expectations about their course grades. From a "bias" point of view, the students with good grades tend to highly rate their instructors on teaching evaluations. Thus, instructors are likely to have some incentive to give more inflated grades to students since teaching evaluation can ultimately affect the promotion of instructors. According to this "grading leniency" hypothesis, the faculty tries to "relax" grading standards in order to receive higher evaluations (e.g., Brockx, Spooren, & Mortelmans, 2011). Several studies employ an expected grade in order to empirically test the grading leniency hypothesis on student evaluations because students do not know their own course grades at the time of the evaluations (Aigner & Thum, 1986; Ewing, 2012; Ginexi, 2003; Greenwald & Gillmore, 1997; Isely & Singh, 2005; Krautmann & Sander, 1999; Matos-Diaz & Ragan, 2010; McPherson, 2006). However, since the expected grade denoted in the student evaluation questionnaires tends to be noisy (i.e., imprecise or perhaps biased), some studies alternatively use the course grades (Brockx et al., 2011; Marsh, 1984; Spooron, 2010; Weinberg, Hashimoto, & Fleisher, 2009). For example, Weinberg et al. (2009) use the actual grade in the current course as a measure of the expected grade because students can have some idea of what grades they will receive based on midterm results, homework scores, and other objective information on their course performance, as well as any possible "signals" from the instructor, although students generally do not receive perfect information on final grades before completing their evaluations.

Instead of utilizing the expected or actual grades directly, some studies use the composite terms. For example, Isely and Singh (2005) use the gap between the expected grade and cumulative GPA as the relative expected grade, and Davies, Hirschberg, Lye, Johnston, and Mcdonald (2007) calculated the difference between the students' course grades and average grade for other courses being taken during the same semester. The students who obtained higher grades, relative to their expectation, would hence give a psychological "gift" to their teachers by giving higher evaluations; whereas, it is the opposite with those students receiving lower grades than their expectations.

Under the assumption that student with better grades are likely to give more favorable evaluations, this study focus on the possibility of a hedonic (implicit) exchange between the students' course grades and student evaluations. In this paper, the reservation grade is defined as the minimum grade expected by students. If the actual grade is higher than the reservation grade, then a grade surplus is realized; as a result, students provide higher evaluations based on their hedonic value of the grade surplus. On the other hand, if the actual grade is lower than the reservation grade or in the face of a negative grade surplus, the students pay back by rating teachers through lower evaluations. Since students with better grades are likely to have a positive hedonic value of grade surplus, the empirical estimation (without appropriately considering this component) may be biased. This study will identify the very existence of the bias factor, suggesting that the positive influence of the students' grades on teaching quality may have to be discounted as much as the bias factor.

This paper is organized as follows. Section 2 demonstrates how the data for teaching evaluation was created. Section 3 describes the summary statistics of the data used in this study. Section 4 describes a theoretical model of this analysis. Section 5 describes the empirical models. Section 6 reports the results and Section 7 is the conclusion.

2. Data of teaching evaluation

In order to empirically identify the existence of the aforementioned bias factor, which might exist in estimating teaching evaluations, this study exploited a very novel data set. In order for this empirical work to be done, there should be two groups of students. One group is composed of students who are informed of their grades in class, prior to submitting the evaluations, whereas another group includes those who are not informed of their grades. The novel data set was created by a system-related technical error that happened at one of the major universities in Korea, Sungkyunkwan University. The system error occurred in 87 classes at the College of Engineering during the spring semester of 2012. At the College of Engineering, course evaluations are largely divided into the two major types: one is for ABEEK programs[1] and the other is for general courses required for the major. The Information and Communications Center responsible for building a course evaluation system was supposed to classify course evaluations according to prescribed course type by setting "ABEEK" for the ABEEK programs and setting "null" for other major-related courses, before students evaluate their courses for the semester. However, the center failed to properly mark "null" for general courses, even though it properly set "ABEEK" for the ABEEK programs.

After the student evaluations and final exams were completed, students were able to check their grades. During the grade announcement period, the Information and Communications Center discovered that the course evaluation type was not properly set for other major-related courses. After correcting this error, the Information and Communications Center sent the data to the administrative division, which manages course evaluation for calculating evaluation scores.

[1] ABEEK programs refer to the engineering education programs accredited by the Accreditation Board for Engineering Education of Korea (ABEEK). Those programs are designed to nurture highly qualified engineers who are needed by major companies in some industries.

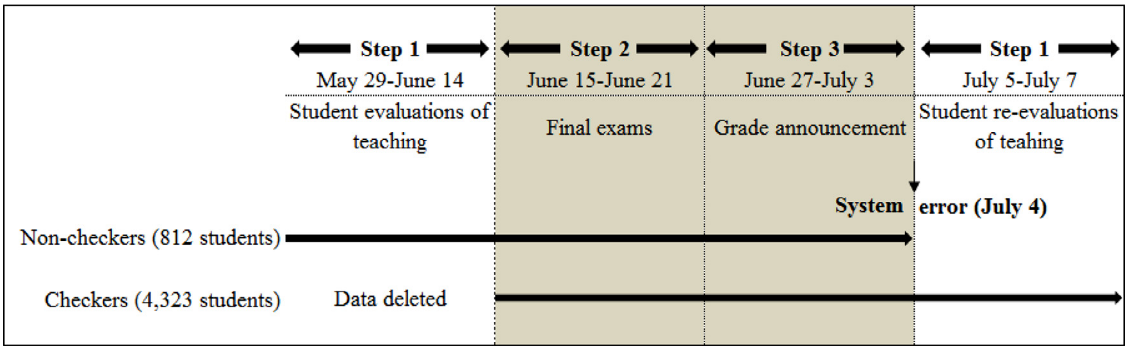

Fig. 1. Process of natural sample formation of student teaching evaluations at Sungkyunkwan University.

However, the course evaluation response rate was extremely low, raising doubt that there might be another error in the evaluation system. An investigation into the cause of the problem showed that parts of the course evaluation data submitted by some students were not properly saved and consequently deleted when the Information and Communications Center was re-setting “null” for the general courses of a major.

An examination of the distribution for the course evaluation respondents, whose responses were deleted, the results indicated that these errors did not occur only for some particular courses or majors. The errors occurred randomly among 87 general courses taken by students whose majors spanned five different fields in the College of Engineering.[2] The Information and Communications Center had to reconstruct the course evaluation program, and the administrative division had to make an announcement that course evaluations would be scheduled to take place again. The announcement was conducted via SMS, email, and phone calls to students whose course evaluations were missing. As a result of this administrative effort, approximately 86.5% of students who checked their grades took part in the re-evaluations. That is about 4.8% points lower than 91.3%, which was the response rate of students who provided their evaluations in accordance with the normal procedures.

Fig. 1 displays the process of how this novel data set was created. The data set consists of one group of students whose responses were randomly deleted because of a system error, as a result of inevitably having checked out their grades before conducting re-evaluations on the teaching; and the other group was of students who normally conducted their course evaluations without knowing their grades beforehand. In Fig. 1, the group, who provided their evaluations on teaching in normal course evaluation procedures, first gave evaluations for teaching before checking their grades (step 1→step 2→step 3). However, the other group, whose course evaluations were deleted because of a system error, already checked out their grades before providing re-evaluations for teaching (step 2→step 3→step 1). That is, students in both groups went through the steps of final exams (step 2) and grade announcement (step 3) in the same period. However, there is a difference as to whether the student evaluations of teaching come before the grade announcement or not.

This uniquely created data set enable this study to conduct a natural experiment where one can empirically test whether there can be a significant bias in the student evaluations of teaching. We can also examine the effects of the difference between the students’ expected grades and the actual course grades on the student evaluations of teaching.

3. A theoretical model

This study estimates the link between grades and student evaluations of teaching (SET) to determine the effect of bias on SET, according to the status of grade checking. In this section, a simple theoretical model can be established for the following empirical analysis. The basic estimating determinant of SET is:

$$SET = F(X) \tag{1}$$

where $F(X)$ is a SET mechanism based on the general characteristics discussed in previous research, including instructor characteristics (e.g., age, sex, tenured or not); student characteristics (e.g., age, gender, major, student’s expected performance); and course characteristics (e.g., size, whether or not it is required).[3]

This study estimates an innovative experiment that goes beyond the results of previous research regarding the determinants of SET. The central purpose of this study is to seek the existence of bias on SET because of the students’ expectations of their course grades and the actual course grades, which depends on the available information to the students at the moment of SET. The SET data of Sungkyunkwan University for this study is appropriate for the analysis because the sample of students who evaluated the course, after checking their grades, were created randomly.[4] The basic model with

[2] As for the percentage of students whose course evaluation responses were deleted, the deletion rate for students majoring in architectural engineering is 38.9%, and the rate for students majoring in mechanical engineering is 58.5%. Also, those rates for students majoring in systems management, advanced materials sciences, and chemical engineering, are 50.1%, 53.7%, and 46.8%, respectively.

[3] See for example Krautmann and Sander (1999), Isely and Singh (2005), McPherson (2006), Davies et al. (2007), Bedard and Kuhn (2008), and Langbein (2008).

[4] Previous research, especially in economics, supports the speculation that grades can influence SET, but it rarely distinguishes between the influence of actual and expected grade due to the difficulty of collecting data on actual or expected grades by separated groups whether knowing their actual grades or not.

the bias term can be described as follows:

$$SET = F(X) + B(EG) \tag{2}$$

where *EG*is the expected grade in the current course and this captures both the instructional quality of a course (as *EG* is included in *X*) and the bias effect of the grade on SET (*EG* as an argument of $B(\cdot)$).

One claim is that instructors who teach more effectively receive better SET because their students learn more in their courses, thereby earning good grades (Brockx et al., 2011; Marsh & Roche, 1997). In this case, a student's expected grade indicates good teaching and more learning. Another claim is that bias term captures the SET reward from a high expected grade or the SET penalty from a low expected grade (Ewing, 2012; Isely & Singh, 2005; Krautmann & Sander, 1999; McPherson, 2006). Since students generally do *not* know the actual grade at the point of teaching evaluation, the source of bias should be the expected grade in the current course. In the teaching evaluation analyzed in our paper, the expected grade is asked to students. Thus, this study utilized the expected grade as a source of the students' learning and bias term as well.[5]

Because of a random glitch, some of students had a chance to know their actual grade before doing their SET. As they recognized the existence of a gap between their actual grades and expected grades, the additional bias mechanism may be activated for the students experiencing the random glitch. The general model, including this additional bias term for students who are affected by the glitch, is:

$$SET = F(X) + B(EG) + dH(AG - EG) \tag{3}$$

where *AG* is the actual grade in the current course, and *d* is a design parameter for the course evaluation mechanism. The $d = 0$ means the SET in the state of "not checking" one's own grade, and $d = 1$ means SET in the state of "checking" one's own grade. Students, who are affected by the glitch, react to actual grades in their responses to SET because they fill out SET after getting their actual grades. The source of bias for students who are affected by the glitch can be two types. The first one is the expected grade and the second one is the difference between the actual grades and expected grades in the current course. After observing the actual grade, the students adjust their bias by the expected grade with the gap between actual grades and expected grades.[6]

If grades affect SET, students will give a reward (or penalty) to the instructor in return for the higher or lower grades. When the bias term in $B(EG)$ is positive, it suggests that students can evaluate the course *subjectively* in addition to the quality of the course or their own learning outcome, depending on the grade surplus that they did not expect.[7] It means that when students evaluate a course subjectively, a tacit agreement occurs between the student's grade surplus and SET for the instructor, in a manner that when the surplus grade is positive, the students provide the instructor with a premium of higher SET scores on the basis of the grade surplus. On the other hand, if the term is negative, the student penalized the faculty by giving low SET. Furthermore, the bias of reward and penalty is created by $B(EG)$ and $H(AG - EG)$ for the group of students experiencing the glitch. As $B(EG) + H(AG - EG)$ is positive, the students experiencing the glitch provide the instructor with a premium of higher SET scores. There is an extra source of bias for the students with the glitch, who can adjust $B(EG)$ by the term of $H(AG - EG)$, even if the students group experiencing no glitch should rely solely on $B(EG)$ for their bias.

This study empirically uses creative experiment data to test the hypothesis as to whether the subjective factor of students can operate in SET. The experiment was possible when two groups were compared in a situation, where some evaluate SET after checking their grades and others evaluate SET without checking their grades. The experiment is free from the sample selection or self-selection bias only when $d = 0$ and $d = 1$ are randomized in the model.

4. Descriptive statistics

The data for student evaluations of teaching includes individual students who were in 87 classes influenced by the system errors happened in the College of Engineering at Sungkyunkwan University, Seoul, South Korea, during the spring semester of the 2012 academic year. The SET is normally conducted on-line after students complete their final examinations. The students have to answer a set of teacher evaluation questionnaires, if they want to check their grades for the courses prior to receiving a formal transcript via mail. Hence, the response rate for SET is usually higher than 90 percent because most students wish to know their grades as soon as possible.

As aforementioned, a very interesting event, from the perspective of researchers, happened at Sungkyunkwan University, which is one of the major universities in South Korea. Some of the students were randomly allowed to see their grades before they participated in the teaching evaluations because of technical problems in the campus-electronic system. This event was only observed for those undergraduate students whose majors were related to several engineering fields. Those fields are mechanical, system management, advanced materials sciences, chemical, and architectural engineering. This natural experiment provides a good opportunity for testing the hypothesis on whether the SET *objectively* measures the effectiveness of the instructors' teaching in the college classroom. If students objectively evaluate the instructors' teaching performance, then SETs should not depend on the students' course grades or on whether they knew their grades in advance.

Table 1 summarizes the summary statistics by dividing the samples of students based on whether they knew their

[5] Some authors use the expected grades as given by students on SET form, and students have some idea of what grades they may receive, based on objective information such as midterm results and homework scores on their course performance (Aigner and Thum, 1986; Greenwald and Gillmore, 1997; Krautmann and Sander, 1999; Ginexi, 2003; Isely and Singh (2005); McPherson, 2006; Matos-Diaz and Ragan, 2010; Ewing, 2012).

[6] A relative measure of grades is the preferred explanatory variable in terms of representing a determinant of SET (Isely and Singh 2005; Davies et al., 2007; Ewing, 2012)

[7] Langbein (2008) use a Hausman test of simultaneity to examine the claims that SET are generally a valid indicator of teaching effectiveness or student learning. In this test, the course and faculty fixed effect dummied and the expected grade are assumed to be exogenous to both the actual grade and the SET. The results support the hypothesis that faculty are rewarded with higher SET if they reward students with higher grades.

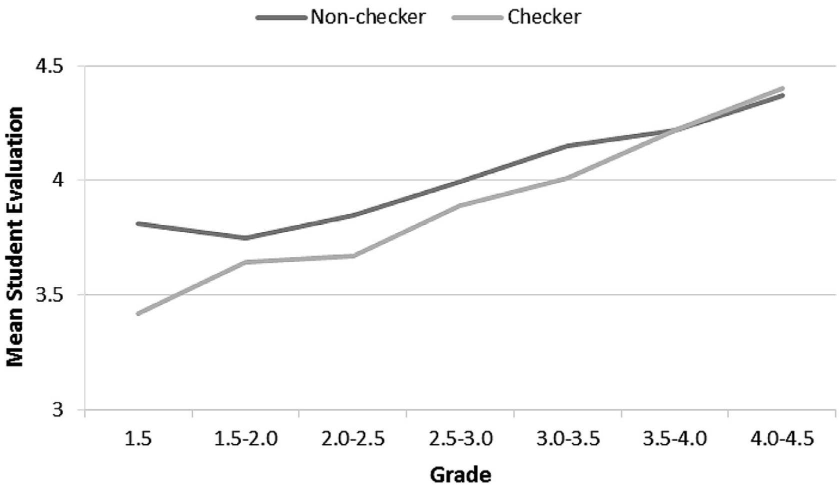

Fig. 2. Mean student evaluation across current course grades.

Table 1
Summary statistics.

	Non-checker	Checker	*p*-value
Class size < 30	0.04 (0.19)	0.03 (0.17)	0.18
30 ≤ class size < 50	0.29 (0.45)	0.27 (0.44)	0.24
50≤class size<70	0.14 (0.35)	0.17 (0.38)	0.03
Class size 70+	0.52 (0.49)	0.52 (0.49)	0.87
Student's age	22.69 (2.15)	22.82 (2.13)	0.12
Female student	0.25 (0.43)	0.19 (0.40)	0.00
Instructor's age	49.60 (7.41)	50.08 (7.47)	0.09
Assistant	0.08 (0.27)	0.07 (0.25)	0.24
Associate	0.22 (0.41)	0.21 (0.41)	0.21
Professor	0.66 (0.48)	0.70 (0.46)	0.02
Lecturer	0.03 (0.16)	0.02 (0.13)	0.04
Expected grade	3.95 (0.67)	3.95 (0.71)	0.93
Previous GPA	3.31 (0.69)	3.36 (0.73)	0.08
Grade in the current course	3.39 (1.01)	3.32 (1.02)	0.07
Sample size	812	4,323	

Standard deviations are in parentheses.

grades in advance, as from the 87 classes influenced by technical problems. The basic characteristics between two groups do not seem to be much different. For example, the average grade in the current course for grade-checkers (who already knew their current course grades at the time of the teacher evaluation period) is 3.32, while the average grade is 3.39 for non-checkers (who do not know their grades). The distribution of the previous GPA between the two groups shows a very similar pattern. Furthermore, the expected grades obtained from SET questionnaires, are shown to be the same between grade-checkers and non-checkers.[8] By examining those three measures of grades patterns, we are able to reasonably hypothesize that the system error randomly happened for the 87 classes analyzed in the paper. [9]

Appendix A shows the distributions of the difference in the grade between checkers and non-checkers, for each of 87 classes influenced by the system errors. Even though the overall difference of grade was not different between the two groups, a difference (of checkers and non-checkers) within a class seems to appear for many cases. This observed gap was expected since both of the class size and the proportion of checkers within a class are different for each class. For example, it is difficult to see a similar grade between the checkers and non-checkers within a very small size of class. Under the same reasoning, the grade gap is more likely to be observed in a class where the proportion of checkers is severely dominating. Thus, class fixed effects were accounted for in this empirical analysis, in order to control for the unobserved different characteristics of each class because of its small sample property.

Fig. 2 shows that the distributions of SETs are affected by the levels of the students' course grades, when they know them in advance. The measures of the teaching evaluation

[8] The p-values shows the results from hypothesis testing of statistical differences of means for each variable by two groups. Most of the key variables shown in the Table 1 are not statistically different between checkers and non-checkers at 5% significant level.

[9] Even though the system error occurred randomly, the distribution of the re-evaluation response rate shows that there may be some degree of selection bias, when compared to the initial response rate of the entire student body. However, this possibility of self-selection because of the re-evaluation process, which took place after the random system error, is not expected to significantly affect the empirical results for the following reasons. First, the student response rate for the teaching evaluation was very high in both student group, those that checked their grades and those who did not. The response rate for the group that checked is 86.5% while the rate for the group that did not check is 91.3%, thus indicating a small difference. Second, the expected grade surveyed in the questionnaire may reflect an unobserved heterogeneity of students in the teaching evaluation. Finally, in the regression analysis of this study, class fixed effects were used, in addition to using the expected grades, in order to control for any unobservable heterogeneity problem across classes.

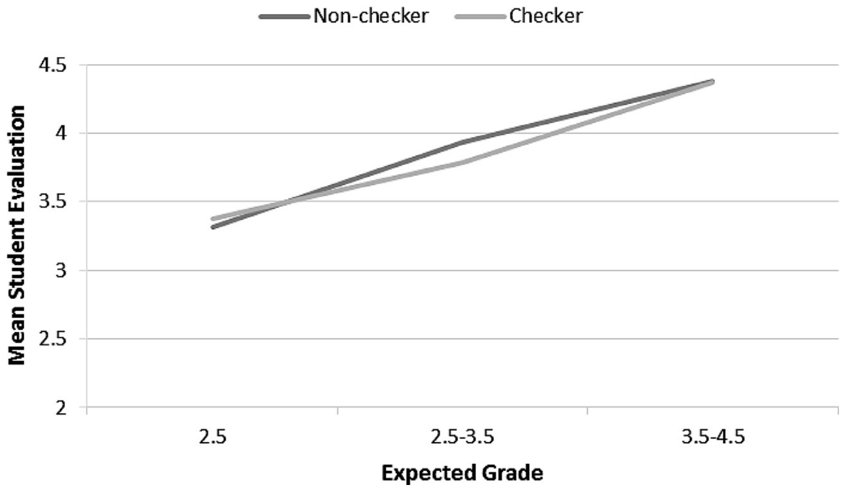

Fig. 3. Mean student evaluation across expected grades.

are coded into five different levels: (1) poor, (2) fair, (3) good, (4) very good, and (5) excellent. For students whose grades in the current courses are quite low (between 1.5 and 2.0 out of a 4.5 scale), the level of SETs of the students, who know their grades in advance, is shown to be quite lower than the level of SETs of students who do not know their grades with regard to the teacher evaluation. The observed gap of SETs between the grade-checkers and non-checkers is becoming narrower when students' grades are getting close to perfect scores.

Fig. 3 shows the level of SET between grade checkers and non-checkers across their levels of expected grades. As the expected grades of students rise, the SET becomes large as well, for both of the two groups. The gap of SETs between grade checkers and non-checkers are not much different regardless of their expected grades.

5. Empirical models

The following SET ratings were estimated in order to test as to whether the college students objectively evaluated their teacher performance:

$$SET_i = \alpha + \beta_{is} \sum_{s=1}^{n} X_{is} + \delta K_{is} + \varepsilon_i \tag{4}$$

where X_{is} is the vector of characteristics affecting SET and ε_i is a residual error term. The dependent variable is originally coded from (1) highest rating to (5) lowest rating but the order was reversed for interpretational purposes.

Vector X_{is} includes class size, the students' information (age, gender, and expected grade) and the instructor's information (age, tenured or not, and teaching position). The class size is categorized as 1–29, 30–49, 50–69, and 70+ students. The criterion for the cut-off points is to make sure that each group has a balanced sample size. The instructor's teaching position is originally categorized as lecturer (which is a semester-based contractual position), assistant professor, associate professor, and full professor position. As a full professor can be regarded as a tenured position in Korea, the position variable (tenured or not) is expected to capture the instructor teaching productivity (Krautmann & Sander, 1999). In addition to the position variable, this study controlled for a tenure variable that reflects an instructor's teaching experience (McPherson, 2006).

The key variable K_{is} is the interacting term between whether students knew their course grades at the time of the teacher evaluations and the information that students newly received from the system errors in SET. In other words, the information is described as the difference between grades in the current course and expected grade. Thus, K_{is} captures the students' attitude toward the teacher evaluation when they receive information on their grades. The δ is expected to be statistically non-significant if students objectively evaluate teacher performance since the role of the expected grade will capture the bias term in SET determination. If the students' attitudes toward teacher evaluation change as they know their grades in the current courses, δ is expected to be statistically significant. In particular, any retaliatory attitude of students will enforce the estimated coefficient of δbeing positive. In other words, the students will give a generous teaching evaluation when they happen to obtain a relatively higher actual grade than their expected grade and vice versa.

6. Main results

In Table 2, the empirical results of the SET determinant are presented, focusing on the students' behavior when they know their current course grades. First of all, column (1) shows that class size does not seem to be an important factor for determining teacher evaluations. However, the students' gender and age are shown to affect teacher evaluations. The "check" variable indicates whether students already knew their current course grades when they report their scores for the teacher evaluation. As can be seen, students who happen to know their grades in advance seem to report relatively lower scores, compared to the non-checkers. The estimated coefficient of the expected grade variable shows a strong positive correlation between the expected grades obtained from survey questionnaires and the scores for teacher evaluations.

Column (2) adds one interaction variable between the indicator of recognizing grades in advance and the gap of

Table 2
The determinant of student teacher evaluations.

	(1)	(2)
Class size 30–49	−0.273	−0.294
	(0.235)	(0.235)
Class size 50–69	−0.275	−0.297
	(0.238)	(0.233)
Class size 70 or more	0.035	0.027
	(0.254)	(0.254)
Student’s age	0.023**	0.025**
	(0.008)	(0.008)
Female student	−0.171**	−0.171**
	(0.034)	(0.034)
Instructor’s age	0.001	0.001
	(0.008)	(0.008)
Tenure (in years)	−0.010	−0.010
	(0.008)	(0.008)
Tenured professor	0.024	0.031
	(0.151)	(0.151)
Check	−0.101**	−0.060
	(0.036)	(0.037)
Expected grade	0.513**	0.536**
	(0.018)	(0.019)
Check* (grade-expected grade)	No	0.096**
		(0.023)
Class fixed effects	Yes	Yes
Sample size	4,376	4,376
R-squared	0.223	0.241

Notes: Robust standard errors are in parentheses.
* statistically significant at the 5% level
** statistically significant at 1%

grades between the current course and expected grade. Given that students already knew their grades in the courses, the empirical result indicates that as students obtained relatively higher grades compared to their expected grades, they gave higher points on their instructors’ teaching evaluation. In addition to their current grades, students seem to care about how low (or high) their grades are compared to their expectation of actual grades. The empirical result shows that the when students receive relatively higher grades than they expected, they provide higher teacher evaluation scores, given that the students knew their grades beforehand. On the contrary, students seem to show some type of retaliatory behavior through teacher evaluations when they receive lower grades than they expected in the classroom.

7. Conclusion

This study examined the students’ attitude toward grade-based teaching evaluation using SET data for the spring semester in the 2012 academic year. It compared and analyzed the behavior of students who evaluated a course under an exogenous experiment, including students who knew their grades versus those who did not.

With the unique data set created by a natural experiment, this study tried to empirically test whether there can be a significant bias in student evaluations of teaching by examining the effects of the difference between the students’ expected grade and the actual course grade, as reflected by the student evaluations of teaching. More specifically, SET was estimated, depending on before and after checking their grades, not in an artificial experiment, but in a natural environmental situation in order to explain the relationship between SET and the students’ inner characteristics.

This paper introduced the concept of surplus grades, which implies a gap between the course grade and the reservation grade that is determined by the expected grade. As a result, this study examined the reward mechanism of a surplus grade in SET. It demonstrated that as the surplus grade became greater, it had a statistically significant positive effect on SET. This result suggests that the trade-off between excess gain and teacher evaluation scores takes place in the form of students’ rewarding faculty with high teaching evaluation scores for the gain of the surplus grade, and in the opposite situation, giving a penalty. The fact, that the discrepancy between the grade received from the course and the expected grade affects SET, demonstrates that a subjective factor operates in SET. Furthermore this paper argues that the estimated size of the positive influence of the students’ grades on the teaching quality in the previous studies may be biased.

Given that the results of SET becomes a reference for the quality improvement of course content and can influence personnel decisions, these empirical results suggest that the psychological and subjective factors, such as the students’ attitude toward grades, can have a significant effect on SET.

Appendix A

Table A.1.

Table A.1
Distributions of difference in grade between checkers and non-checkers within each class.

Class	Grade Gap	Class	Grade Gap	Class	Grade Gap
1	−0.23	30	−0.26	59	−0.21
2	−0.02	31	+0.38	60	+0.60
3	−0.04	32	+0.19	61	−0.57
4	+0.51	33	−0.10	62	−0.15
5	−0.20	34	+0.02	63	−0.08
6	−0.70	35	+0.21	64	−0.72
7	−0.31	36	+0.21	65	+0.44
8	+0.12	37	−0.11	66	−0.80
9	+0.17	38	+0.36	67	+0.00
10	−0.46	39	+0.26	68	+0.07
11	−0.38	40	+0.23	69	−1.53
12	+0.13	41	+0.06	70	−0.38
13	+0.07	42	+0.75	71	+0.03
14	−0.37	43	+0.13	72	−0.47
15	−0.29	44	−0.04	73	−0.04
16	+0.16	45	−0.41	74	+0.52
17	−0.13	46	−0.03	75	+0.10
18	−0.30	47	+0.29	76	−0.24
19	−1.04	48	−0.05	77	+0.00
20	+0.31	49	+0.53	78	−0.10
21	−0.08	50	+0.29	79	−0.65
22	−0.14	51	−0.08	80	−0.35
23	+0.29	52	+0.00	81	+0.21
24	−0.17	53	−0.54	82	−0.79
25	−0.12	54	+0.48	83	+0.15
26	+0.48	55	−0.26	84	−0.36
27	+0.28	56	+1.28	85	−0.47
28	+1.51	57	−0.67	86	−0.04
29	−0.14	58	+0.50	87	−0.70

References

Arnold, Ivo J. M. (2009). Do examinations influence student evaluations? *International Journal of Educational Research, 48*(4), 215–224.

Aigner, Dennis J., & Thum, Frederick D. (1986). On student evaluation of teaching ability. *The Journal of Economic Education, 17*(4), 243–265.

Bedard, Kelly, & Kuhn, Peter (2008). Where class size really matters: Class size and student ratings of instructor effectiveness. *Economics of Education Review, 27*(3), 253–265.

Brockx, Bert, Spooren, Pieter, & Mortelmans, Dimitri (2011). Taking the grading leniency story to the edge. The influence of student, teacher, and course characteristics on student evaluations of teaching in higher education. *Educational Assessment, Evaluation and Accountability, 23*(4), 289–306.

Davies, Martin, Hirschberg, Joe, Lye, Jenny, Johnston, Carol, & Mcdonald, Ian (2007). Systematic influences on teaching evaluations: the case for caution. *Australian Economic Papers, 46*(1), 18–38.

Ewing, Andrew M. (2012). Estimating the impact of relative expected grade on student evaluations of teachers. *Economics of Education Review, 31*(1), 141–154.

Feldman, Kenneth A. (1984). Class size and college students' evaluation of their college instructors. *College Student Journal, 40*, 691–703.

Fernández, Juan, & Angel Mateo, Miguel (1997). Student and faculty gender in ratings of university teaching quality. *Sex Roles, 37*(11–12), 997–1003.

Ginexi, E. M. (2003). General psychology course evaluations: differential survey response by expected grade. *Teaching Psychology, 30*(3), 248–251.

Greenwald, Anthony G., & Gillmore, Gerald M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*(11), 1209–1217.

Heckert, Teresa M., Latier, Amanda, Ringwald, Amy, & Silvey, Brenna (2006). Relation of course, instructor, and student characteristics to dimensions of student ratings of teaching effectiveness. *College Student Journal, 40*(1), 195–203.

Isely, Paul, & Harinder, Singh (2005). Do higher grades lead to favorable student evaluations? *The Journal of Economic Education, 36*(1), 29–42.

Krautmann, Anthony C., & Sander, William (1999). Grades and student evaluations of teachers. *Economics of Education Review, 18*(1), 59–63.

Langbein, Laura (2008). Management by results: Student evaluation of faculty teaching and the mis-measurement of performance. *Economics of Education Review, 27*(4), 417–428.

Lee, Jaeseong, & Cho, Joonmo (2014). Who teaches economics courses better?: Using student-professor matched data for the principle of economics course. *Applied Economics Letters, 21*(13), 934–937.

Marsh, Herbert W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76*(5), 707–754.

Marsh, Herbert W., Hau, Kit-Tai, Chung, Choi-Man, & Siu, Teresa L. P. (1997). Students' evaluations of university teaching: Chinese version of the students' evaluations of educational quality instrument. *Journal of Educational Psychology, 89*(3), 568–572.

Marsh, Herbert W., & Roche, Lawrence A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*(11), 1187–1197.

Marsh, Herbert W. (2007). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology, 99*(4), 775–790.

Matos-Diaz, Horacio, & Ragan, James F. (2010). Do student evaluations of teaching depend on the distribution of expected grade? *Education Economics, 18*(3), 317–330.

McPherson, Michael A. (2006). Determinants of how students evaluate teachers. *The Journal of Economic Education, 37*(2), 3–20.

Ory, John C. (2001). Faculty thoughts and concerns about student ratings. *New Directions for Teaching & Learning, 87*, 3–15.

Spooren, Pieter (2010). On the credibility of the judge: A cross-classified multilevel analysis on students' evaluation of teaching. *Studies in Educational Evaluation, 36*(4), 121–131.

Ting, Kwok-fai (2000). A multilevel perspective on student ratings of instruction: lessons from the Chinese experience. *Research in Higher Education, 41*(5), 637–661.

Weinberg, Bruce A., Hashimoto, Masanori, & Fleisher, Belton M. (2009). Evaluating teaching in higher education. *The Journal of Economic Education, 40*(3), 227–261.