A Closer Look at the Relationship between Grades and Teacher Evaluation: The Role of Grade Knowledge

Tin-Chun Lin*

ABSTRACT. Economic behavior theory was used to frame an investigation of the role of grade knowledge in student ratings of professors. A two-group experiment was conducted in which one group had midterm exams and thus received grade feedback before completing student evaluations of teaching (SET), while the other group did not have a midterm exam and thus had no specific grade knowledge before completing the SET. Both groups had a final exam and received exam feedback after the SET was administered. Results revealed that only in the midterm condition were grades significantly associated with SET. That is, SET is more strongly related to grades when students have had clear grade feedback prior to SET administration, implying that both students and professors engage in economic behavior and that a reciprocal relationship exists between students and professors. Moreover, two implied debatable issues on this topic are briefly discussed in the conclusion. (A20; A22; C30)

I. Introduction

The identification of precise methods for assessing a teacher's teaching quality (or performance) is a perennially important issue in higher education because teaching quality is a primary factor in student performance (e.g., De Paola, 2009; and Lin, 2010). Thus far, student evaluation of teaching (SET) is the assessment system most commonly and widely used by administrators in almost all U.S. universities and colleges. Nevertheless, a question remains: *Do grades significantly affect student behavior in rating professors*? Indeed, this question has been broadly discussed and investigated in several previous studies (e.g., Seiver, 1983; Nelson and Lynch, 1984; Krautmann and Sander, 1999; Clayson, 2004; Isely and Singh, 2005; McPherson, 2006; Langbein, 2008; and Matos-Diaz and Ragan, 2010). Most of these studies have concluded that students' expected grades exert a positive and significant

School of Business and Economics, Indiana University–Northwest, 3400 Broadway, Gary, IN 46408. I wish to thank the editor, and two anonymous referees for very helpful discussion and advice. Additional correspondence address: Phone : (219) 980-6634; Fax: (219) 980-6916; E-mail: tinlin@iun.edu.

effect on SET.

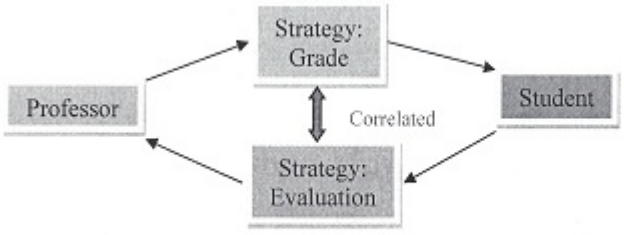
The positive relationship between student evaluations of teaching and student expected grades (or effective grades) rests on two factors: (1) students tend to offer more favorable evaluations in exchange for good grades, or (2) excellent instructors improve student skills to the point that they obtain good grades. Thus, if the positive effect is due to the first factor, both students and professors are engaging in economic behavior and a reciprocal relationship exists between students and professors. To verify that a positive relationship may be due to the first factor, it is necessary to use the theory of economic behavior as a basic framework when investigating the role of grade knowledge in student rating of professors. If specific grade knowledge positively and significantly affects students' behavior in rating professors, then a reciprocal relationship does exist between students and professors.

The reciprocal relationship has been broadly discussed in the business education literature (e.g., Clayson, 2004; Clayson, Frost, and Sheffet, 2006; and Lin, 2009a). The reciprocal relationship model based upon economic behavior theory shows that both groups will choose their best strategies to maximize their payoffs. In this paper we used the theory of economic behavior to look into the relationship between professors and students. Professors give their students grades while students also give their professors grades (i.e., student evaluation of teaching). In order to receive better evaluations from students, it is possible that some professors might adopt strategies (e.g., lowering the grading standard, creating easier exams, giving students extra bonuses [e.g., an attendance bonus], curving students' grades, avoiding some harder teaching materials that should be taught, being very nice to students, etc.) to ensure their students' approval. Strategies adopted by these professors may likely influence students' rating behavior. One of the most influential factors (or strategies) is students' grades, because grades may directly affect students' feelings (or emotions) and reflect their satisfaction with their professors. Clayson, Frost, and Sheffet (2006) explained the phenomenon as the reciprocal relationship between professors and students, in which students reward professors who give them good grades and punish professors who give them bad grades.

In 2009, Lin applied a static game of complete information to address the economic behaviors manifested between professors and students, and developed a model for the reciprocal relationship between professors and students. In the model, he created a production function of education

product constructed by both the professor and the student jointly and simultaneously. Each student and each professor has his/her own payoff function. His theoretical analysis showed that both the professor and the student will choose his/her best strategy to maximize his/her payoff and determine the Nash equilibrium. Consequently, professors' evaluations and students' grades are positively and endogenously correlated (see Figure 1). His theoretical evidence suggests that students' grades are one of the primary factors influencing students' rating behavior.

Figure 1–The Reciprocal Relationship between Professor and Student



Therefore, for this study we developed two research questions: (1) if students do not have grade knowledge before completing SETs, do course grades significantly affect student behavior in rating professors? And (2) If students have clear grade knowledge before completing SETs, do course grades significantly affect student behavior in rating professors?

To investigate these two research questions, we designed an experiment in which students were separated into two different groups: (1) no midterm exams, only one final exam; and (2) both midterm and final exams. In the first group, students had no grade knowledge before filling out the SETs, while students in the second group had clear grade knowledge (i.e., midterm exam grades) before filling out the SETs. We acknowledge that giving only a final exam may not be a good instructional method and do not promote it, but this strategy may be a better way to investigate the role of grade knowledge in student rating of professors.

This paper is organized as follows. First, we clarify several critical points regarding previous studies and briefly review some selected previous studies. Second, we describe our experiment design and data source. Third, we report the development of econometric models based upon our hypotheses. Fourth, we report our empirical results and provide a detailed discussion regarding the results. Fifth, we briefly discuss the

limitations of this research. Finally, conclusions may be found in the final section.

II. Clarification of Critical Points Regarding Previous Studies

Several critical points from previous studies should be clarified before moving on to the next section. Most previous studies have used self-reported grade expectations (i.e., expected grades) as a proxy for students' final grades (e.g., Seiver, 1983; Nelson and Lynch; Krautman and Sander, 1999; Isely and Singh; McPherson, 2006; Langbein, 2008; Matos-Diaz and Ragan, 2010). The reasons are understandable—student evaluations of teaching are anonymous, which means that it is impossible to match students' final grades with their SETs, and a student's expected grade may reflect grades on one or more midterm exams, problem set scores, term paper grades, class presentations, and/or attendance policies. Further, the expected grade also reflects anticipated information about the final exam and its grading.

To learn more about students' expectations for their final grades we conducted a preliminary study (via a survey)[1] which found that a student's expected grade covers two effects: (1) the student's midterm average grade; and (2) the student's satisfaction with the instructor's grading policy. For example, some instructors upwardly curve students' grades on the final exam. Other instructors give students several midterm exams but may drop the worst one or two scores. Such grading policies definitely benefit students and in turn impact their expectations about grades and hence their rating behavior. A student who is satisfied with an instructor's grading policy may be more likely to expect a higher grade because the grading policy appears to be favorable to his/her grade.

Based upon the survey evidence from our preliminary study, SET could relate to students' satisfaction with an instructor's grading policy. Therefore, it is reasonable to assume that the positive and significant effect between students' expected grades and SET scores may partially (or mostly) come from students' satisfaction with an instructor's grading policy rather than from "grades". To verify whether or not our assumption is correct, the effect of students' satisfaction with a grading policy needs to be extracted. Hence, in this study one question was added to a survey to ascertain students' satisfaction with the grading policy. If

the effect of students' satisfaction with a grading policy is expressed in a positive and significant manner on evaluations, this verifies that our assumption is correct. In addition, individual actual grades (i.e., midterm/final average grades) rather than expected grades were adopted.

It should be noted that while most researchers have used expected grades, some researchers have used actual grades. For example, Lanbein (2008) used both actual and expected grades for her investigation. However, the actual grade she used was the average actual grade in the class rather than individual student's actual grade.

Moreover, some researchers, such as Seiver (1983) and Nelson and Lynch (1984), have adopted a simultaneous framework to investigate this issue. Their basic argument in favor of simultaneity runs something like the following: instructors may have an "expected overall evaluation" that may be influenced by their grading policies and thus may influence students' "expected grades" and reported SET score. Readers may think that it seems far-fetched that an instructor's SET expectation can influence the students' expected grades. Indeed, it is not far-fetched. For example, if an instructor expects positive SET scores from students, the instructor may adopt a grading policy that is favorable to students' grades, such as curved, providing attendance bonus, dropping the worst one or two midterm-exam scores, etc. Due to a favorable grading policy, students may expect higher final course grades, which in turn may influence their SET scores. The instructor's "expected evaluation" is not observed but the class or student's "expected grade" is. The use of "expected grade" in their studies was required in order to run a simultaneous framework.

Additionally, expected grades were not adopted in this study; rather, the effects of midterm average grade and student's satisfaction with the grading policy were emphasized. This raises another question. In the second group, when the instructor gives students both midterm exams and a final exam, can students' final course grades and SET be simultaneously determined since midterm average grades may affect their emotions and in turn influence their rating behavior? That is, some (or many) students may use the SET to reward or exact revenge on their professors. Simply speaking, using midterm average grades, students can figure out their potential final grades for a course. Therefore, grades and SET are no longer simultaneously determined in this case. Thus, single-equation estimation was used for the second group. However, for the first group without midterm exams, students' final course grades and SETs

were simultaneously determined. Hence, a simultaneous framework was adopted for the first group. Some may be skeptical that these two variables (i.e., grades and SETs) can exert a simultaneous impact on one another because the prevailing thought is that a time sequence is generally in effect. The reason these two items are regarded as simultaneously determined is because students do not know their course grades when they fill in the SETs; similarly, the professor does not know the results of the SETs when he/she assigns students grades for the course. Unlike the group with midterm exams, students know their midterm average grades when they fill in the SETs so that they can forecast their course grades, which in turn may affect their rating behavior. Therefore, the effects for the group without midterm exams may be regarded as simultaneous even if technically they are not so.

Furthermore, the emphasis in the "typical SET study" was on cross-class and cross-instructor differences. The objective for the "typical SET study" was to understand why SETs differ across instructors and how grading policies affect that outcome. However, the macro data being used may fail to reflect individual differences across students. Therefore, an alternative course of action was selected for this study. Individual micro data were adopted and only one instructor was chosen in order to better understand why SETs differ across students given the same instructor and grading policy.

In addition to those critical points discussed above, some previous studies provided valuable findings that need to be pointed out.

1. Prior to 1983, almost all empirical work focused on single-equation estimation. Seiver (1983) adopted both a single-equation model and a simultaneous equation model. As a result, in his two-equation model estimated with two-stage least squares (2SLS), he did not suggest that students' expected grades related to their overall evaluations of teachers, but he found a statistically significant relationship between students' expected grades and SET in the ordinary least squares model.
2. Nelson and Lynch (1984) tested three hypotheses and concluded that: (1) the student evaluation process may create grade inflation (i.e., the easier the grading, the higher the teaching evaluations); (2) faculty who have been experiencing falling real incomes from teaching will adopt easier grading policies; and (3) the grade-evaluation relationship should be estimated using a simultaneous equations

model rather than a single equation model.

3. Krautmann and Sander (1999) considered grades as an endogenous determinant of SET. Therefore, they used the Instrumental Variable (IV) approach to solve the endogeneity problem. Findings indicated that grades positively affect student evaluations, which implies that faculty may intend to "buy" higher evaluations by lowering their grading standards. In addition, they pointed out an important conclusion: grade inflation may hurt the quality of higher education and weaken the signaling role of educational credentials in screening workers.
4. Clayson (2004) found that the reciprocity relationship is most critical in validating student rating behavior.
5. Isely and Singh (2005) showed that if an instructor for a particular course had some classes in which students expected higher grades, s/he would receive a more favorable average SET in these classes. In addition, they created another alternative variable—relative expected grade (i.e., expected grade relative to the incoming GPA of students). They concluded that the expected grade relative to incoming students' GPA provided more explanatory power.
6. McPherson (2006) tested for endogeneity and controlled for unobserved heterogeneity. He concluded that expected grades and SET are positively and significantly related, implying that instructors may "buy" higher evaluations by giving students higher grades. In his empirical work, he also found that an instructor's level of experience and class size were significant determinants of SET scores in principal classes.
7. Lanbein (2008) revealed that actual grades are positively and significantly related to SET scores, controlling for expected grade and fixed effects for both faculty and courses. After discussing the implications of this issue, she concluded that the SET has become a flawed measure of teaching quality and grades are a flawed signal of future job performance.
8. Matos-Diaz and Ragan (2010) applied risk aversion to the SET issue. They hypothesized that students are risk-averse, preferring less variability in grades. Thus, students' rating behavior will not only depend on their expected grades but also on the variance in expected grades (i.e., distribution of expected grades) because variance provides information on grading policy. Their empirical work showed that student ratings and expected grades are positively and

significantly related, but student ratings and variance in expected grades are negatively and significantly related. Their findings imply that faculty may be able to improve their SET scores by reducing the distribution of expected grades and increasing the mean for expected grades.

In short, the main differences between this study and previous studies may be briefly summarized as follows: (1) individual actual grades were adopted rather than expected grades; (2) a new variable—students' satisfaction with an instructor's grading policy—was developed to extract from the effect of expected grades; (3) cross-student data (individual micro data) were employed rather than cross-class data (macro data); and (4) an experiment was created by separating students into two groups—with and without midterm exams.

III. Experiment Design and Data Source

3.1. Experiment Design

In this experiment, four undergraduate Introduction to Microeconomics classes during the fall 2011 and spring 2012 semesters were chosen for the study. The group without midterm exams was enrolled in the fall 2011 semester class, while the group with midterm exams was enrolled in the spring 2012 semester class. That is, students in the group without midterm exams were given only one final exam, while students in the group with midterm exams were given three exams (i.e., two midterm exams and one final exam). Therefore, students' final grades in the group without midterm exams depended on the final exam, while students' final grades in the group with midterm exams were based on the average of these three exam scores.[2] There were 85 student participants in the fall 2011 semester and 90 student participants in the spring 2012 semester. There was no sufficient statistical evidence of a difference in the means for GPA (i.e., grade point average) between the fall 2011 and spring 2012 groups based on a two-tailed test at the 0.05 level of significance, implying that the quality of students in these two groups was not significantly different.

To conduct this experiment, the following factors needed to be held constant:

(1) *Teacher's instructional style and teaching materials.* Only one teacher was chosen to ensure consistent instructional style and teaching materials.
(2) *Incentive to attend class.* Students were given complete freedom to make their own choices about attending class. Thus, there were no mandatory attendance policies, no attendance bonus, and no quizzes.
(3) *Quality of classroom.* These two different groups (four different sections in two semesters) met in the same classroom to ensure the same classroom quality.
(4) *Same exam for each group and each section.* The same final exam was created for these two different groups (four different sections in two semesters), and the same midterm exams were also developed for two different sections in the second group (i.e., with midterm exam group, spring 2012). The final exam was comprehensive. All exams were collected when students turned in their answers and no student was allowed to use a cellphone/iphone. This practice decreases the probability that an exam will become public and makes it difficult for students to obtain information from a previous year's exams—the best and only ways to minimize the probability that students will gain information from a previous year's exams.

3.2. Data Source

The data used in this study included non-self-reported and self-reported. To avoid data collection problems or potential bias and to ensure that student data were confidential and anonymous, another instructor from a different institution was chosen.

The following four variables led to non-self-reported data:

(1) *Exam scores.* Both groups' exam scores were recorded. These exam scores can be used as a proxy for a student's grade performance. Students' grades used here were original grades without a curve.
(2) *Attendance record.* Daily attendance was taken by the instructor throughout both the fall 2011 and spring 2012 semesters. Students were aware that daily attendance was being taken but they were told by the instructor that it was just for record-keeping purposes and would not affect their final grades due to a lack of mandatory attendance policies. We understand that awareness of attendance monitoring could cause students to treat the variable as something

that might ultimately affect their grades. However, it is the only way to ensure correct attendance records. We believe that significant errors in data would result from asking students to self-report their attendance records. In addition, student's attendance may be used as a proxy for student's interest in the class. If a student is interested in the class, he/she will attend it more often. Alternatively, this variable may also be a proxy for conscientiousness—for example, even if a student has zero interest in the class topic. Moreover, it should be pointed out that the attendance might be a function of how good the teacher is, not just an exogenous measure of interest or conscientiousness.

(3) *Grade Point Average (GPA).* A student's grade point average (GPA) can be used to proxy his/her quality because GPA, regardless of a student's major, is a measure of a student's motivation and scholarly ability. Each student's GPA was provided by the office of the registrar.

(4) *Student's age.* Each student's age was supplied by the administration office. The office provided each student's birth date so that each student's age could be figured out. This variable can be used to proxy a student's maturity. In general, it is assumed that a student's maturity and age are positively related.

To collect the self-reported data, we created questionnaires for both semesters—these surveys were custom-designed for this experiment. Additionally, it was a challenge to match non-self-reported data with self-reported data due to the student data being confidential and anonymous. Thus, we adopted a strategy to keep the student data confidential and anonymous and also to avoid any possibilities of negativity. The strategy is described below.

In both semesters, on the day of the final exam, a proctor handed out the questionnaire to each student a few minutes before the exam began. Students were told that their ratings would be completely confidential and anonymous, so students understood and expected that the instructor would never see their individual SET scores. After all students finished the survey, the proctor went to each student to collect the response by herself, row by row and seat by seat. That is, students did not submit or pass their responses forward to the proctor, and there was no identification of each student on each response. After the proctor collected all responses from students, the proctor put the students'

responses in a big envelope and sealed it, and then the proctor handed out the final exam to each student. Students were told to place their answer sheets on their tables (but cover their answers) and then leave the classroom. After all students left, the proctor collected each student's answer sheet by herself and followed the same order as she had in collecting students' rating responses, row by row and seat by seat. Note that there was no seat number, but the proctor remembered the order in which she had collected students' responses. More importantly, when the instructor graded students' final exam, the instructor maintained the order created by the proctor and did not sort their exam answer sheets alphabetically; the proctor did not provide a sealed envelope containing students' rating responses to the instructor until the instructor had posted students' final grades online and submitted them to the office of the registrar. That is, the proctor gave students' exam answer sheets to the instructor first; then after the instructor posted students' grades online, she gave the sealed envelope to the instructor. We then picked up the sealed envelope and other information needed for the study from the instructor. This procedure ensured that students' final grades absolutely were not influenced by their rating responses.

The following nine variables led to self-reported data.

(1) *Student's efforts.* Two variables indicated this factor: (1) Frequency of studying for this class. There were five choices for this question: *1* = I study 1–5 hours before the test; *2* = I study 6–10 hours before the test; *3* = I study 11–15 hours before the test; *4* = I study 16–20 hours before the test; *5* = I study more than 20 hours before the test. (2) Frequency of practicing the study-guide before the exam. On a weekly basis, students were provided a study-guide with answers. There were five choices for this question: *1* = I never use the study-guide; *2* = I practice only once before the test; *3* = I practice 2 times before the test; *4* = I practice 3 times before the test; *5* = I practice more than 3 times before the test.

(2) *Student's work hours per week.* This variable can serve as a proxy for a student's opportunity cost of studying for the course. If a student works more, his/her opportunity cost will be higher. Students were asked to write down the total number of hours worked per week.

(3) *Student's math background.* Students were asked whether or not they had finished college algebra and calculus classes. This was a dummy variable so "yes" was set as 1 and "no" as 0. This variable was

considered because a math background is needed to learn economics since economics is more mathematical than other business and social sciences classes.

(4) *Depth of understanding of the lecture.* Students were asked: *how much do you usually understand the lecture in the class?* There were five choices for this question: *1* = Below 30%; *2* = 30 – 49%; *3* = 50 – 69%; *4* = 70 – 89%; *5* = Over 90%.

(5) *Professor's instruction skill.* Students were asked: *Do you agree that the instructor well organizes the lecture?*

(6) *Professor's communication skill.* Students were asked: *Do you agree that the instructor's speech and communication are clear and understandable?*

(7) *Professor's efforts.* Two questions were asked: (1) *Do you agree that the instructor is well prepared for the class?* (2) *Do you agree that the instructor is enthusiastic about teaching?*

(8) *Overall evaluation.* Students were asked: *Overall, I would rate the quality of this instructor as excellent.*

(9) *Student's satisfaction with grading policy*. Students were asked: *Overall, I am satisfied with the instructor's grading policy.*

The response options are the same for questions 5–9, which were five choices: *1* = No, I strongly disagree; *2* = No, I disagree but not strongly; *3* = Undecided; *4* = Yes, I agree but not strongly; *5* = Yes, I strongly agree.

In addition, a short summary of the data sources is presented in Table 1. There are three sources: instructor-supplied secondary data, university-supplied secondary data, and student-supplied primary data.

TABLE 1–Summary of the Data Source

Non-Self-Reported Data	Self-Reported Data
Instructor-supplied Secondary Data:	Student-supplied Primary Data:
Exam Scores	*Professor's instruction skill*
Attendance record	*Professor's communication skill*
University-supplied Secondary Data:	*Professor's efforts*
Student GPA	*Overall Evaluation*
Student age	*Student's efforts*
	Student's work hours per week
	Student's math background
	Depth of understanding of the lecture
	Student's satisfaction with grading policy

3.3 Descriptive Statistics

Table 2 reports means and standard deviations for the variables used in this study, t statistic, and K-S statistic. In addition, the reliability (i.e., Cronbach's alpha) of exams was measured. The Cronbach's alpha for the group with midterm exams was 0.87, which is high and indicates strong internal consistency among these exams. We were not able to compute Cronbach's alpha for the group without midterm exams, as there was only one exam.

In addition, two important points need to be mentioned before we present the results of the regressions.

(1) Satisfaction with grading policy was much lower (on average) when a midterm was not given, yet the overall evaluation was much higher (on average) when a midterm was not given. This is a very important finding, and we will provide a detailed discussion in a later section after presenting the empirical results.
(2) The average grade on the first midterm was 69, implying that many of the students failed the first midterm. Similarly, the final course grade was a C on average. Based upon the average GPA of 2.8, many of these students might receive a lower grade than they expected. It seems that the instructor may be an unusually "hard" teacher. Moreover, the teacher received worse evaluations from students when a midterm was given. That might be because a midterm was given and because it was graded hard. When these two reasons

coexist, students' rating behavior may be significantly related to grade feedback—the lower the grades, the worse the evaluations. That is, a reciprocal relationship exists between students and teachers—students reward teachers who give them good grades and punish teachers who give them poor grades. We will provide a detailed econometric investigation in the following section to prove our belief.

TABLE 2–Mean, Standard Deviation, t-Statistic, and K-S Statistic

Variables	Mean and Standard Deviation		Two-Tailed t Test	K-S Test K-S Statistic	
	(1) Fall 2011	(2) Spring 2012	(3) t Statistic	(4) Fall 2011	(5) Spring 2012
Overall Evaluation	4.08 (0.82)	3.56 (1.30)	3.23 (0.002)	0.295 (<0.010)	0.267 (<0.010)
First exam (scores)		69.08 (14.98)			0.067 (>0.150)
Second exam (score)		78.92 (12.96)			0.080 (>0.150)
Final exam (scores)	71.81 (12.99)	74.71 (13.08)	-1.47 (0.143)	0.095 (0.057)	0.085 (0.103)
Final course grade (scores)	71.81 (12.99)	74.03 (12.20)	-1.16 (0.247)	0.095 (0.057)	0.068 (>0.150)
Work hours per week	28.74 (14.29)	29.03 (13.77)	-0.14 (0.891)	0.194 (<0.010)	0.195 (<0.010)
Dummy variable-algebra	0.71 (0.44)	0.69 (0.47)	0.24 (0.808)	0.445 (<0.010)	0.437 (<0.010)
Dummy variable-calculus	0.35 (0.48)	0.36 (0.47)	-0.04 (0.971)	0.416 (<0.010)	0.414 (0.010)
Grade Point Average (GPA)	2.88 (0.49)	2.82 (0.49)	0.88 (0.379)	0.099 (0.044)	0.126 (<0.010)
Depth of understanding of the lecture	3.35 (01.08)	3.34 (1.08)	0.05 (0.959)	0.291 (0.010)	0.294 (<0.010)
Number of attendance (whole)	27.04 (3.50)	26.62 (3.97)	0.73 (0.466)	0.214 (<0.010)	0.216 (<0.010)

TABLE 2–Mean, Standard Deviation, t-Statistic, and K-S Statistic

Variables	Mean and Standard Deviation		Two-Tailed t Test	K-S Test K-S Statistic	
	(1) Fall 2011	(2) Spring 2012	(3) t Statistic	(4) Fall 2011	(5) Spring 2012
(continued)					
Number of attendance (midterm)		18.12 (2.21)			0.254 (<0.010)
Frequency of studying for the class	3.01 (1.14)	2.47 (1.13)	3.17 (0.002)	0.177 (<0.010)	0.315 (<0.010)
Frequency of practicing the study guide	3.52 (1.06)	3.67 (1.12)	-0.90 (0.369)	0.181 (<0.010)	0.183 (<0.010)
Well organized the lecture	4.29 (1.06)	4.30 (1.04)	-0.04 (0.970)	0.348 (<0.010)	0.349 (<0.010)
Well prepared for the class	4.62 (0.67)	4.61 (0.70)	0.12 (0.905)	0.430 (<0.010)	0.433 (<0.010)
Speech clear and understandable	3.44 (1.17)	3.47 (1.10)	-0.18 (0.856)	0.238 (<0.010)	0.252 (<0.010)
Enthusiastic about teaching	4.47 (0.85)	4.46 (0.84)	0.12 (0.907)	0.403 (<0.010)	0.387 (<0.010)
Student age	24.66 (5.18)	23.14 (4.33)	2.09 (0.038)	0.209 (<0.010)	0.193 (<0.010)
Satisfied with the grading policy	2.91 (0.63)	3.92 (0.71)	-10.06 (0.000)	0.359 (<0.010)	0.344 (<0.010)

Note: Number in parentheses in Columns (1) and (2) is standard deviation, while in Columns (3) - (5) is *p*-value.

IV. Econometric Models

In light of the research questions and given the data available for this study, we developed two testable hypotheses. Based upon these two hypotheses, we created econometric models to investigate this issue.

Hypothesis 1: *Without midterm exams (i.e., only one final exam), students' grades will not be significantly associated with SETs.*

To investigate Hypothesis 1, a simultaneous-equation model is required. Here, we used the Two-Stage Least Squares (2SLS) procedure to correct for simultaneous questions and to obtain unique estimates that were consistent and asymptotically efficient. Thus, in the first stage:

$$OEV = a_0 + a_1 WOR + a_2 WPR + a_3 ENU + a_4 DEP + a_5 ATD + \varepsilon_1 \text{ , and} \quad (1)$$

$$FGD = b_0 + b_1 FRS + b_2 FRP + b_3 ALG + b_4 CAL + b_5 GPA + b_6 ATD + b_7 WHR + \varepsilon_2 \text{ ,} \quad (2)$$

where *OEV* = overall evaluation; *WOR* = well organized the lecture; *WPR* = well prepared for the class; *ENU* = enthusiastic about teaching; *DEP* = depth of understanding of the lecture; *ATD* = total number of attended classes; *FGD* = student's final grade; *FRS* = frequency of studying for the class; *FRP* = frequency of practicing study guide; *ALG* = finished college algebra class; *CAL* = finished calculus class; *GPA* = grade point average; *WHR* = total work hours a week; and ε_1 , ε_2 = stochastic disturbance with a mean 0 and a variance σ^2 .

The results for Equations (1) and (2) are reported in Columns (1) and (2) of Table 3. $\hat{OEV}$ and $\hat{FGD}$ were saved—they are the predicted values of *OEV* and *FGD* as obtained from the reduced form estimates. The structural equations were estimated but *OEV* and *FGD* were replaced by $\hat{OEV}$ and $\hat{FGD}$. $\hat{OEV}$ and $\hat{FGD}$ are the instrumental variables (IV) here. Therefore, in the second stage the model for the professor's overall evaluation and student's final grade can be estimated in a linear form. The econometric models in the second stage can be expressed as follows.

$$OEV = \alpha_0 + \alpha_1 \hat{FGD} + \alpha_2 SPH + \alpha_3 AGE + \alpha_4 SAF + u_1 \text{ , and} \quad (3)$$

$$FGD = \beta_0 + \beta_1 \hat{OEV} + \beta_2 GPA + \beta_3 SPH + \beta_4 AGE + u_2 \quad (4)$$

where *SPH* = speech and communication are clear and understandable; *AGE* = student's age; *SAF* = student's satisfaction with the instructor's grading policy; and u_1 , u_2 = stochastic disturbance with a mean 0 and a variance σ^2.

In this formulation, the null hypothesis is that the parameters estimated by coefficients a_1 and β_1 are zero, while the alternative hypothesis is that the parameters are not zero.

Moreover, it should be noted that in the simultaneous equations model there is presumably an assumption of no correlation between the errors in both Equations (3) and (4).

Hypothesis 2: *With midterm exams, students' grades will be significantly associated with SETs*.

To investigate Hypothesis 2, a single-equation model is required. However, student's grade was an endogenous variable in the model[3]. When an endogeneity problem occurs, the Two-Stage Least Squares (2SLS) procedure is needed. Therefore, the student's grade was estimated in the first stage. The regression model was created in a linear form, such as:

$$GRD = d_0 + d_1 ATD + d_2 FRS + d_3 FRP + d_4 ALG + d_5 CAL + d_6 GPA + d_7 WHR + \mu_1 \quad (5)$$

where $GRD = GRD_F$ or GRD_M; GRD_F = final average grade (= mean of two midterm-exams and one final-exam scores); GRD_M = midterm average grade (= mean of two midterm-exam scores); $ATD = ATD_F$ or ATD_M; ATD_F = total number of attended classes in a semester; ATD_M = total number of attended classes in the midterm; and μ_1 = stochastic disturbance with a mean 0 and a variance σ^2.

The results for Equation (5) are reported in Columns (3) and (4) of Table 3. $\hat{GRD}$ was saved; the predicted value of GRD was obtained from the reduced form estimates. $\hat{GRD}$ was the instrumental variable (IV) here. Hence, in the second stage, the model for the professor's overall evaluation can be estimated in a linear form. The econometric model (Model 1) in the second stage can be expressed as follows.

$$OEV = \lambda_0 + \lambda_1 \hat{GRD} + \lambda_2 ENU + \lambda_3 WOR + \lambda_4 WPR + \lambda_5 DEP + \lambda_6 SPH + \lambda_7 AGE + \lambda_8 SAF + \tau_1 \quad (6)$$

where τ_1 = stochastic disturbance with a mean 0 and a variance σ^2.

TABLE 3–Determinants of *OEV, FGD, GRD$_F$*, and *GRD$_M$* in the First Stage

	Fall 2011 (No Midterm Exams)		Spring 2012 (With Midterm Exams)	
Explanatory Variables	OLS Explained Variable: *OEV* (1)	OLS Explained Variable: *FGD* (2)	OLS Explained Variable: *GRD$_F$* (3)	OLS Explained Variable: *GRD$_M$* (4)
Constant	0.814 (1.31)	-1.71 (-0.17)	11.82 (1.25)	9.78 (0.83)
WOR	0.395*** (3.94)			
WPR	0.116 (0.75)			
ENU	0.153* (1.70)			
DEP	0.082 (1.24)			
ATD	0.003 (0.16)	0.988*** (3.29)	1.05*** (4.04)	1.51*** (2.90)
FRS		4.558*** (4.82)	0.538 (0.57)	0.505 (0.47)
FRP		1.617 (1.60)	0.161 (0.17)	0.427 (0.40)
ALG		-0.323 (-0.14)	-0.636 (-0.29)	-1.067 (-0.43)
CAL		3.34 (1.57)	5.363** (2.50)	5.937** (2.46)
GPA		10.12*** (4.88)	12.50*** (6.13)	12.91*** (5.61)
WHR		-0.094 (-1.30)	-0.15* (-1.97)	-0.14* (-1.64)
R^2 $\overline{R^2}$	0.568 0.541	0.545 0.503	0.453 0.406	0.394 0.343
F-Statistic	20.81	13.17	9.69	7.63
Observations	85	85	90	90

Note: Number in parentheses is *t*-value; *OEV* = overall evaluation; *FGD* = student's final grade; *GRD$_F$* = final average grade; *GRDM* = midterm average grade; *WOR* = well organized the lecture; *WPR* = well prepared for the class; *ENU* = enthusiastic about teaching; *DEP* = deprth of understanding of the lecture; *FGD* = student's final grade; *FRS* = frequency of studying for the class; *FRP* = frequency of practicing study guide; *ALG* = finished college algebra class; *CAL* = finished calculus class; *GPA* = grade point average; *ATD* = total number of attended classes (*ATD$_F$* is for *GRD$_F$*; while *ATD$_M$* is for GRD$_M$); *WHR* = total work hours a week. ***p< .01; **p< .05; *p< .10

We also replace $\hat{FGD}$ in Equation (3) by $\hat{GRD}$, which would allow us to easily compare these two groups (without midterm and with midterm exams) due to the same explanatory variables. Thus, the econometric model (Model 2) in the second stage is shown below.

$$OEV = \gamma_0 + \gamma_1 \hat{GRD} + \gamma_2 SPH + \gamma_3 AGE + \gamma_4 SAF + \tau_2 \quad (7)$$

where τ_2 = stochastic disturbance with a mean 0 and a variance σ^2.

In this formulation, the null hypothesis is that the parameters estimated by coefficients λ_1 and γ_1 are zero, while the alternative hypothesis is that the parameters are not zero.

In addition, it should be pointed out that all of the error terms listed in the discussion of the econometric methods are based on an assumption of homoscedasticity rather than identical variances.

V. Results and Discussion

5.1. Results

Hypothesis 1

The results for Equations (3) and (4) are reported in Table 4. As Table 4 shows, student's final grade did not exert a statically significant effect on overall evaluation at any significant level in the all-students, Section I-students, and Section II-students groups.[4] Similarly, overall evaluation also did not exert a statistically significant effect on student's final grade at any significant level in the all-students, Section I-students, and Section II-students groups. These results imply that students' grades and SETs are not correlated. However, except for the Section II-students group, student's satisfaction with grading policy had a positive and statistically significant effect on overall evaluation at the 5% level in the all-students group and the Section I-students group, implying that student's satisfaction with grading policy and SET are correlated.

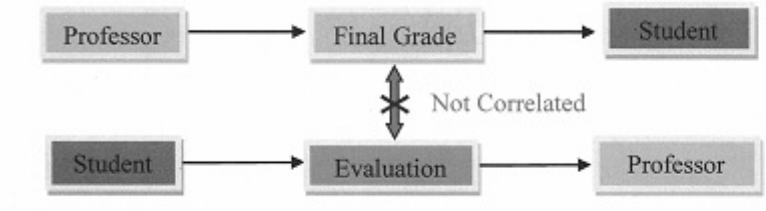
In short, Hypothesis 1 is supported. When students did not have grade knowledge prior to SET administration, their course grades were not significantly associated with SET. Figure 2 also explains this result.

TABLE 4–No Midterm Exams: Determinants of *OEV* and *FGD* in the Second Stage (Fall 2011)

	All Students		Section I		Section II	
Explanatory Variables	2SLS Explained Variable: *OEV* (1)	2SLS Explained Variable: *FGD* (2)	2SLS Explained Variable: *OEV* (3)	2SLS Explained Variable: *FGD* (4)	2SLS Explained Variable: *OEV* (5)	2SLS Explained Variable: *FGD* (6)
Constant	0.702 (0.98)	24.16** (2.21)	0.149 (0.13)	10.93 (0.66)	1.884* (1.93)	44.06** (2.54)
$\hat{FGD}$	0.012 (1.38)		0.011 (0.72)		0.010 (1.01)	
$\hat{OEV}$		3.184 (1.20)		3.789 (1.01)		0.039 (0.01)
GPA		7.682** (2.61)		10.756** (2.34)		5.996 (1.43)
SPH	0.230*** (3.11)	2.349* (1.75)	0.187 (1.35)	3.022 (1.54)	0.186** (2.28)	1.670 (0.85)
AGE	0.035** (2.43)	0.180 (0.71)	0.044* (1.73)	0.199 (0.51)	0.028* (1.75)	0.215 (0.63)
SAF	0.296** (2.37)		0.466** (2.19)		0.080 (0.52)	
$\frac{R^2}{R^2}$	0.334 (0.301	0.288 0.252	0.385 0.316	0.400 0.334	0.219 0.139	0.134 0.045
F-Statistic	10.04	8.08	5.63	6.01	2.73	1.51
Observations	85	85	41	41	44	44

Note: Number in parentheses is *t*-value; *OEV* = overall evaluation; *FGD* = student's final grade; *GPA* = grade point average; *SPH* = speech and communication are clear and understandable; *AGE* = student's age; and *SAF* = student's satisfaction with the instructor's grading policy.
***p< .01; **p< .05; *p< .10

Figure 2. Grade and Evaluation with out Midterm Exams



Hypothesis 2

The results for Equation (6) are reported in Table 5. As indicated there, both student's final average grade and midterm average grade exerted positive and statistically significant effects on overall evaluation at the 5% level in the all-students and at the 10% level in the Section I-students groups (but the effect is not statistically significant at any level in the Section II-students group).[5] Additionally, student's satisfaction with the grading policy exerted a positive and significant effect on overall evaluation at the 10% or 5% level in the all-students and Section I-students groups (while the effect is also not statistically significant at any level in the Section II-students group).

In addition, the results for Equation (7) are reported in Table 6. As Table 6 shows, both student's final average grade and midterm average grade exerted positive and statistically significant effects on overall evaluation at the 1% level in the all-students group and at the 5% level in the both Section I-students and Section II-students groups. Moreover, student's satisfaction with the grading policy exerted a positive and significant effect on overall evaluation at the 1% level in the all-students and Section I-students groups and at the 10% level in the Section II-students group.

Moreover, it should be noted that the coefficients of final and midterm average grades are much smaller than the other explanatory variables' coefficients (such as *ENU, WOR, WPR, DEP, WPR, DEP, SPH, and SAF*). This is because both grades are 100-point scale, while the other explanatory variables are 5-point scale. If we convert both grades to 5-point scale, the effect size of grades will become much bigger. For example, 0.021 (in Column 2 of Table 5) means that additional one point in midterm average grade is estimated to raise overall evaluation by approximately 0.021 points. If we convert it to 5-point scale, the effect size will become 0.42 (= (100/5) x 0.021), which means that additional twenty points in midterm average grade is estimated to raise overall evaluation by approximately 0.42 points.

Consequently, these results imply that grades and SETs are correlated, and that students' rating behavior may be affected by their satisfaction with the grading policy. In summary, Hypothesis 2 is supported. When students had grade knowledge prior to SET administration, their course grades were significantly associated with SET. Figure 3 also explains this result.

TABLE 5–With Midterm Exams: Determinants of *OEV* in the Second Stage (Spring 2012)–Model 1

	All Students		Section I		Section II	
Explanatory Variables	2SLS Explained Variable: *OEV* (1)	2SLS Explained Variable: *OEV* (2)	2SLS Explained Variable: *OEV* (3)	2SLS Explained Variable: *OEV* (4)	2SLS Explained Variable: *OEV* (5)	2SLS Explained Variable: *OEV* (6)
Constant	-4.280*** (-5.78)	-4.266*** (-5.63)	-4.345*** (-4.01)	-4.539*** (-4.02)	-5.568*** (-4.03)	-5.419*** -3.95)
$\hat{GRD}_F$	0.023** (2.50)		0.024* (1.79)		0.022 (1.51)	
$\hat{GRD}_M$		0.021** (2.36)		0.025* (1.88)		0.019 (1.37)
ENU	0.384*** (3.87)	0.388*** (3.89)	0.314** (2.09)	0.310** (2.07)	0.546*** (3.70)	0.550*** (3.66)
WOR	0.105 (0.99)	0.089 (0.83)	0.142 (1.03)	0.128 (0.93)	0.002 (0.01)	-0.023 (-0.13)
WPR	0.084 (0.54)	0.096 (0.61)	-0.004 (-0.02)	0.009 (0.04)	0.567* (2.00)	0.588** (2.07)
DEP	0.520*** (6.09)	0.522*** (6.08)	0.592*** (4.84)	0.586*** (4.80)	0.375*** (2.80)	0.389*** (2.93)
SPH	0.107 (1.39)	0.112 (1.45)	0.078 (0.70)	0.077 (0.68)	0.132 (1.18)	0.143 (1.27)
AGE	0.024 (1.56)	0.025 (1.63)	0.024 (0.93)	0.028 (1.06)	0.032 (1.61)	0.033 (1.62)
SAF	0.238* (1.91)	0.249** (2.01)	0.336* (1.79)	0.357** (1.96)	-0.024 (-0.12)	-0.031 (-0.16)
R^2 / $\bar{R}^2$	0.815 0.797	0.814 0.795	0.845 0.812	0.846 0.813	0.806 0.760	0.804 0.758
F-Statistic	44.64	44.20	25.85	26.08	17.66	17.42
Observations	90	90	47	47	43	43

Note: Number in parentheses is *t*-value; *OEV* = overall evaluation; GRD_F = *final average grade;* GRD_M = midterm average grade; *WOR* = well organized the lecture; *WPR* = well prepared for the class; *DEP* = depth of understanding of the lecture; *ENU* = enthusiastic about teaching; *SPH* = speech and communication are clear and understandable; *AGE* = student's age; and *SAF* = student's satisfaction with the instructor's grading.

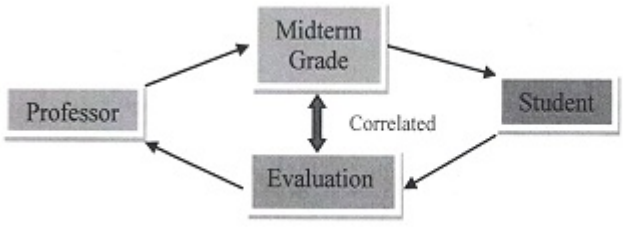
***p< .01; **p< .05; *p< .10

TABLE 6–With Midterm Exams: Determinants of *OEV* in the Second Stage (Spring 2012)–Model 2

	All Students		Section I		Section II	
Explanatory Variables	2SLS Explained Variable: *OEV* (1)	2SLS Explained Variable: *OEV* (2)	2SLS Explained Variable: *OEV* (3)	2SLS Explained Variable: *OEV* (4)	2SLS Explained Variable: *OEV* (5)	2SLS Explained Variable: *OEV* (6)
Constant	-4.637*** (-4.99)	-4.693*** (-4.99)	-4.827*** (-3.35)	-5.312*** (-3.61)	-4.131*** (-2.84)	-3.794** (-2.64)
$\hat{GRD}_F$	0.0423*** (3.53)		0.0418** (2.29)		0.0417** (2.29)	
$\hat{GRD}_M$		0.0420*** (3.54)		0.0471** (2.61)		0.0364** (2.05)
SPH	0.399*** (4.20)	0.404*** (4.27)	0.383*** (2.72)	0.374*** (2.69)	0.396*** (2.73)	0.419*** (2.88)
AGE	0.057*** (2.79)	0.060*** (2.92)	0.054 (1.57)	0.060* (1.76)	0.065** (2.28)	0.066** (2.27)
SAF	0.597*** (3.86)	0.605*** (3.94)	0.681*** (2.95)	0.690*** (3.11)	0.440* (1.76)	0.442* (1.74)
R^2 $\overline{R}^2$	0.618 0.601	0.619 0.601	0.647 0.614	0.659 0.626	0.530 0.480	0.518 0.467
F-Statistic	34.45	34.48	19.27	20.28	10.70	10.20
Observations	90	90	47	47	43	43

Note: Number in parentheses is *t*-value; *OEV* = overall evaluation; GRD_F = final average grade; GRD_M = midterm average grade; *SPH* = speech and communication are clear and understandable; *AGE* = student's age; and *SAF* = student's satisfaction with the instructor's grading.
***p< .01; **p< .05; *p< .10

Figure 3. Grade and Evaluation with Midterm Exams



5.2. Discussion

In addition to findings described above, we are left with one concern to clarify. Based upon the statistical evidence from the data used in this study, the professor's overall evaluations from the group without midterm exams were higher than those for the professor from the group with midterm exams.[6] This evidence might lead readers to argue that only giving one final exam would ensure better evaluations from students.

Additionally, finding that the SET is higher when there are fewer exams makes it easier for us to argue that students prefer less exam-work but does not necessarily indicate that we should argue that this grading procedure is somehow more fair, or led to better feedback, or for some other reason was highly desirable to students. More importantly, it also does not mean that we have support for the belief that fewer exams lead to better SET. To clarify this concern, we took a look at the correlation between SET and final course grade for these two groups. If the correlation was higher when students had grade feedback prior to SET administration, this may indicate that students knew their grades (i.e., midterm grades) prior to SET administration but not that having more exams alone would increase the correlation. To test the hypothesis, we used the one-tailed test (the upper tail test) by formulating the null (H_0) and alternative (H_a) hypotheses as below:

$$\begin{cases} H_0 : r_{12} - r_{11} \leq 0 \\ H_a : r_{12} - r_{11} > 0 \end{cases}$$

where r_{11} the population correlation between SET and final course grade for the group without midterms; and r_{12} the population correlation between SET and final course grade for the group with midterms.

We used "Fisher's *r*-to-*z* transformation" to transform correlations (r) into standard normal (z). As a result, the sample correlation between SET and final course grade for the group with midterms was 0.54, while the sample correlation between SET and final course grade for the group without midterms was 0.31. The *p*-value was 0.03, which is statistically significant at the 5% level, implying that the null hypothesis is rejected. Therefore, the answer is clear now—SET is more strongly related to grades when students have had clear grade feedback. That is, one of the main reasons for a lower SET when there are more exams is that students

in the group with midterms have had clear grade feedback prior to SET administration, and hence grade feedback affects their rating behavior.

Furthermore, on average, satisfaction with grading policy was much lower when midterm exams were not given to students. This might be because students would experience increased feelings of risk when only one exam determines their final course grades. In other words, a riskier grading procedure makes this grading policy less desirable to students. Nevertheless, based upon the evidence shown in Table 4, students' satisfaction with grading policy and overall evaluations are still positively correlated. This means that even though the grading policy is less desirable to students, students who were relatively more satisfied with the grading policy would still give their professor a better evaluation.

VI. Limitations

This research had two limitations—(1) experiment design; and (2) peer effect—that could possibly result in some potential errors. We illustrate these two limitations below.

6.1. Limitation of Experiment Design

Although we held constant important variables such as the teacher, classroom, exam, and double-blind protocol for collecting data in our experiment design, alternative designs could have been used to test the hypotheses while minimizing differences across the control and treatment groups. Our design led to some important differences across groups besides the treatment variable of two midterms: (1) the control group sections (no midterms) were all in the Fall 2011 semester classes, whereas the treatment group sections (with midterms) were all in the Spring 2012 semester classes; and (2) the class meeting time and number of times across sections could vary across groups. Of greatest concern was the treatment variable itself: the control group only had a final exam while the treatment group had two midterms and a final exam (equally weighted). This led to a considerable difference in grading policy and may have influenced students' rating behavior. Fortunately, we included "satisfaction with grading policy" as an explanatory variable in the regressions, making it much more strongly correlated with the Spring 2012 (treatment-group) semester courses (see Tables 4 and 6) and significantly higher in Spring 2012 (see Table 2). Furthermore, based

upon the one-tailed t-test at the 10% significant level, the final exam grade in the treatment-group was significantly higher because more frequent testing has been shown to have a positive impact on grades.

Our design created a considerable difference in grading policy and may have influenced SET, but it allowed us to verify that both expected grades and final course grades were not the "grades" affecting SET, and still provided convincing enough evidence and a step in the right direction for explaining the causes of grade inflation—a controversial issue in higher education today. In a future study, we will consider other alternative designs. For example, in one design we could have a midterm SET administered before the midterm in one group and after in the other. Another design could involve students who know that only the professor will observe the SET results, thereby reducing the incentive to reciprocate since there are no consequences to the professor. Both of these designs would be a bit cleaner and enables the hypotheses to be tested without changing the grading schema across groups.

6.2. Limitation of the Peer Effect

A "*peer effect*" may be found in students' rating behavior. For example, if a student had a bad experience with the professor for some reason, this student might share negative comments about the professor with other students. Therefore, the rating behavior of students may likely be influenced by the student's negative comments, and hence lead to a bias. Similarly, former students' comments (either positive or negative) about the professor may also affect current students' rating behavior. This is because students could simply distribute information and comments to all other students. Although the "*peer effect*" may exist, it cannot be easily observed and measured, creating a significant challenge in data collection. Thus, to collect such data, we will have to develop a strategy that maintains the confidentiality and anonymity of student data and also avoids any possibility of negativity. These will be left to future research on this issue.

VII. Conclusion

In this paper we developed a two-group experiment—one group had midterm exams and hence received grade feedback before completing

student evaluations of teaching, while the other group did not have a midterm exam and thus had no specific grade knowledge before completing student evaluations of teaching. Both of these two groups had a final exam and received final course grade feedback after the SET was administered.

In light of the empirical results described previously, three major findings may be summarized as follows:

1. Student's final grade was not significantly correlated with professor's overall evaluation in the no-midterm exams group, while student's final and midterm average grades were positively and significantly associated with professor's overall evaluation in the midterm-exam group. That is, neither expected grades nor final grades affect students' rating behavior, because expected grades are not real grades and final grades are determined after students fill out the evaluation. In fact, midterm average grades (or midterm grades) may significantly affect students' behavior in rating professors.
2. Overall, student's satisfaction with a grading policy exerts a positive and statistically significant effect on professor's overall evaluation in both groups (without and with midterm exams). The evidence confirms our assumption that the positive and significant effect between students' expected grades and SET scores may partially (or mostly) come from students' satisfaction with an instructor's grading policy rather than from "grades".
3. According to the correlation test, student evaluation of teaching is more strongly related to grades when students have had clear grade feedback prior to SET administration, which explains why the professor received a lower evaluation when there were more exams in the group with midterms because grade feedback prior to SET administration would affect students' rating behavior.

While our main finding seems to be similar to the main finding from past studies—students' grades are positively associated with SETs—our study indeed neither replicates nor refutes the main findings of past studies done on class averages (not the individual level). This is because we used individual actual grades and cross-student data rather than expected grades and cross-class data, which were commonly used in previous studies. Thus, students' rating behavior actually is influenced by midterm grades rather than expected grades.

Furthermore, before we wrap up this paper, two important implied debatable issues regarding SETs need to be pointed out.

1. In addition to the issue of *halo error*[7] in student evaluation of teaching (Orsini, 1988; Clayson, 1989, 1999; Clayson and Haley, 1990; Simpson and Siguaw, 2000; Clayson and Sheffet, 2006; and Madden, Dillon, and Leak, 2010), this assessment system ignores the existence of economic behaviors between teachers and students. Both teachers and students are economic individuals so both will respond to each other via economic behaviors. Therefore, the best strategy for receiving good grades and feedback is "*collusion*". In other words, teaching evaluations may lead professors to intentionally inflate grades[8] in order to receive good comments (a form of "*cheating*") and thus foster "*collusion*". As Simpson and Siguaw (2000) reported, some (or many) faculty members may use "*halo effects*" to their advantage by managing student evaluation. Similarly, some faculty members could also use "*collusion*" to manipulate evaluations.
2. Obermiller, Fleenor, and Raven (2005) pointed out that the role of students and their relationship with professors are always complex. Their survey showed that students generally prefer the customer orientation. A recent online survey on campus also found that the majority of students perceive themselves as "*customers*" in school[9]. If students are customers, then professors become "*servers*" or "*sellers*". Then the question is: how can servers evaluate (or judge) their customers? (Only customers evaluate servers.) That is, professors (i.e., "*servers*") are not allowed to test their students (i.e., "*customers*") and thus cannot provide grades for their students (i.e., "*customers*"). Rather, if grades must be given to students, the only grade that "*servers*" are allowed to give "*customers*" must be A (or A+), because "customers are always right". Therefore, it is impossible for "*servers*" to fail "*customers*". Hence, Franz (1998) stated that this belief could lead higher education to emphasize entertainment and professors whose role is to delight students rather than truly teach them. On the other hand, if it is inappropriate to regard students as "*customers*", then professors should not be regarded as "*servers*" and a customer orientation is no longer relevant or appropriate.[10] For that reason, the question here is: how can students evaluate (or judge) professors? In other words, under the circumstances described here, students do not have the right to evaluate professors. Unfortunately, in a system focusing on student

evaluation of teaching, both professors and students evaluate one another at the same time, creating a "*paradox of roles*" between professors and students.

In summary, the main contribution of this study to economic education is our verification of an important fact: midterm grades (grade feedback prior to SET administration) actually affect students' rating behavior, implying that both students and professors do engage in economic behaviors and that a reciprocal relationship does exist between students and professors. Hence, grades (especially midterm grades) could have been frequently used as a strategy to influence students' rating behavior. The school authority should be aware of the importance of this fact and identify a better form of faculty performance assessment in order to avoid the existing reciprocal relationship between students and professors.[11]

References

Clayson, D. E. 1989. "Halo Effects in Student Evaluation of Faculty: A Question of Validity." Paper presented at "Positioning for the 1990s," *Proceedings of the Southern Marketing Association*, New Orleans, LA.

Clayson, D. E. 1999. "Students' Evaluation of Teaching Effectiveness: Some Implications of Stability." *Journal of Marketing Education*, 21(1): 68–75.

Clayson, D. E. 2004. "A Test of the Reciprocity Effect in the Student Evaluation of Instructors in Marketing Classes." *Marketing Education Review*, 14(2): 11–21.

Clayson, D. E., Frost, T. F., and Sheffet, M. J. 2006. "Grades and the Student Evaluation of Instruction: A Test of the Reciprocity Effect." *Academy of Management Learning and Education*, 5(1): 52–65.

Clayson, D. E., and Haley, D. A. 1990. "Student Evaluation in Marketing: What is Actually Being Measured?" *Journal of Marketing Education*, 12(3): 9–17.

Clayson, D. E., and Haley, D. A. 2011. "Are Students Telling us Truth? A Critical Look at the Student Evaluation of Teaching." *Marketing Education Review*, 21(2): 101–112.

Clayson, D. E., and Sheffet, M. J. 2006. "Personality and the Student Evaluation of Teaching." *Journal of Marketing Education*, 28(2): 149–160.

De Paola, Maria. 2009. "Does Teacher Quality Affect Student Performance? Evidence from an Italian University." *Bulletin of Economic Research*, 61(4): 353–357.

Franz, R. 1998. "Whatever You Do, Don't Treat Your Students Like Customers." *Journal of Management Education*, 22(1): 63–67.

Isely, P., and Singh, H. 2005. "Do Higher Grades Lead to Favorable Student Evaluations?" *Journal of Economic Education*, 36(1): 29–42.

Jones, E. B., and Jackson, J. D. 1999. "College Grades and Labor Market Rewards." *Journal of Human Resources*, 25(2): 253–266.

Krautmann, A. C., and Sander, W. 1999. "Grades and Student Evaluations of Teachers." *Economics of Education Review*, 18(1): 59–63.

Langbein, L. 2008. "Management by Results: Student Evaluation of Faculty Teaching and the Mis-measurement of Performance." *Economics of Education Review*, 27(4): 417–428.

Lichty, R. W., Vose, D. A., and Peterson, J. M. 1978. "The Economic Effects of Grade Inflation on Instructor Evaluation: An Alternative Interpretation." *Journal of Economic Education*, 10(1): 3–11.

Lin, T.-C. 2009a. "Application of a Static Game of Complete Information: Economic Behaviors of Professors and Students." *Economics Bulletin*, 29(3): 1683–1691.

Lin, T.-C. 2009b. "Implications of Grade Inflation: Knowledge Illusion and Economic Inefficiency in the Knowledge Market." *Economics Bulletin*, 29(3): 2321–2331.

Lin, T.-C. 2010. "Teacher Quality and Student Performance: The Case of Pennsylvania." *Applied Economics Letters*, 17(2): 191–195.

Madden, T. J., Dillon, W. R. Dillon, and Leak, R. L. 2010. "Students' Evaluation of Teaching: Concerns of Item Diagnosticity." *Journal of Marketing Education*, 32(3): 264–274.

Matos-Diaz, H., and Ragan, J. F. 2010. "Do Student Evaluations of Teaching Depend on the Distribution of Expected Grade?" *Education Economics*, 18(3): 317–330.

McPherson, M. A. 2006. "Determinants of How Students Evaluate Teachers." *Journal of Economic Education*, 37(1): 3–20.

Nelson, J. P. and Lynch, K. A. 1984. "Grade Inflation, Real Income, Simultaneity, and Teaching Evaluations." *Journal of Economic Education*, 15(1): 21–37.

Obermiller, C., Fleenor, P., and Raven, P. 2005. "Students as Customers or Products: Perceptions and Preferences of Faculty and Students." *Marketing Education Review*, 15(2): 27–36.

Orsini, J. L. 1988. "Halo Effects in Student Evaluations of Faculty: A Case Application." *Journal of Marketing Education*, 10(2): 38–45.

Sabot, R., and Wakeman-Linn, J. 1991. "Grade Inflation and Course Choice." *Journal of Economic Perspectives*, 5(1): 159–170.

Seiver, D. A. 1983. "Evaluations and Grades: A Simultaneous Framework." *Journal of Economic Education*, 14(3): 32–38.

Shapiro, C., and Stigliz, J. 1984. "Equilibrium Unemployment as a Worker Discipline Device." *American Economic Review*, 74(3): 433–444.

Simpson, P. M., and Siguaw, J. A. 2000. "Student Evaluations of Teaching: An Exploratory Study of the Faculty Response." *Journal of Marketing Education*, 22(3): 199–213.

Spense, M. 1973. "Job Market Signaling." *Quarterly Journal of Economics*, 88(3): 355–374.

Endnotes

1. A survey was conducted in spring 2008 to learn more about students' expectations for their final grades. This survey was distributed and collected sometime after the midterm exams but before the final exam. (Note: responses were anonymous.) Students were asked two questions and for one explanation: "*What grade for this class do you expect to receive? Why do you expect that grade? Express your*

reasons." The results showed that students' expected grades basically depended on their midterm average grades and a probable curve on final grades after the final exam (Note: the instructor's grading policy states: each exam weights one third of the final grade, and I reserve the right to curve up your final grades if the overall final average of the class is below 78%.) They also believed that they would do at least as well on their final exams as they did on their midterm exams, or even better.

2. The instructor's grading policy for fall 2011 and spring 2012 is described below. Fall 2011: (1) *Graded items*: the final exam is the only graded item for this course, and this item is graded on a 100-point scale. The final exam is comprehensive. There are no projects, no term papers, no mandatory attendance policy, no attendance bonus, and no quizzes. Participation in class discussions is encouraged but not part of the course grade. (2) *Grade scale*: following the standard straight scale (no curved). (3) *Weights of graded items*: the final exam is 100% of the course grade. The grading policy for spring 2012 is the same as the other group in fall 2011 except for graded items consisting of two midterm exams and one final comprehensive exam. Each exam is one-third of the course grade.
3. Based upon the Hausman specification test, the null hypothesis that student's grade is an exogenous variable is rejected, implying that it is an endogenous variable.
4. The main difference between the Section I-students and the Section-II students is that the Section I-students meet in the morning, while the Section II-students meet in the afternoon. We report their regressions separately instead of just creating a Section-II dummy variable, because it allows us to see how each explanatory variable impacts the explained variable in the regressions in each section.
5. The same reason as in Endnote 4 for why we report their regressions separately rather than just creating a Section-II dummy variable.
6. We used the one-tailed test. As a result, the mean for the overall evaluation was significantly higher in the group without midterm exams than in the group with midterm exams at the 1% level.
7. The halo error is "a mistake or bias that can occur in evaluating an individual's performance where they are consistently rated based on the evaluator's overall impression, rather than on their actual performance in various categories" (see http://www.businessdictionary.com.) According to evidence in previous studies (e.g., Orsini, 1988; Clayson, 1989, 1999; Clayson and Haley, 1990; Clayson and Sheffet, 2006; and Madden, Dillon, and Leak, 2010), a number of factors that determine student evaluations of teaching are not related to professors' actual teaching performance. For example, Clayson and Sheffet (2006) discovered that a professor's personality would significantly influence student behavior in rating professors. In addition, Clayson (1989) showed that, for marketing students, the halo error in student evaluation of teaching is highly related to expected grade.
8. Grade inflation has a number of negative impacts on both the labor and knowledge markets. Briefly, it can: (1) make education a less efficient signal (Spence, 1973); (2) lead universities/colleges toward the Giffen good case (Lichty, Vose, and Peterson, 1978); (3) exacerbate information asymmetry and raise monitoring costs (Shapiro and Stigliz, 1984); (4) lead to signals of comparative advantage to bias (Sabot and Wakeman-Linn, 1991); (5) create a biased signal in the labor market (Jones and Jackson, 1999); and (6) create knowledge illusion and economic inefficiency in the knowledge market (Lin, 2009b).
9. In March 2010, an online survey on campus was conducted. The question was: *"Do*

you have the perception that students are customers in school?" There were two choices: yes or no. In total, 1,016 students responded to this survey within a few days. A total of 54% of the sample chose "yes", while 46% chose "no".

10. The George Mason University Faculty Senate passed a resolution officially stating that it is inappropriate to regard students as customers. The Faculty Senate Statement: *"Corporate models" of education in which students are viewed as "customers" are not appropriate. Education is a unique activity in a democratic society that differs markedly from both business and government. Universities are absolutely essential in contemporary society as centers of free inquiry, free expression, open discovery, and dissent. Any attempt to force education into a corporatist mold devalues faculty, lowers academic standards, and harms both students and the institution itself"* (GMU Faculty Senate, September 2002).
11. In addition to the problem of the reciprocal relationship between students and professors and the halo error in student evaluations of teaching, students may not offer honest responses on evaluations. Clayson and Haley (2011) found that a majority of students admitted that an estimated 30% of their answers on evaluations were not true.