



Repositorio Institucional de la Universidad Autónoma de Madrid

<https://repositorio.uam.es>

Esta es la **versión de autor** del artículo publicado en:
This is an **author produced version** of a paper published in:

The Journal of Economic Education 44.1 (2013): 32-46

DOI: <http://dx.doi.org/10.1080/00220485.2013.740387>

Copyright: © 2013 Taylor & Francis Group, LLC

El acceso a la versión del editor puede requerir la suscripción del recurso
Access to the published version may require subscription

Experimental Evidence on the Effect of Grading Incentives on Student Learning in Spain

Joaquín Artés and Marta Rahona

In this article, the authors aim to identify the causal effect of the use of graded problem sets on academic performance of Spanish students. The identification strategy relies on an experiment in which the authors exploit variation arising from observing the performance of nearly 300 students taking the same class during the same semester and with the same instructors. Academic performance is measured through a multiple choice final exam in which some questions are related to graded problem sets and others are related to non-graded problem sets given through the semester. After accounting for potential biases and selection concerns, the results show that graded problem sets increase test scores by eight percentage points, or close to a letter grade.

Keywords *classroom experiment, homework, student performance*

JEL codes *A22, C93, I21*

It has been shown in many articles that one of the most important determinants of students' academic achievement is teacher quality (e.g., Rivkin, Hanushek, and Kain 2005). Very little is known, however, about the specific characteristics or teaching techniques that make a teacher better. This lack of knowledge makes education policies aimed at improving teacher quality likely to fail. This article is focused on the effect of a specific teaching technique—the use of graded problem sets—on academic achievement.

In studies dealing with the effect of teacher quality on academic performance, differences in teacher quality are measured by comparing the results on standardized tests of different students within the same school over years (e.g., Rockoff 2004; Aaronson, Barrow, and Sander 2007; Rivkin, Hanushek, and Kain 2005). They have found that there are significant differences in teaching effectiveness even among teachers within the same schools. They also found, however, that differences among teachers' effectiveness are not generally correlated with objective

characteristics such as teachers' education, age, salary, or experience (Hanushek 1986; Rivkin, Hanushek, and Kain 2005; Kane, Rockoff, and Staiger 2008; Carrell and West 2010). The inability to identify a set of objective characteristics that good teachers share remains a challenge for researchers and policy makers.

Joaquín Artés is an associate professor of economics at the Universidad Complutense de Madrid, and the corresponding author (e-mail: jartes@der.ucm.es). Marta Rahona is an assistant professor of economics at the Universidad Complutense de Madrid.

The authors are grateful to Associate Editor Sam Allgood and the referees for their helpful comments. They acknowledge financial support from Universidad Complutense de Madrid.

In order to solve this challenge, in recent articles on academic achievement the focus has moved from teachers' characteristics to teaching methods. Based on theoretical models and empirical articles that suggest that student effort is the single most important determinant of student success (e.g., Stinebrickner and Stinebrickner 2008), identification of teaching methods and techniques that create adequate incentives for students to put forth the effort needed to succeed has been attempted in recent empirical articles on teaching effectiveness. For example, Figlio and Lucas (2004) and Betts and Grogger (2003), using U.S. data, suggested that teachers with higher grading standards improved the average academic performance of their students on standardized tests (although the gains were not equally distributed across students). Similarly, Bonesrønning (2004) evaluated different grading practices in a sample of Norwegian schools and also concluded that hard graders obtain better results than easy graders. Oettinger (2002) studied the effect of different grading schemes on academic performance and found that absolute grading schemes (e.g., the A–F scale) and a continuous grading scheme do affect student effort differently, and thus academic performance as well.

Another teaching method, the use of graded problem sets, is the subject of analysis in this article. Graded problem sets have the potential to improve student learning because they create a stronger incentive to work early on the material compared to non-graded problem sets, and because they provide students with regular feedback (the grade) on their level of understanding. The effectiveness of graded problem sets in improving student learning has been analyzed in Grove and Wasserman (2006), Grodner and Rupp (2011), and Geide-Stevenson (2009). Grove and Wasserman (2006) analyzed whether students performed better in college economic classes when graded problem sets counted toward their grade. In order to identify the causal effect of graded problem sets on academic achievement, they took advantage of a natural experiment that allowed them to compare the performance of two similar groups of students who took the same class under different grading requirements. They found that the group of students for which problem sets did not count toward their grade did significantly worse on the same final exam than the group of students for which problem sets were part of the grade. In particular, they found a learning gain of one-third of a letter grade for the average student in the treatment group.

Grodner and Rupp (2011) used a randomized experimental design and found that students in the homework-required group performed half of a letter grade better on tests. Geide-Stevenson (2009) used a methodology similar to Grove and Wasserman (2006), but applied it to a sample of older and more experienced students and found no difference between students in the treatment group and students in the control group, suggesting that more experienced students may not need the regular feedback that graded problem sets provide in order to obtain a similar learning gain.

Several other articles have used survey data and non-experimental designs and also have found a positive relation between homework assignment and test performance for samples of students other than college level (e.g., Eren and Henderson 2008, 2011; Cooper, Robinson, and Patall 2006). Although this research has mainly focused on U.S. schools, in recent articles like Falch and Rønning (2011) similar conclusions were reached using a sample of 16 OECD countries. While this literature suffers from identification concerns compared to the use of experimental

designs, their results contribute to a body of evidence that indicates that at least part of the learning gain that students obtain from teachers may come from objective teaching methods that can be learned by instructors and applied in the classroom.

This article contributes to the above literature by providing the results from an experiment carried out in college economic classes in Spain. To our knowledge, no article before has investigated the relationship between problem sets and academic achievement at the college level in countries other than the United States. The choice of the country and the timing of the study also are particularly relevant. During the last decade, Europe has undertaken a major reform of its higher education system by implementing the European Space for Higher Education (ESHE). While the most publicized goals of the reform are to improve the quality of European universities and to gain convergence and mobility across them, a practical consequence of the reform, in the case of Spain, has been to transform a lecture-based instruction system with little emphasis on active learning into a system more similar to the United States in which labs, sections, applied learning, and thus problem sets are heavily emphasized. Given the magnitude of this reform, the scarcity of empirical studies quantifying the potential learning gains from the teaching methods favored by the ESHE is surprising.

Our method of estimation is close to the experimental designs used in Grove and Wasserman (2006), Grodner and Rupp (2011), and Geide-Stevenson (2009). In our study, the estimation strategy also relies on a carefully designed experiment. We exploit variation arising from observing the performance of nearly 300 students taking the same class during the same semester and with the same instructors. A major advantage, however, is that due to the design of the experiment, the identification of the causal effect comes from within-student variation. The students in the control group are *exactly* the same ones as in the treatment group, which eliminates most selection concerns. In particular, academic performance is measured through a long multiple choice final test in which some questions are related to graded problem sets and others are related to non-graded problem sets given throughout the semester. In addition, the identification strategy is further refined by the fact that the problem sets that are required for credit are different in each section, which allows us to control for differences in the difficulty of the exam questions. After controlling for potential remaining biases and selection concerns, our results show that graded problem sets increase test scores by 8 percentage points, or close to a letter grade. We also found that the result is particularly strong among weaker students.

The remainder of the article is organized as follows. We describe the research design and data in the next section and explain the econometric methods in the subsequent section. The next two sections are devoted to the results, followed by the concluding section.

RESEARCH DESIGN AND DATA

The study was conducted at Complutense University of Madrid, Spain, during the fall of 2009. Complutense University is the largest public university in Spain. Approximately 80,000 students were enrolled during the academic year 2009–10.

In order to adjust to the needs of the ESHE, Complutense University undertook a major reform of its academic offering during 2009 and 2010. New curricula were developed in all areas of study. The new curricula were officially launched in the fall of 2010.¹ The reform was supposed to transform a system based on large classrooms, in which lectures were the main teaching method,

into another system in which smaller classes would be the norm. The new curricula significantly reduced the amount of professors' lectures and drastically increased the number of hours of instruction through sections, labs, and applied work. Prior to the reform, grades were mostly determined by the results obtained by students in only one final exam a year (to be taken in June or September). On the other hand, instructors under the new system can use many more methods of evaluation due to the reduction of class size and the increase in the number of lab or section hours. As a consequence, the use of graded problem sets has increased significantly since the implementation of the ESHE.

During the years prior to the launching of the reform, the European Union gave funding to some universities to conduct research studies measuring the potential impact of the changes favored by the ESHE. Complutense University received some of these funds and organized several competitive calls to distribute the funding. We describe the results of a research experiment that was awarded funding in 2009 in this article.

The classes in which the experiment was carried out took place at the Law School during the year prior to the introduction of the new curriculum. The Law School at Complutense offers both undergraduate and graduate degrees. This study was focused on the undergraduate degree. Unlike the higher education systems of other countries (e.g., the United States), the curriculum of undergraduate studies at Complutense in 2009 offered little flexibility to students in choosing their courses. In order to obtain the degree, all the students had to take exactly the same 25 yearly courses. Both having yearly courses instead of semesters and having a fixed curriculum were common across many universities and undergraduate degrees prior to the development of the ESHE. The law curriculum at Complutense was designed in 1953, and it is known as "Plan 53." At the time of this study, the plan had been in place for 56 years.

Economics courses comprise 2 out of the 25 courses of "Plan 53." The first of these classes is Principles of Economics. Students take this class during the second year of their five-year program. The course is equivalent to a semester of principles of microeconomics and a semester of principles of macroeconomics at American universities. The level of instruction is similar to that of Mankiw's (2011) *Principles of Economics* textbook. Mankiw's book is indeed widely used among instructors teaching this class.

The second economics course that students are required to take is "Public Finance." Students enroll in this class during their third year of studies. The class is a typical public economics course divided into the equivalent of a semester of public expenditure and a semester of taxation. The class is taught at the level of Rosen's (2004) *Public Finance* book, which is also a widely used text among instructors teaching this class.

The research experiment described in this article was conducted during the public expenditure part of the Public Finance course offered in 2009. The total number of students who enrolled and completed the class was 316. The class was offered at five different times, three days a week. Three times were offered in the morning (from 8:30 a.m. to 11:30 a.m.) and two in the afternoon (from 5:30 p.m. to 7:30 p.m.). Two professors taught the five classes. One of the professors was in charge of the morning classes while the other was responsible for the afternoon classes. At the time of signing up for the class, students chose whether they wanted to take the class in the morning or in the afternoon, but they could not decide in which specific group. The group was assigned to the student by the university, and students could not change this assignment. The assignment is alphabetical. This means that students whose last names start with the same letter are allocated to the same class. We can consider this allocation as random, as there is

no evidence of any correlation between last name initials and academic performance. However, the distribution of students between morning and afternoon classes is not random because it depends on students' choices. This is a concern for estimation purposes if students with different ability or effort levels self-select disproportionately into either morning or afternoon groups. The summary statistics of table 1 show that this might be the case because afternoon students are, on average, older, obtained lower grades on the university entrance exam, and are more likely to work part-time.²

The total number of students for which we were able to obtain data on all the relevant variables used in the analysis was 289. The distribution of students among the different groups was 67, 69, and 55 in the three morning classes and 60 and 38 in the afternoon classes.

On the first day of class, instructors gave students the syllabus and explained the grading system and the organization of the class. Both instructors followed the same textbook and grading system. The course grade was determined by the grade obtained on the final exam (90 percent) and the graded problem sets (10 percent). Four problem sets were distributed throughout the semester.

TABLE 1
Summary Statistics: Mean (Standard Deviation)

	All Students	Morning	Afternoon	Morning Group 1	Morning Group 2	Morning Group 3	Afternoon Group 4	Afternoon Group 5
Overall grade	58.76 (21.06)	62.70 (21.02)	49.94 (18.21)	65.48 (19.24)	63.72 (21.70)	57.67 (21.84)	51.78 (17.22)	48.88 (18.95)
Entrance exam	6.32 (1.01)	6.43 (1.04)	6.08 (0.88)	6.39 (1.02)	6.52 (1.01)	6.34 (1.11)	6.07 (0.94)	6.08 (0.86)
Times enrolled	1.81 (1.32)	1.56 (1.02)	2.37 (1.69)	1.50 (0.87)	1.47 (0.77)	1.76 (1.41)	2.65 (1.84)	2.21 (1.60)
Year entrance exam	2005.40 (2.72)	2005.92 (2.15)	2004.15 (3.40)	2006.68 (1.43)	2006.01 (1.89)	2005.62 (3.05)	2003.90 (3.44)	2004.30 (3.43)
Collective decision making	48.75 (29.68)	55.55 (28.30)	33.31 (26.94)	57.57 (25.97)	57.24 (28.52)	50.77 (30.63)	30.75 (28.25)	34.77 (26.30)
Public goods	68.07 (30.27)	69.13 (30.13)	65.66 (30.61)	72.66 (28.40)	70.13 (28.51)	63.65 (33.84)	66.56 (28.93)	65.14 (31.78)
Externalities	67.09 (29.05)	68.15 (29.20)	64.69 (28.73)	70.40 (29.19)	69.32 (28.27)	63.75 (30.47)	70.37 (27.77)	61.44 (29.01)
Monopolies	54.36 (33.76)	60.17 (31.97)	41.87 (34.46)	64.28 (31.00)	61.27 (29.36)	53.56 (35.82)	49.74 (32.12)	37.37 (35.21)
Treated questions	60.01 (30.22)	57.86 (30.25)	65.17 (29.61)	60.95 (28.70)	59.25 (28.92)	57.86 (33.21)	68.46 (28.20)	63.29 (30.35)
Control questions	59.18 (33.30)	65.87 (29.64)	37.59 (31.13)	71.53 (28.71)	69.72 (28.30)	63.56 (32.05)	40.24 (31.49)	36.07 (30.97)
Gender (% of females)	57.05 (49.52)	61.92 (48.77)	47.73 (50.01)	62.85 (48.66)	53.33 (50.22)	69.64 (46.39)	53.12 (50.70)	44.64 (50.16)
Number of observations	289	191	98	67	69	55	38	60

Notes: Standard deviations are in parentheses. The numbers in bold mean that the block corresponds to the treated questions. *Collective decision making* and *Monopolies* are Treated questions for morning students and Control questions for afternoon students. *Public goods* and *Externalities* are Control questions for morning students and Treated questions for afternoon students.

Each of them referred to one of the four main blocks in which the course was structured. The first problem set referred to collective decision-making and voting rules, the second one referred to public goods, the third one dealt with externalities, and the fourth one with monopolies and regulation. In each class, only two of the problem sets had to be turned in and were used to compute the corresponding 10 percent of the final grade. The other two problem sets did not have to be turned in. The four problem sets were exactly the same for all students, but the ones that had to be turned in and graded were different in the morning and in the afternoon classes. The specific problem sets that would count for a grade in the morning and afternoon groups were randomly decided by instructors at the beginning of the experiment. Instructors heavily emphasized that the four problem sets were equally important for the final exam regardless of whether they would be graded or not. To make this point clearer, they made final exams from previous years available to students.

The final exam consisted of 40 multiple choice questions divided into five blocks of approximately equal numbers of questions. The first block of questions was a “theory block.” In this block, students had to show knowledge of definitions and concepts explained during the class, but not directly related to the material covered in the problem sets. Each of the other four blocks of multiple choice questions was directly related to one of the four problem sets. All the 289 students took the same final exam in the same classroom during finals week. For grading purposes, each correct answer was given one point, incorrect answers were penalized with -0.333 point, and blank answers did not add or subtract points. The grading scheme was clearly explained in the first page of the exam and also by the instructors during the exam.

ECONOMETRIC ISSUES

The class organization and experimental set up just described is similar to that of Grove and Wasserman (2006). An important difference, however, is that in our study each of the students in the class is, at the same time, part of both the treatment and the control group, while in Grove and Wasserman each student belonged to either the control or the treatment group. In their study, the identification of the causal effect of having graded problem sets came from comparing two very similar groups of students and observing the differences in performance in the treated group versus the control group. In our case, we were able to observe the performance of the same student under the treatment (questions for which problem sets were required) and without the treatment (questions for which problem sets were not required). Identification arises from comparing the performance of the same student on the questions subject to the treatment and the questions for which the student did not receive the treatment. While comparing the performance of the same student with and without the treatment is close to an ideal situation for the purposes of causal identification, several potential threats must be addressed before we could be certain that we were capturing a causal effect.

The most important concern was related to the difficulty of each of the problem sets. Ideally, the perfect experimental design would consist of comparing the performance of the same student on exactly the same questions on the exam, both with and without the treatment. This cannot be done, because for any given student, a problem set cannot be both required and not required at the same time. The problem then is that some problem sets are harder than others. If this is the case, just comparing the performance of students on the questions related to graded and non-graded

problem sets would not capture the causal effect of requiring and grading problem sets. A higher average score on the graded problem set questions could possibly mean that those questions were simply easier than the questions on the test that referred to non-graded problem sets. Similarly, a lower average score on the treatment questions could just imply that those questions were harder. Our experimental design allowed us to take this into account because the required problem sets were different in each of the classes. This means that each specific question on the exam was answered both by students for which the corresponding problem set was required and graded and students for which it was not. In addition, it is worth noting that the write-up of the problem sets and the decision on which problem sets each group would have to turn in was made at the beginning of the term prior to observing student characteristics in each group, so that the assignment of each group into the treatment was uncorrelated with student characteristics.

A second concern was that morning and afternoon classes were taught by different instructors. We knew from previous research that unobserved teacher characteristics have an impact on academic performance of students (e.g., Rockoff 2004; Rivkin, Hanushek, and Kain 2005). If the two instructors who participated in the experiment possess different unobservable characteristics that affect the performance of the students on the questions related to graded problem sets, it is possible that students taking the class with one of the instructors might perform better on some blocks of the exam for reasons other than just the assignment of the problem set. For example, if one of the instructors is particularly good at explaining public goods and this part of the material is not included in one of the required problem sets in this class, finding no difference between the performance of the students in this block compared to the graded problem sets blocks of the exam might just mean that the teacher effect is counterbalancing the problem set effect.

A third concern was potential peer effects. Students took the class during the same semester and with the same instructor but at different times and with different peers. In addition, student composition in the morning classes and in the afternoon classes differed, with a higher rate of part-time students and students with lower entrance exam grades in the afternoon class. Previous literature suggests that working part-time is likely to have a detrimental effect on the academic success of students (e.g., Stinebrickner and Stinebrickner 2003). It is conceivable, therefore, that requiring and grading problem sets may have differential effects on student performance due to the different group dynamics. If students in one group are very cooperative, these students may be performing better on certain parts of the exam just because they benefited from their peers' help throughout the term. That is, two otherwise identical students taking the same course with the same instructor and with the same requirements, but with different peers, could potentially perform differently. Additionally, it is also possible that the group dynamics in some of the groups are better, favoring also the performance of the instructor during the lectures in the better groups. As group dynamics and instructor performance are likely to change during the term and thus during the presentation of different parts of the course, a better performance on the problem-set-required parts of the exam may be then explained, at least in part, by these peer and time effects.

The concerns just outlined prevented us from being able to use a simple mean comparison to identify the effect of interest. Each of these concerns, however, could be addressed econometrically in a regression framework by controlling for the relative difficulty of each problem set, and for teacher and group effects. In order to do so, we first rearranged the data into student/block observations so that for each student we had four data points, one referring to each of the four

blocks of the exam. We then estimated a student fixed effects-model of the form:

$$\text{Points}_{ij} = b_0 + b_1 \text{Treatment}_{ij} + b_2 \text{Block}_j + x_i + e_{ij}, \quad (1)$$

where *Points* is the number of points (as a percentage) obtained by student *i* in block *j* of the exam; *Treatment* is a dummy that takes value 1 if the block refers to a graded problem set; *Block* is a set of four problem set dummies that capture invariant differences among each block of exam questions, such as their inherent difficulty; and x_i is the set of student fixed-effects dummies that controls for student invariant characteristics. Note that teacher and group are invariant within student, so the student fixed-effects dummies control for these factors as well as for other student invariant characteristics such as their relative ability level. In this model, therefore, we are exploiting the variation that arises from observing the same student answering different blocks of questions, holding the difficulty level constant across blocks.

A potential problem with the model of equation 1 is that the fixed-effects specification does not control for the fact that students chose whether to be in the morning or afternoon classes. If the reasons that led students to choose either morning or afternoon are correlated with how they perform in the class and how the treatment impacts them, the estimates of equation 1 could be biased. As mentioned before, this was a concern because the summary statistics of our sample showed that students in the morning classes were younger, less likely to work part-time, and did better on both the entrance exam and in the class (see table 1). An alternative to the model in equation 1 that solves this problem is to estimate a regression without fixed effects but controlling specifically for teacher effects, group effects, student effort and ability level, and other demographic characteristics such as gender or age. In this case, the corresponding regression would be:

$$\begin{aligned} \text{Points}_{ij} = & b_0 + b_1 \text{Treatment} + b_2 \text{Block}_j + b_3 \text{Teacher}_i + b_4 \text{Group}_i + b_5 \text{EntranceExam}_i \\ & + b_6 \text{TimesEnrolled} + b_7 \text{YearEntranceExam}_i + b_8 \text{Female} + e_{ij}, \end{aligned} \quad (2)$$

where, *Points*, *Treatment*, and *Block* are defined as in equation 1; *Teacher* is a dummy that takes the value 1 for classes taught by one of the instructors and 0 otherwise; *Group* is a set of dummy variables corresponding to each of the five groups of students; *EntranceExam* is the grade obtained by the student in the nationwide general admission test that students have to take prior to be admitted into college (called *selectividad*); *TimesEnrolled* is the number of times that the students have pre-registered for the class before actually taking it, and is a proxy for some student unobserved characteristics such as effort or part-time work;³ *YearEntranceExam* is the year in which the student took the entrance exam and controls for age; and *Female* takes value 1 for female students and 0 otherwise. This specification does not control for all unobserved within-student characteristics, as the fixed-effects specification does. It is, therefore, more likely to suffer from an omitted variable bias. This problem is alleviated by the inclusion of *EntranceExam*, *TimesEnrolled*, *YearEntranceExam*, and *Female*. Provided that the omitted variable bias is small, the specification in equation 2 allowed us to identify teacher and peer effects separately as well as to quantify the effect of some observable student characteristics that would otherwise be subsumed into the fixed effects term. In addition, as explained before, the specification in equation 2 allowed us to explicitly control for systematic differences in ability between morning and afternoon students that are not captured by the fixed-effects regression and avoided, therefore, the problem of weaker or stronger students selecting themselves into either

morning or afternoon classes due to reasons not captured by the fixed-effects term. In the results section below, we discuss the estimates of both equations 1 and 2.

The coefficient of interest in both equations is b_1 , which measures the differential effect of requiring and grading problem sets on student performance. Positive and significant estimates of b_1 would support the claim that requiring and grading problem sets leads to an improvement of student performance. The key identifying assumption of the model is that b_1 would have been zero if none of the problem sets had been required and counted towards the grade, or if all of them had been required. If this is the case, then an unbiased estimate of b_1 would be capturing the causal effect of requiring problem sets.

Looking at within-student variation holding difficulty levels constant excludes some of the validity threats that would question our key identifying assumption. A remaining selection concern might be, however, that students prepare more for the parts of the exam that are related to graded problem sets, not because having to turn in the problem sets and receiving feedback through grades help them understand the material better, but because requiring problem sets sends a signal that those parts of the material are more important for the exam. This kind of strategic reasoning would lead to some students prioritizing the “treatment” part of the material for reasons other than the treatment itself. This selection concern is likely to be small or inexistent due to three facts. First, instructors emphasized several times during the term that the exam would consist of an even number of questions from each problem set. Secondly, students had access to exams from previous years, so they could verify that all four problem sets were equally weighted on the exam. Third, even if students thought that requiring a problem set might be correlated with the content of the exam, it is not clear what the sign of the correlation would be. Some students may think that requiring the problem set signals that the material is more likely to be on the exam, but some other students may think that, as they have already been tested on those parts of the course, the exam would not emphasize them. For all these reasons, we believe that our identification strategy is free from this kind of selection concern.

RESULTS

Table 1 provides the summary statistics for the relevant variables used in the analysis. The average overall grade in the class for the students is 58.76 points out of 100. The result may seem low for people not accustomed to the Spanish higher education system. Most Spanish universities, and Complutense in particular, are large public universities where admission requirements are relatively low even in prestigious schools. This means that in the same class there are very strong students and relatively weak students. This is different from, for example, most American universities in which strict admission procedures result in the student bodies of each university being much more homogeneous. For example, “weak” students at Harvard University are still strong students, while at Complutense, Harvard-type students coexist in the same classroom with second-tier students. This heterogeneity also implies a big dispersion in grades and relatively higher failing rates in each class. An average grade of 58 points is in line with historical rates of success in the department and can be considered as normal.

Turning to the difference between morning and afternoon students, table 1 shows that morning students perform, on average, better than afternoon students. This can be observed by looking at the overall performance variable (e.g., the mean of the *Overall Grade* variable is 62.7 for morning

students and 49.94 for afternoon students). If we look at the average performance of each of the five groups of students, the differences still exist. Every morning group performs better than the best of the afternoon groups. Morning students are also stronger on average according to the *EntranceExam* grade. They are also younger (e.g., took the entrance exam more recently), and have pre-registered fewer times for the class before completing it. These differences are consistent with the fact that a higher rate of part-time students enroll in the afternoon classes and with the results of Stinebrickner and Stinebrickner (2003), who argued that working part-time had net detrimental effects on academic performance. Teacher and peer effects could also explain, at least partially, some of the differences in performance in the class. The fact that the student composition and performance differs in morning and afternoon classes supports the use of controls and within-student variation to avoid selection concerns.

The summary statistics of table 1 also indicate that there seem to be important differences in the difficulty across the different blocks of questions. Both morning and afternoon students in all five groups performed significantly better on the questions related to public goods and externalities compared to the other two blocks of problem-set questions. This is true regardless of whether the block was treated or not. The summary statistics, however, anticipate that requiring and grading the problem sets may have an impact on students' learning. The difference in student performance between the two relatively easier blocks of questions and the two harder ones is as high as 27.58 percentage points for the groups in which public goods and externalities were treated but is only 8 percentage points for the groups in which these two blocks were not treated (the numbers are obtained by subtracting the rows labeled *Treated questions* and *Control questions* in table 1).

Overall, the summary statistics show that student ability, teacher, peer, and question difficulty effects are likely present. For this reason, a simple mean comparison between the performance of students on the treatment and on the control questions is not meaningful for the purposes of identification.

Table 2 shows the results of the estimation of two sets of models which control for student ability, difficulty level, teacher, and peer effects. The first column corresponds to the fixed-effect model described above in equation 1. The dependent variable in this model is the grade obtained by the student in each block of questions.⁴ Among the covariates, the fixed-effects term controls for student-specific variation such as student ability, teacher and peer effects, and the exam block dummies control for the difficulty of the problem set. The variable of interest is *Treatment*. The estimate of the coefficient of this variable is 0.084 and is significant at the 1-percent level. According to this coefficient, requiring and grading problem sets has a positive impact on student performance on the final exam of approximately 8.4 percentage points, or almost a letter grade, *caeteris paribus*.

The ordinary least squares specifications of columns 2 to 5 in table 2 confirm the same result. Column 2 shows the estimates of the model without any control variables except for the treatment variable. Columns 3 to 5 add different control variables. These specifications show that once we controlled for the difficulty level of each of the blocks of questions, the coefficient of the treatment variable was very stable and was always significant at the 1-percent level. The magnitude of the coefficient in these regressions is identical to the one in the fixed effect (FE) specification of column 1. Thus, the causal effect of requiring and grading problem sets seems to be slightly less than a full letter grade.

According to the results in table 2, student ability and effort level, as measured by *EntranceExam*, is an important determinant of student achievement. The *R*-square of the regression

TABLE 2
Determinants of Exam Block Grade as a percentage

	1 FE	2 OLS	3 OLS	4 OLS	5 OLS
Treatment	0.084*** (0.016)	0.009 (0.017)	0.084*** (0.016)	0.084*** (0.016)	0.084*** (0.016)
Public goods	0.225*** (0.023)		0.226*** (0.023)	0.226*** (0.023)	0.226*** (0.023)
Externalities	0.215*** (0.021)		0.216*** (0.021)	0.216*** (0.021)	0.216*** (0.022)
Monopolies	0.057** (0.024)		0.058** (0.024)	0.058** (0.024)	0.058** (0.024)
Entrance exam				0.081*** (0.010)	0.070*** (0.010)
Year entrance exam					-0.005 (0.005)
Times enrolled					-0.066*** (0.025)
Female					-0.004 (0.024)
Teacher					0.135*** (0.036)
Group 1					-0.029 (0.030)
Group 2					-0.079** (0.036)
Group 5					0.054 (0.042)
Constant	0.430*** (0.018)	0.592*** (0.017)	0.429*** (0.021)	-0.081 (0.068)	9.824 (9.581)
Observations	1,156	1,156	1,156	1,156	1,156
R-squared	0.150	0.000	0.080	0.150	0.180

Note: Clustered (by student) standard errors in parentheses.

*Significant at 10%; **significant at 5%; ***significant at 1%.

increases from 0.08 to 0.15 when this variable is included. The coefficient of this variable is significant at the 1-percent level in all the specifications, and has a value of 0.07 in the most inclusive of them (last column of table 2). A 10-percent increase in performance on the entrance exam is associated with an average increase of 7 percent on the performance in each block of the exam. Other variables related with student ability and effort such as *TimesEnrolled* and *Year-EntranceExam* behave as expected, although only *TimesEnrolled* is significant. According to the coefficient of *TimesEnrolled*, students who signed up for the class in previous years but dropped it early in the term perform worse in the class compared to students who have never before dropped the class.

Unobserved teacher effects are also present. Column 5 shows that the teacher effects variable is sizable and significant, although a word of caution is needed when interpreting this coefficient. While differences in student ability and effort level in morning and afternoon classes are captured

by the *EntranceExam*, *YearEntranceExam*, and *TimesEnrolled* variables, the coefficient of the *Teacher* variable may be capturing, in part, some unobserved student characteristics inadequately controlled for by the variables that measure student ability.

The group dummies, on the other hand, are not significant except for one group. This is consistent with the idea that the assignment of students to group by the university is random. The gender dummy of column 5 is not significant either, pointing to no differences in class performance related to gender differences. The literature on gender differences in student learning suggests that men usually outperform women in economic classes and that this difference in learning develops mainly during adolescence (Siegfried 1979; Heath 1989). We found no such differences in our sample. This result may not be representative of potential learning differences among the whole population of students if students selected into the Law School share a similar set of analytical abilities that are different from those of the whole population of students (Heath 1989).

EXTENSIONS

Previous literature suggests that students with different academic aptitudes may benefit differently from graded assignments. According to this reasoning, better students would be less affected by problem sets because they need fewer incentives to study during the term and less feedback from instructors to do well in the class. On the other hand, the continuous effort that problem sets require and the feedback provided by grading would benefit weaker students more by creating an incentive to work continuously throughout the term. Grove and Wasserman (2006) found no evidence of such a differential effect. They reached this conclusion by estimating a model in which the treatment variable was interacted with student GPA. This coefficient is small and non-significant in their study, which led them to conclude that there were no differential effects across students. Such differences may exist, however, in a group of more heterogeneous students such as the ones at Complutense. In addition, it is also possible that the effects across different groups of students are nonlinear, in which case simply adding an interaction would not fully capture differences across groups of students.

In the first column of table 3, we show the estimates of the same models as in table 2, but add an interaction term. We use performance on the entrance exam as a proxy for student ability. Contrary to Grove and Wasserman (2006), this term is significant and relevant in magnitude, showing evidence of a differential effect of the treatment across student ability. The interaction term is negative and has a value of -0.042 . This means that better students (students who obtained a higher grade on the entrance exam) benefit less from having graded problem sets compared to weaker students. For a student who barely passed the entrance exam (e.g., 5 out of 10 in the entrance exam), the effect of the treatment is 13 percentage points, or more than a full letter grade.

To account for the possibility of nonlinear effects, columns 3 to 8 in table 3 show the estimates of the FE and OLS models when the regressions are run separately for three different sub-samples of students. Students in the first sub-sample are those with lower academic ability (those that score in the lower 33rd percentile on the entrance exam). The second sub-sample includes students between the 33rd and 66th percentile, and the third includes only the top-performing

TABLE 3
Determinants of Exam Block Grade as a Percentage, by Student Ability

	1 OLS All	2 FE Lower	3 FE Medium	4 FE Top	5 OLS Lower	6 OLS Medium	7 OLS Top
Treatment	0.346*** (0.081)	0.139*** (0.026)	0.060** (0.030)	0.061** (0.030)	0.139*** (0.027)	0.060* (0.030)	0.061** (0.030)
Public goods	0.219*** (0.023)	0.212*** (0.039)	0.131*** (0.041)	0.318*** (0.037)	0.212*** (0.039)	0.133*** (0.041)	0.318*** (0.038)
Externalities	0.209*** (0.022)	0.185*** (0.038)	0.171*** (0.039)	0.275*** (0.036)	0.185*** (0.038)	0.173*** (0.039)	0.275*** (0.036)
Monopolies	0.058** (0.024)	0.041 (0.038)	0.020 (0.045)	0.111*** (0.040)	0.041 (0.038)	0.022 (0.046)	0.111*** (0.041)
Entrance exam	0.091*** (0.013)				0.127 (0.108)	0.017 (0.090)	0.049** (0.023)
Year entrance exam	−0.005 (0.005)				−0.011 (0.008)	−0.002 (0.007)	0.011 (0.007)
Times enrolled	−0.066*** (0.025)				−0.046 (0.045)	−0.048 (0.059)	−0.100** (0.0451)
Female	−0.004 (0.024)				−0.002 (0.042)	0.020 (0.048)	−0.035 (0.032)
Teacher	0.135*** (0.036)				0.002 (0.060)	0.105 (0.088)	0.104** (0.047)
Group 1	−0.029 (0.030)				−0.002 (0.057)	−0.046 (0.062)	−0.031 (0.037)
Group 2	−0.019** (0.036)				−0.008 (0.058)	−0.088 (0.075)	−0.121** (0.054)
Group 5	0.054 (0.042)				−0.106 (0.066)	−0.014 (0.079)	−0.021 (0.060)
Treatment*Entrance exam	−0.042*** (0.012)						
Constant	9.694 (9.587)	0.327*** (0.027)	0.473*** (0.033)	0.491*** (0.033)	22.179 (15.692)	3.675 (13.041)	−22.574 (14.152)
Observations	1156	384	383	388	384	383	388
R-squared	0.190	0.180	0.080	0.270	0.140	0.080	0.280

Note: Clustered (by student), standard errors in parentheses.

*significant at 10%; **significant at 5%; ***significant at 1%.

students. The coefficient of the *Treatment* variable in these three regressions indicates that top students benefited comparatively less from the assignment and grading of problem sets. Both FE and OLS models yield similar results. Top and middle students increased their performances by approximately half a letter grade due to the treatment, while students in the bottom range of the distribution increased their performance by more than a full letter grade. Overall, the results of table 3 suggest the presence of a significant differential effect of grading problem sets on student performance, with weaker students benefiting more.

CONCLUSION

The goal of many education policies is to improve students' academic achievement. Educational research has shown that students' achievement improves with good teachers. Few policy implications have been derived from this fact, however, because it is difficult to identify, *a priori*, what characteristics make one teacher better than others.

In this article, we have estimated the degree to which requiring and grading problem sets may lead to an improvement of student performance. Our results show that the use of this technique increases student performance in our sample by close to a letter grade. We also found that this improvement is not equally distributed across students. Weaker students benefit relatively more compared to better students.

The main policy implication of the results is that the use of graded problem sets appears to be a powerful technique to increase students' learning. As the ESHE promotes and facilitates the use of frequent graded assignments as a pedagogical technique, large learning gains should be expected, *caeteris paribus*. Our results also suggest that some objective characteristics of the teaching style can indeed be identified as having potential to improve teaching performance. To the extent that we can train instructors in the use of these techniques, we could achieve a substantial improvement in student learning outcomes.

In this article, the effect of the use of graded problem sets on student performance has been identified using data from an experiment. The main advantage of such a design is that it allows attributing the improvement in test scores specifically to the use of problem sets and not to other potential confounding factors. Nevertheless, as with other experimental designs, the results of this analysis have limited generalizability because they refer to a specific economics course during a specific semester in a specific university and country. The quantitative and qualitative importance of the implications that can be derived from the results of this study can motivate further research in this field so that the results can be better generalized. A natural extension of this research would be to perform a similar analysis using data from different universities, countries, and fields of study.

NOTES

1. Many other Spanish universities offered new programs during the fall of 2010, as that was the deadline suggested by the Spanish government for all universities to adjust their academic offerings to the ESHE.
2. We were not able to obtain exact data on the number of part-time students. The University Registrar, however, did provide us with information regarding the number of times that a student had been pre-enrolled in the class. This variable is measured as the number of times that students signed up for the class without actually taking it. Some students drop the class during the semester without taking any of the exams. It is reasonable to expect that part-time students or students with time constraints are more likely to drop some their classes if they realize that they do not have enough time to complete all the requirements.
3. See note 4 for a more detailed description of this variable.
4. To study the robustness of the model to the use of a different dependent variable, we also estimated the different regressions using as the dependent variable the percentage of correct questions in each block, which does not penalize incorrect questions. The results are both qualitatively and quantitatively the same. These results are available upon request from the authors.

REFERENCES

- Aaronson, D., L. Barrow, and W. Sander. 2007. Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics* 25: 95–135.
- Betts, J. R., and J. Grogger. 2003. The impact of grading standards on student achievement, educational attainment, and entry-level earnings. *Economics of Education Review* 22(4): 343–52.
- Bonesrønning, H. 2004. Can effective teacher behavior be identified? *Economics of Education Review* 23: 237–47.
- Carrell, S. E., and J. E. West. 2010. Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy* 118: 409–31.
- Cooper, H., J. G. Robinson, and E. A. Patall. 2006. Does homework improve academic achievement? A synthesis of research, 1987–2003. *Review of Educational Research* 76: 1–62.
- Eren, O., and D. J. Henderson. 2008. The impact of homework on student achievement. *Econometrics Journal* 11: 326–48.
- . 2011. Are we wasting our children's time by giving them more homework? *Economics of Education Review* 30: 950–61.
- Falch, T., and M. Rønning. 2011. Homework assignment and student achievement in OECD countries. Working Paper Series No 5/2011. Trondheim, Norway: Norwegian University of Science and Technology, Department of Economics.
- Figlio, D. N., and M. E. Lucas. 2004. Do high grading standards affect student performance? *Journal of Public Economics* 88: 1815–34.
- Geide-Stevenson, D. 2009. Does collecting and grading homework assignments impact student achievement in introductory economics course? *Journal of Economics and Economic Education Research* 10(3): 3–14.
- Grodner, A., and N. G. Rupp. 2011. The role of homework in student learning outcomes: Evidence from a field experiment. ECU Working Paper ecu1001. Greenville, NC: East Carolina University, Department of Economics.
- Grove, W. A., and T. Wasserman. 2006. Incentives and student learning: A natural experiment with economics problem sets. *The American Economic Review* 96: 447–52.
- Hanushek, E. 1986. The economics of schooling. *Journal of Economic Literature* 24(3): 1141–77.
- Heath, J. 1989. An econometric model of the role of gender in economic education. *American Economic Review* 79: 226–30.
- Kane, T. J., J. E. Rockoff, and D. O. Staiger. 2008. What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review* 27: 615–31.
- Mankiw, N. G. 2011. *Principles of economics*. 6th ed. Cincinnati, OH: South-Western College Pub.
- Oettinger, G. S. 2002. The effect of nonlinear incentives on performance: Evidence from “ECON 101.” *Review of Economics and Statistics* 84: 509–17.
- Rivkin, S. G., E. A. Hanushek, and J. F. Kain. 2005. Teachers, schools, and academic achievement. *Econometrica* 73: 417–58.
- Rockoff, J. E. 2004. The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review* 94: 247–52.
- Rosen, H. S. 2004. *Public finance*. 7th ed. New York, NY: McGraw-Hill.
- Siegfried, J. J. 1979. Male-female differences in economic education: A survey. *Journal of Economic Education* 10: 1–10.
- Stinebrickner, R., and T. Stinebrickner. 2003. Working during school and academic performance. *Journal of Labor Economics* 21: 473–91.
- . 2008. The causal effect of studying on academic performance. *The B.E. Journal of Economic Analysis & Policy* 8, Article 14.

