barcelona gse
graduate school of economics

The Importance of Relative Performance Feedback Information: Evidence from a Natural Experiment using High School Students

Ghazala Azmat
Nagore Iriberri

March 2010

Barcelona Economics Working Paper Series

Working Paper nº 444

The Importance of Relative Performance Feedback Information: Evidence from a Natural Experiment using High School Students[*]

Ghazala Azmat[+] Nagore Iriberri[**]

Universitat Pompeu Fabra and Barcelona GSE

First Draft: January 2009
This Version: March 2010

Abstract

We study the effect of providing relative performance feedback information on performance, when individuals are rewarded according to their absolute performance. A natural experiment that took place in a high school offers an unusual opportunity to test this effect in a real-effort setting. For one year only, students received information that allowed them to know whether they were performing above (below) the class average as well as the distance from this average. We exploit a rich panel data set and find that the provision of this information led to an increase of 5% in students' grades. Moreover, the effect was significant for the whole distribution. However, once the information was removed, the effect disappeared. To rule out the concern that the effect may be artificially driven by teachers within the school, we verify our results using national level exams (externally graded) for the same students, and the effect remains.

Keywords: school performance, relative performance, absolute performance, feedback, natural experiment, social comparison, self-perception, competitive preferences.

JEL classification: I21, M52, C30.

[*] We are grateful to Oiartzo Ikastola (former Orereta Ikastola) and the University of the Basque Country for giving us access to the data. We also thank Manuel Arellano, Samuel Bentolila, Stephane Bonhomme, Antonio Cabrales, Antonio Ciccone, Vincent P. Crawford, Vicente Cuñat, Sergi Jimenez-Martin, Ignacio Palacios-Huerta, and Pedro Rey-Biel for their comments. Ghazala Azmat acknowledges financial support from ECO2008-06395-C05-01 and the support of the Barcelona GSE Research Network and of the Government of Catalonia. Nagore Iriberri acknowledges financial support from Fundación Rafael del Pino, Ministerio de Educación y Ciencia (SEJ2006-05455, SEJ2007-64340 and ECO2009-11213) and the support of the Barcelona GSE Research Network and of the Government of Catalonia.

[+] Ghazala Azmat. Departament d'Economia i Empresa. Universitat Pompeu Fabra, Ramón Trías Fargas 25-27, 08005 Barcelona (Spain). Tel: (+34) 935421757. E-mail: ghazala.azmat@upf.edu.

[**] Nagore Iriberri. Departament d'Economia i Empresa. Universitat Pompeu Fabra, Ramón Trías Fargas 25-27, 08005 Barcelona (Spain). Tel: (+34) 935422690. E-mail: nagore.iriberri@upf.edu.

1. Introduction

Improving students' performance has been an important concern for academics and educational policy makers alike. Given the recent introduction of the OECD coordinated *Programme for International Student Assessment* (PISA), improvements in students' performance, measured by their grades, is at the heart of governmental reform.[1] The education literature has focused on school inputs as the principal means to improve students' performance. In particular, by looking at the effects of reducing the pupil/teacher ratio, improved quality of teacher (experience and education), and extended term length (see Krueger, 1999, Card and Krueger, 1992). There is however, a lively debate regarding the effectiveness of school inputs, largely due to their associated costs (Hanushek, 1996 and 2003). Moreover, the PISA reports do not show a strong positive relationship between the amount spent per student and the performance in the standardized tests in mathematics, science and reading. For example, the US ranks second in expenditure per pupil (the cumulative expenditure on educational institutions up to age 15 is 91,770$) but ranked twenty-second (out of 30) in performance (see OECD PISA report, 2006).

More recently, there has been interest in analyzing the relevance of performance evaluations and *feedback information* regarding these evaluations. The effect of interim feedback information regarding own performance on subsequent performance has been studied mostly in labor settings.[2] Bandiera et al. (2008) study this effect empirically on students' performance. They find that by providing university students with interim feedback information regarding own performance has a positive effect on their final performance. However, feedback information involving *relative performance* has received less attention. The provision of relative performance feedback information allows for social comparison (individuals can evaluate their own performance by comparing themselves to others, Festinger, 1954). While this has been extensively studied in management and psychology literature (see Festinger, 1954, Locke and Latham, 1990, and Suls and Wheeler, 2000, for an overview), it has not been fully explored in economics.[3]

[1] For example, Germany is considering a complete revamp of their traditional education system, *Gymnasium*, in response to PISA reports (See Economist, Oct 17th 2008); in 2008, United Kingdom extended the compulsory school leaving age by one year; since 2001 the United States has implemented the *No Child Left Behind* Act.

[2] Many papers analyze the optimal provision of interim feedback information on own performance using a principal-agent model in a tournament setting (Aoyagi, 2007, and Ederer, 2010) and under piece-rate and flat-rate incentives (fixed-wage) (Lizzeri et al., 2002, and Ertac, 2006).

[3] The provision of relative performance feedback information has received attention mostly in the tournament literature. Gershkov and Perry (2009), Kräkel (2007) and Lai and Matros (2007) study the optimal provision of relative performance feedback information in tournaments. For empirical work see Casas-Arce and Martinez-Jerez (2009) and Young et al. (1993). Finally, for experimental work see Muller and Schotter (2003), Hannan et

In our paper, we investigate both theoretically and empirically the role that relative performance feedback information plays on students' performance. We use a natural experiment that took place in a high school, where for one year only, students were provided with relative performance feedback information in addition to the usual individual performance information. Typically, students received report cards containing the grades for each subject, where grades measured absolute performance since there was no grade curving (i.e., no tournament incentive scheme).[4] However, during the academic year 1990-1991, students also received in their report card their own average (over all subjects), as well as the class average (over all subjects and students), such that they could observe whether they were performing above or below the class average, as well as the distance from this average. The relative performance information based on the class average allowed for social comparison, since students could observe whether they were performing better or worse than their classmates. This information was removed after one academic year. The question we address is whether this additional information had any effect on students' real effort and, therefore, on their performance.

The importance of social comparison and relative performance feedback information has been studied under flat-rate incentives (fixed-wage) (Falk and Ichino, 2006, Mas and Moretti, 2009, and Kuhnen and Tymula, 2008). However, when individuals are rewarded according to their absolute performance, to our knowledge this is the first paper that addresses the effect of relative performance feedback information on performance in a *real effort* and *natural* setting, such as schooling. There are two experimental papers that pose also a similar question. Hannan et al. (2008) use an experiment (with no real effort) to compare the impact of providing relative performance feedback information to subjects under both, the tournament and piece-rate incentives. They find that for the subjects participating under piece-rate incentives, information regarding their position relative to the average, increases efforts for all subjects. Eriksson et al. (2009) on the other hand, use a two-person experiment (with real effort) to test the effect of providing the information about the other person's performance under both the tournament and piece-rate incentives. Under piece-rate incentives, they find no significant effect when the information is provided. However, they do find that the subject

al. (2008), Fehr and Ederer (2007) and Eriksson et al. (2009). Empirical work finds ambiguous results; while some authors find that the provision of relative performance feedback increases all participants' effort, others find that once the relative performance feedback is provided, the leading participants slack off and participants who are lagging behind give up.

[4] When individuals are rewarded according to their absolute performance, we could also say that they are rewarded according to a *piece-rate* scheme. Similarly, when individuals are rewarded according to their relative performance, they are rewarded according to a *tournament* scheme.

who is lagging behind makes significantly more mistakes. In addition, Blanes i Vidal and Nossol (2009) look at the effect of relative performance feedback information (whole ranking is made available) on individual performance in a labor setting and they find that employees work harder when this information is provided.[5]

The natural experiment that took place in a high school offers a unique opportunity to study the importance of relative performance information, when individuals are rewarded according to their absolute performance. There are many important features that should be highlighted. First, the experiment takes place in a natural setting and allows us to measure real effort through their grades. Second, the provision of the additional information took place for exogenous reasons in the academic year 1990-1991. In particular, the adoption of a new application to produce report cards offered the possibility of including the extra information and the administrative staff (not the teachers) decided to use it. It was untargeted, that is, it was not introduced as a response to any initiative to affect performance. Third, there is no systematic difference between the year 1990-1991 and any other year in terms of class-sizes, number of teachers, subjects taught and/or the evaluation system. Fourth, we have panel data on 1,313 students (3,414 grades) registered at the high school between the years 1986 and 1994. Typically students would complete four years in high school before going to University. Since we can follow the same students over time, we are able to control for individual fixed effects. Finally, the additional information was removed from the report cards after one year, allowing us to exploit the variation in students' performance before, during and after the treatment.

We consider two alternative explanations for why students would react to the relative performance information and the empirical analysis allows us to test the relevance of each explanation.

On the one hand, students might react to the additional information because individuals have inherently competitive preferences or that the provision of the relative performance information stimulates this type of preferences.[6] Given the existence of competitive preferences, when information that allows for social comparison is provided, that is, relative performance feedback information is provided, people get utility (disutility) from being ahead

[5] Recently, Bandiera et al. (2009) and Delfgaauw et al. (2009) have also considered the impact of relative performance feedback at the team level.

[6] There is extensive work on preferences that include social comparison, such as negative interdependent preferences (Kandel and Lazear's (1992), a type included in Charness and Rabin (2002), Ok and Kockesen (2000)), and preferences over relative income (Duesenberry, 1949, Easterlin, 1974, Layard, 1980, Frank, 1984, 1985, Clark and Oswald, 1996, Hopkins and Kornienko, 2004, Dubey and Geanakoplos, 2004 and 2005, and Moldovanu et al., 2005). Note that competitive preferences are different from preferences that show *inequity aversion* (Fehr and Schmitd, 1999, and Bolton and Ockenfels, 2000). This is further explained in Section 3.

(behind) of others. There are two important features to note. First, there is no explicit reward (penalty) derived from being above (below) the class average. Unlike in a tournament, students are not explicitly rewarded according to their relative performance but according to absolute performance (see footnote 4).[7] Second, the relative performance information is private information such that it is different from status-seeking preferences. We will refer to this theory as the preferences-based or competitiveness theory. We show that based on this explanation, when the relative performance information is provided, *all* students would choose higher effort, and therefore higher performance would be observed.

On the other hand, students might react to the additional information because individuals have an imperfect knowledge of their own ability, such that the additional information is informative of one's own ability. Moreover, if performance is a function of both, ability and effort, where ability and effort are complements in performance, then the self-perceived ability will affect the optimal choice of effort. Relative performance feedback affects the self-perceived ability and, therefore, it affects the choice of effort. We will refer to this theory as the self-perception theory. We show that based on this explanation, top (bottom) performing students would choose higher (lower) effort, because this information encourages high ability (discourages low ability) students.

We find that the feedback information on relative performance had a strong positive effect on students' performance. Overall, we find a 5% increase in their grades. This is comparable, if not better than the effects found by the literature on improving school inputs. For example, Krueger (1999) uses an experimental study (the Tennessee Student/Teacher Achievement Ratio (STAR)), where students were randomly assigned to small or regular size classes, to show that reducing class size from 22 to 15 leads to internal rate of return of 6%. More importantly, contrary to improving school inputs, providing feedback information on relative performance involves no additional cost. Moreover, this positive effect is significant throughout the grade distribution, where the strongest effects are found at the tails of the distribution. This supports the competitive preferences rather than the self-perception hypothesis, since we do not observe any discouragement effect on the bottom performing students. It has also important policy implications, as it implies that the students at the top as well as bottom of the distribution react positively. In addition, we find that when the relative performance feedback information is removed, the effect disappears, such that there is no lasting effect of the treatment.

[7] There is an extensive literature on tournaments and contests as optimal contracts (See Prendergast, 1999, for a literary review).

A more detailed analysis shows that there are heterogeneous effects from the provision of relative performance feedback information. First, the effect is significant only for students in the first and fourth years of high school. It is reasonable to assume that in the first year of high school the relative performance feedback information provides *new* information and that students might be more reactive to it than in the subsequent years. As for the fourth year effect, given that the grades during this year are especially important in determining the final university entry grade, one may believe that any additional information regarding grades will provoke a stronger reaction. Second, the positive effect is strongest in science subjects such as Mathematics, as well as in language subjects. This has important policy implications since there has been a special interest in improving grades in more technical subjects such as Math. Third, we also find that, although girls overall obtain better grades than boys there is no significant gender difference in the reaction to the relative performance feedback information.[8]

One may question whether the positive effect is driven by students, parents or teachers. Report cards have to be signed by parents, such that the additional information is provided to both parents and students. It is, therefore, impossible to disentangle whether the effect is coming from students or their parents. In the rest of the paper, both theoretical and empirical parts, to avoid repetition we will only refer to students. We are, however, able to rule out the possibility that teachers are artificially driving the effect. We use an external source of variation coming from national level exams, *Selectividad*, (similar to the Scholastic Aptitude Tests (SAT) used in the United States), completed at the end of the fourth year of high school. Selectividad differs from SAT in that it tests the knowledge on the topics covered during the last year of high school, such that effort and performance in this year should be highly correlated with the performance on the Selectividad test. The similarity is that both exams are written and graded by external bodies, such that the teachers in the school have no way to affect these grades. We replicate the analysis using the grades from the national level exams and find the same positive and strong effect.

We find additional evidence for our results in a companion paper that replicates this experiment in a controlled environment such as the laboratory. In Azmat and Iriberri (2010), we conduct an experiment with real effort, where treated subjects are informed about their own performance, as well as the group average performance, while the non-treated are only given feedback on own performance. In line with our findings in this paper, the provision of

[8] This finding is consistent with the Niederle and Yestrumskas (2008) finding that women and men do not differ in their preferences over receiving relative performance feedback information.

the relative performance feedback information leads to an increase on performance although this increase slows down over time. These findings suggest that the effect has external validity. In addition, they provide evidence that the change in behavior is, at least in part, coming from self motivation (students) and not solely from parents and/or teachers.

The paper is organized as follows. Section 2 describes the natural experiment in detail. Section 3 derives theoretical predictions for why and how we would expect students to react to the additional information. Section 4 describes the data and presents the main descriptive statistics. Section 5 presents the results from the empirical analysis. As well as identifying and quantifying the treatment effect, we thoroughly investigate the impact of the information treatment. Finally, we conclude in Section 6. The web appendix provides further details.[9]

2. Description of Natural Experiment

The natural experiment took place in a high school located in the province of Gipuzkoa in the north of Spain (the Basque Country) during the academic year 1990-1991. The high school was a private, but subsidized, school where education was provided in Basque, while Spanish and English (or French) were taught as language subjects. The alternative to this private Basque school was the public school where three different language options were offered: an education in Basque (with Spanish as a language subject), an education in Spanish (with Basque as a language subject) and a mixed education in Basque and Spanish (where some subjects are taught in Basque and others in Spanish).[10] The main deciding factor to choose among these competing alternatives was the preference for an education in Basque.[11] Tables A.1-A.4 of the web appendix show the comparison of the main macro variables between Gipuzkoa, Basque Country and Spain. Overall, we see that Gipuzkoa is no different from other provinces in the region in terms of the main demographic variables.[12]

The natural experiment occurred in the academic year 1990-1991. Typically, students would receive a report card at the end of each quarter (November, February, April and June). These report cards would provide the list of subjects taken and the grade obtained in each of the subjects. Grades measure absolute performance, since there was no grade curving (see Section 4, Figures 1a and 1b). In the academic year of 1990-1991, the treatment year, the

[9] The web appendix is available at http://www.econ.upf.edu/~azmat/ and http://www.econ.upf.edu/~iriberri/ .

[10] There was no annual fee to attend public school, while for the private Basque school the fees were subsided by the local government.

[11] Basque private schools, known as *Ikastolas*, were first founded in the 1960s. For more information about the history of Ikastolas, see www.ikastola.net .

[12] This is important for external validity purposes, since it suggests that we are not dealing with an outlier school or region.

computer application used to produce students' report cards changed.[13] This change resulted in students being provided with additional information that facilitated social comparison. In particular, as well as the list of grades obtained in each of the subjects, students were also provided with their own average grade across all subjects and the *class* average grade across all subjects. This allowed a direct comparison between the students' own average grade with the average grade of the class. Moreover, students could observe whether they were performing above or below the class average, as well as the distance between their own average grade and the class average grade. Given that students received report cards four times during the academic year, they received this additional information four times during the treatment year. However, we only have one grade per subject for each academic year, which is an average over the four quarter grades (see Section 4 for a discussion on how this may affect the parameter estimates). Finally, the information treatment was removed after the academic year 1990-1991, lasting for only one year, and consequently the additional information was simply omitted from academic year 1991-1992 onwards. The removal was due primarily to parents' and teachers' complaints.[14] See Figures A.1, A.2, and A.3 in the web appendix for an example of the report cards before, during and after the experiment, respectively.

There are several important features of this natural experiment that make it almost like a randomized field experiment. First, it was an experiment that took place in a real environment, where grades can be used as a measure of real effort. Second, the introduction of the additional information was exogenously applied, without being a meditated decision of school officials or teachers. The new computer application offered the possibility of providing the extra information and the administrative staff decided to incorporate it. Third, it was untargeted, that is, it was not introduced as a response to any initiative to affect performance. Finally, it took place in an arbitrary year that was not systematically different from any other year in our sample. In principle, no other significant differences occurred in 1990-1991 with regard to class-sizes, teachers, subjects/material taught and the evaluation system, as we will justify in Section 4 (Table 3).

Although all students in year 1990-1991 were affected by the treatment, the richness of our data in terms of number of years and individual level panel data, as well as the off-on-off nature of the treatment, gives us a quasi-control group. We use all years in our analysis and as

[13] The adopted software was provided by COSPA. For more information see http://www.cospa-agilmic.com

[14] From private communication with school officials, we could find out that the main complaint against providing this additional information was that it fostered *competition* among students, which many parents and teachers considered it to be a negative thing.

a robustness check, we also restrict the analysis to the year prior to and the year post the treatment. If there are any contemporaneous shocks around the treatment period, we would identify them with this analysis.

3. Theoretical Predictions: Why We Would Expect Students to React to Relative Performance Feedback Information

In this section we review two different theoretical frameworks that predict how students would react to the additional information. On the one hand, students may react to the additional information because they have inherently competitive preferences. This implies that students get utility (disutility) from being ahead (behind) of others. On the other hand, considering standard selfish preferences, students may react to the additional information because individuals have an imperfect knowledge of their own ability, and the additional information that allows for social comparison is informative of one's ability. We are unable to disentangle whether the effect will come from students or their parents. Given that parents' preferences are indistinguishable from their children's preferences, the same predictions are expected if we substitute students' preferences by their parents'. Therefore, when we make references to students' preferences, we are referring to both students' and their parents' preferences.

Consider $\mathrm{N} \geq 2$ students who differ in their ability, $a_i \in F\left[\underline{a}, \overline{a}\right]$, and choose effort levels, $e_i \in \left[\underline{e}, \overline{e}\right]$ where $i = 1,2,...N$. For each student *i*, both ability and effort levels yield deterministically their performance at school, given by the expression $p_i(a_i, e_i)$, which is represented by their grades. Performance is assumed to be increasing and strictly concave in both ability and effort. Effort is costly and the cost function, given by $c(e_i)$, is increasing and strictly convex in effort. Moreover, effort and ability are complements in performance, that is, effort is more productive for high ability students than for low ability students.[15] The

[15] The assumptions for the performance function include $\frac{\partial p_i(a_i,e_i)}{\partial a_i} > 0$, $\frac{\partial^2 p_i(a_i,e_i)}{\partial a_i^{\,2}} < 0$, $\frac{\partial p_i(a_i,e_i)}{\partial e_i} > 0$, $\frac{\partial^2 p_i(a_i,e_i)}{\partial e_i^{\,2}} < 0$. Effort and ability being complements in performance means that $\frac{\partial^2 p_i(a_i,e_i)}{\partial a_i \partial e_i} > 0$. The assumptions for the cost function include $\frac{\partial c(e_i)}{\partial e_i} > 0$ and $\frac{\partial^2 c(e_i)}{\partial e_i^{\,2}} > 0$. Note that this specification is equivalent to having a performance function that only depends on effort and a cost function that depends on both ability and effort in a way that effort is at least as costly for low ability students as to high ability students.

predictions from the competitive preferences theory do not rely on the complementarity between ability and effort, but the predictions from the self-perception theory do.[16] Finally, students receive two types of signals: signals containing information about their own performance, s_i, (no relative performance feedback information) and a signal containing information about average performance, $\bar{s}$, (relative performance feedback information). How the signals will be used and interpreted by the students is dependent on the theoretical framework. This will be described in more detail in each of the subsequent sections.

In the following sections, we will compare students' optimal effort levels, when the relative performance feedback is provided (treatment) and when it is not (control), for the two different models.

3.1 Preferences-based Theory: Competitiveness

We will show that the competitiveness theory predicts that students will react to the additional information exerting more effort and therefore, we would expect a higher performance level during the treatment year.

Ability and effort levels are assumed to be privately known to each student, such that students choose their optimal effort level. The utility shown below presents a specific form of competitive preferences.[17] We assume all individuals have homogeneous preferences.

$$u_i = p_i(a_i, e_i) - c(e_i) + \alpha \left[\frac{p_i(a_i, e_i) - E\left[\frac{1}{N} \sum_{k=1}^{N} p_k(a_k, e_k) \right]}{\sigma_{\bar{p}}} \right] \quad \text{for } i = 1,2,...N. \quad (3.1)$$

[16] For competitive preferences, we can get the same results assuming that effort is equally productive for high and low ability students or even assuming that effort is less productive for high ability students than low ability students. However, self-perception theory's predictions are highly dependent on the complementarity of ability and effort in performance. According to self-perceived ability theory, when effort is equally productive for high and low ability students, the provision of relative performance feedback information would have no effect at all, and when effort is assumed to be less productive for high ability students than low ability students, we would get the opposite results, meaning high (low) ability students would be discouraged (encouraged) by the relative performance feedback information. Note that these predictions are not consistent with the empirical findings of this study.

[17] Many specific models that incorporate competitiveness have been proposed. The model proposed in this paper is close to Kandel and Lazear's (1992) model where *peer pressure* enters additively into the utility function. A specific form of peer pressure mentioned by the authors is the difference between the average effort and one's effort, which is the same as our functional form. Charness and Rabin (2002) propose a simple piece-wise linear utility in which others' payoffs affect one's utility. One type of interdependent preferences their utility model includes is that of *competitive* preferences, where others' payoffs enter negatively in one's utility. Dubey and Geanakoplos (2004, 2005) and Moldovanu et al. (2005) assume individuals have knowledge of the complete ranking and they assume individuals get positive utility from the number of individuals below them and negative utility from the number of individuals above them. Hopkins and Kornienko (2004) propose a utility in which "status" or position in the ranking enters multiplying the absolute income.

The first difference compares the benefit and cost of effort. α >0 represents the weight given to the competitiveness. Moreover, $E\left[\frac{1}{N}\sum_{k=1}^{N} p_k(a_k, e_k)\right]$ is the expectation of the average performance and $\sigma_{\bar{p}}$ is the standard deviation, which measures the precision of such an expectation. The second difference captures a competitive game, where students receive a positive utility if they perform above the expected average and a negative utility if they perform below the expected average. The intuition behind this second difference resides in the *appreciation* or *depreciation* of a specific performance level, depending on whether it outperforms or underperforms with respect to the expected class average. For example, a grade of 7, in a scale between 0 and 10, will yield higher utility if the expected average grade in the class was 6, than if it was 8. In other words, any performance level that is above the expected class average is *inflated*, while any performance level that is below is *deflated*. Although the utility in (3.1) shows some resemblance to the utility function presented by Bolton and Ockenfels (2000) to represent *inequity aversion* preferences, there are significant differences.[18]

Finally, although students care about whether they are performing above or below the class average, the importance given to this social comparison or competitive term is also dependent on how precisely they know the class average. When the expected class average is very noisy or imprecisely known, then students give less weight to such comparison. The higher the precision, the lower the standard deviation of the expected class average, students give more weight to the competitive part of the utility function.

Since own performance is privately known, $s_i = p_i(a_i, e_i)$, we can write $E\left[\frac{1}{N}\sum_{k=1}^{N} p_k(a_k, e_k)\right] = \frac{1}{N} p_i(a_i, e_i) + E\left[\frac{1}{N}\sum_{k \neq i} p_k(a_k, e_k)\right]$, where the unknown random variable $\overline{p}_k = \frac{1}{N}\sum_{k \neq i} p_k(a_k, e_k)$ is assumed to be distributed according to $N(\mu_{\overline{p}_k}, \sigma^2_{\overline{p}_k})$. Therefore, when students do not get relative performance feedback information (*NRPFI*), the utility function is as follows:

[18] Individuals who show inequity aversion get disutility when their outcome is different from the average outcome, whether their outcome is above or below the average, since they want to reduce differences and inequalities. However, competitive individuals get disutility only if their outcome is below the average outcome since when their outcome is above the average they want to increase differences and inequalities. The overall prediction according to inequity aversion preferences is that students who find out they are performing below (above) the expected class average would put in higher (lower) effort. Therefore, the overall grade dispersion would decrease.

$$u_i^{NRPFI} = p_i(a_i,e_i) - c(e_i) + \alpha\left[\frac{p_i(a_i,e_i) - \left(\frac{1}{N}p_i(a_i,e_i) + \mu_{\bar{p}_k}\right)}{\sigma_{\bar{p}_k}}\right] \text{ for } i = 1,2,...N. \text{ (3.2)}$$

During the treatment year, students in addition to their own performance, they are provided with relative performance feedback information (*RPFI*), that is, a noisy signal of the average performance:

$\bar{s} = \bar{p}_k + \varepsilon$ (3.3)

where ε is distributed according to $N(0,\sigma_\varepsilon^2)$. The two random variables, $\bar{p}_k$ and ε, are independently distributed. When relative performance feedback information is provided, students will choose their effort level conditioning on the received signal $\bar{s}$, as shown below:

$$u_i^{RPFI} = p_i(a_i,e_i) - c(e_i) + \alpha\left[\frac{p_i(a_i,e_i) - \left(\frac{1}{N}p_i(a_i,e_i) + \mu_{\bar{p}_k|\bar{s}}\right)}{\sigma_{\bar{p}_k|\bar{s}}}\right] \text{ for } i = 1,2,...N. \text{ (3.4)}$$

Result 1: Given competitive preferences shown in (3.1), for any ability level, the optimal effort level when relative performance feedback information is provided $(\bar{s})$ is *higher* than the optimal effort level when no such information is provided.

The proof, shown in the Proof of Result 1 in the web appendix, is straightforward when comparing the first-order conditions for the two informational conditions. The intuition behind this result is that the purely competitive part of the utility function pushes the effort level up. If we focus only on the competitive part, such that effort is costless, regardless of what other students choose, students can do no better than to choose the highest effort level. Under relative performance feedback information, the expected class average becomes more precise, such that more weight is given to the competitive part of the utility function. Since the competitive part pushes the effort choice up, the optimal choice of effort under relative performance feedback information is higher.

3.2 Self-perception Theory: Learning about Own Ability

We will show that the self-perception theory predicts that high ability (low ability) students will react to the relative performance feedback information by exerting more (less) effort and therefore, we would expect a higher (lower) performance level for high ability (low ability) students during the treatment year.

We adapt the model proposed by Ertac (2006) to the type of relative performance feedback information provided in the natural experiment we study.[19] The main feature of this model is the assumption that students do not perfectly observe their own ability, such that they use both own performance feedback information and relative performance feedback information, one's performance in comparison with others' performance, to learn about it. Students receive a noisy signal of their own ability.

$$s_i = a_i + \eta \qquad i = 1,2,...N. \quad (3.5)$$

The shock, η, represents a common shock to performance. It can be interpreted as the easiness of the exam. Ability, a_i, is independently distributed according to $N(\overline{a}, \sigma^2)$ and the common shock, η, is distributed according to $N(0, \psi^2)$. In addition, ability and the common shock are independently distributed. Furthermore, when the social comparison information is revealed, students also observe the average signal.

$$\overline{s} = \frac{\sum_{k=1}^{N} s_k}{N} = \frac{\sum_{k=1}^{N} (a_k + \eta)}{N} = \frac{\sum_{k=1}^{N} a_k}{N} + N\eta \quad (3.6)$$

Both the individually received signal, as well as the average signal (when provided), will be informative about students' own ability. Self-perceived ability in turn determines the optimal effort level. Both ability and effort levels yield deterministically their performance at school and for simplicity we will assume that performance is given by $p_i(a_i, e_i) = a_i e_i$.

On the one hand, in the absence of relative performance feedback information (*NRPFI*), students can only use their private signal about own performance (s_i) to form the expected value of their own ability. The utility function is the same as in (3.1), when the competitive part is absent (α=0).

$$u_i^{NRPFI} = E\left[p_i(a_i, e_i) - c(e_i) \middle| s_i\right] = E\left[a_i \middle| s_i\right] e_i - c(e_i) \qquad (3.7)$$

On the other hand, when relative performance feedback information is provided (*RPFI*), in the form of the average performance of the class composed by *N* students, $\overline{s}$, relative performance information is also used to form the expected value of students' own ability.

$$u_i^{RPFI} = E\left[p_i(a_i, e_i) - c(e_i) \middle| s_i, \overline{s}\right] = E\left[a_i \middle| s_i, \overline{s}\right] e_i - c(e_i) \quad (3.8)$$

[19] Ertac (2006) presents a principal-agent model and analyzes the effect of feedback information regarding own past performance and others' past performance under different types of contracts. Since in the natural experiment we study the treatment variable is relative performance feedback information in the form of average grade of the class, and the incentive structure is fixed where students' performance is evaluated according to their grades (piece-rate), we focus on the effect of the class average grade on students' effort levels.

Result 2: If $s_i > s*$ then $e^{NRPFI*}(s_i) < e^{RPFI*}(s_i, \bar{s})$ and if $s_i < s*$ then $e^{NRPFI*}(s_i) > e^{RPFI*}(s_i, \bar{s})$, where $s* = (\bar{s} - \bar{a})\frac{N(\sigma^2 + \psi^2)}{\sigma^2 + N\psi^2} + \bar{a}$. Students whose signal is *above* (*below*) $s*$ would put in *more* (*less*) effort, when social comparison information is provided.

The proof, shown in Proof of Result 2 in the web appendix, is straightforward when comparing the first-order conditions for the two settings. The comparison reduces to the difference between $E[a_i | s_i]$ and $E[a_i | s_i, \bar{s}]$. Note that when the average signal is equal to the unconditional expected ability, $\bar{s} = \bar{a}$, such that $\bar{s}$ does not inform about the easiness or difficulty of the exam, then the $s* = \bar{s}$. Every student whose signal is above (below) the average signal would put in higher (lower) effort level when the social comparison information is provided. However, when $\bar{s} \neq \bar{a}$, then the average signal is informative about the easiness or difficulty of the exam, which determines the threshold signal, $s*$, to be higher (lower) than the average signal when $\bar{s} > \bar{a}$ and $\bar{s} < \bar{a}$ respectively.

3.3. Testable Hypothesis

We now summarize the main hypothesis regarding the predicted sign of the effect that the relative performance feedback information can have on performance, based on the alternative theoretical models depicted in the previous section.

Null Hypothesis: No effect on grades.

The null hypothesis is that we should find no effect for the additional information provided during the treatment year. There are two main explanations for why this might be the null hypothesis. Firstly, based on the preferences-based explanation, this would suggest that either the students' utility is unaffected by relative performance feedback information (no competitive preferences), or that the students already possess very precise information that allows for social comparison, such that, the fact that it is explicitly provided adds no extra information. Second, based on the self-perception explanation, this would suggest either that students do not have an imperfect notion of their ability or that again this relative performance information is known without the explicit provision of it.

Alternative Hypothesis:

(1) *Positive* effect on grades for *all* students.

(2) *Positive* effect on grades for *high* ability students and *negative* effect on grades for *low* ability students.

We consider two alternative hypotheses. On the one hand, based on the preferences-based explanation and assuming that the additional information that allows for social comparison is really new to the students, then we would expect *all* students' grades to be higher during the treatment year with respect to the other years. Also, we should observe no differences in the dispersion among the grade distribution. On the other hand, based on the self-perception explanation we would expect students' grades to be *higher* (*lower*) for those *high* ability (*low* ability) students because students who find out they are *above* (*below*) certain threshold should be encouraged (discouraged). This implies that the dispersion among the grade distribution should increase. Note that for being able to discriminate among the two hypotheses, it is necessary to look at the effect of the information treatment throughout the distribution of students' grades, as well as to look at the effect of the treatment on the dispersion among students' grades (see Section 5.4).

4. Data Description and Descriptive Statistics

In this section we begin by describing the data from the natural experiment. We have data on students' grades for all subjects between the academic years 1986-1987 and 1994-1995 (3,414 grades). Grades range between 1.5 and 9.5, see Table A.5 in the web appendix for a full list of the possible grades and their numerical conversion. Although students received their report card four times in an academic year, we can only observe their yearly grades by subjects, which are an average over the four quarters. This has two implications. First, since students receive this information for the first time when they receive their grades in the first quarter, they can only react to the additional information from the second quarter onwards. In turn, this implies that any effect that we observe on the average grade over all four quarters will be weaker than the "true" effect. Second, we are unable to observe whether the effect is equally intense in the second, third and fourth quarters or the effect is strongest in the second quarter and then vanishes. There is, however, some other evidence on these dynamics. Hannan et al. (2008) provide laboratory experimental subjects with relative performance information three times, and they observe that the effect does not vanish over time. Azmat and Iriberri (2010) use summation-solving as the task and relative performance feedback information is provided four times. They find that, although there is an effect at each period, the effect does diminish over time.

Students stay in high school for four years, starting at the age of fourteen and finishing at the age of seventeen. We will refer to each of the four high school years as Levels 1 to 4, while *years* will refer to academic years between 1986 and 1994. We are able to identify each student and follow them through each level of school. Overall, we have an unbalanced panel of 1,313 students. In Table 1, we show the structure of our data. The academic year 1990 is the treatment year and four cohorts (in different levels) were affected by the treatment. There are twelve cohorts in total, of which, six are full cohorts (i.e., we follow the students through all four levels). In our analysis we will compare the treated students with the untreated students. In Table 1 we also report the number of students per academic year and per level. Overall, the number of students per year is quite similar over the years. In the period prior to the treatment, the number of students is slightly higher than in other years. This is due to the end of the "baby boom". Also from 1990 onwards a nearby middle school extended its studies to high school level. This should not affect our analysis but we do, however, check that there was no selection or reduced class size effects coming from this change.[20]

In Level 1, students are randomly divided into three (or four) groups, depending on the number of students, such that each group has about 30 students. In Levels 1 and 2, students have a specified set of compulsory subjects that they must undertake. However, in Level 3, students have to choose between Arts or Science specializations, which they usually follow through in Level 4. In Table 2, we list the subjects and their mean grades and standard deviations by level. We can see that the average grades in Mathematics, Physics and Technical Drawing are generally lower while in subjects such as Religion, Music and Physical Education are typically higher.

Grades are not curved and therefore, they measure absolute performance. Figures 1a and 1b provide evidence that grades are not curved. In Figure 1a, we show the grade distribution for Math. Each vertical bar represents the distribution of grades in a particular group and academic year, starting with 1986 and finishing with year 1994. The shaded regions show the proportion of students that obtain a certain grade (1.5, 4, 5.5, 6.5, 7 and 9.5). In line with grades not being curved, the distribution of grades varies widely across group-years. We have repeated this analysis for other subjects and we get quantitatively similar results. Figure 1b shows three examples of grade distributions for a given teacher and level. Since each teacher could have a specific grade curving in a given level, comparing the distribution of grades by the same teacher, who is teaching multiple groups, is an additional way of testing for grade

[20] We have the grades data on the students who chose to stay in their nearby school but exclude these students from our analysis, as we do not know whether they received the additional information. As a robustness check, we have repeated our analysis including these students and our results remain unchanged. Moreover, class sizes do not change as shown in Table 3. The school had three groups in Level 1instead of four but this was the case for all subsequent years and in many other levels over time.

curving. In Figure 1b, we clearly see that the distributions of grades by the same teacher significantly vary across different groups, implying clearer evidence against grade curving.

At the end of Level 4 students also take a standardized final exam called Selectividad (similar to the Scholastic Aptitude Tests (SAT) used in the United States) before they can access University. For the students in our sample, we have data on their Selectividad grades. The final grade, that will determine entry into University, is composed of 50% of the Selectividad grade and 50% of the average grade of Levels 1 to 4. However, the Selectividad exams are based on material covered only in Level 4, which gives Level 4 a much higher weight on determining the University entry grade.

In Table 3, we list the main descriptive statistics for all years combined and separately for the treatment year in 1990. From the table we can see that there were no significant differences between the treatment and the other years regarding class sizes, the number of teachers, students' gender composition and the proportion of repeaters. The only noticeable difference is the drop in attrition from Level 1 to 2 and from Level 3 to 4. Since, the leavers were typically the students who were performing badly we expect that this fall in attrition will dampen any effect from treatment. In Section 5.7 we look at this change in more detail and see that this is not affecting the results. The other difference we see is that the average number of students and therefore the number of groups in Levels 2 and 3 are slightly higher during year 1990 than in other years (a likely consequence of the baby boom cohorts). However, as it is shown in Table 3 class sizes remain unchanged, which is the key variable. Overall, class sizes are on average around 30 students; the number of teachers per year is between 12 and 15; the frequency of girls is slightly higher than 50 percent; and the Science track is more frequently chosen than the Arts track.

5. Econometric Analysis

This section identifies and quantifies the effect the relative performance information feedback had on students' performance. We split the analysis into several parts. First, we analyze whether the treatment had any effect on students' performance. Second, we proceed to quantify this effect and check for its robustness. From these results, we are able to reject the null hypothesis of there being no effect. Third, in section 5.3, we look more closely at the impact of the treatment across different high school levels, students' gender, and across different types of subject. We then move from the mean analysis to the distributional analysis, focusing on the treatment effect along the distribution of students' grades or abilities, as well as the effect on the dispersion among students' grades. This section is of particular interest as it helps us to discriminate among the two proposed theoretical explanations. In section 5.5 we analyze whether the treatment had any lasting effect. In section 5.6, we are able rule out that the effect was artificially driven teachers within the school, to which we refer as external validity. We conclude the section with some robustness checks.

5.1. Identifying the Effect: Kernel Distribution

Figure 2 shows the kernel distribution of grades for all students before (1986-1989), during (1990) and after (1991-1994) the additional information treatment. We observe that the grade distribution is to the right of the grade distributions observed before and after the treatment. This shows that the additional information had an effect and that this effect was positive, resulting in higher grades during the treatment year. Moreover, we see that the treatment affects all parts of the distribution and in particular the tails of the distribution. Once the treatment is removed, we observe that the distribution of grades moves back in the direction of the distribution before the treatment was introduced. However, this post treatment distribution does not completely return to the pre treatment distribution. This may be due to either lasting effects of the treatment after it is removed, or due to grade inflation over time.[21] We disentangle the two effects in the following analysis. Note that grade inflation does not imply grade curving. Grade curving would involve the grade distribution (frequency of Fail, Pass, Good, Very good and Excellent) being kept constant by teachers/exam boards over time. Grade inflation on the other hand, assuming students' ability is constant over time, would imply that shifts in the grade distribution are due to the exams becoming overall easier or that the grading overall becoming more lenient. We are able to rule out that the positive trend in grades is driven by the teachers within the school (see Section 5.7 on external validity).

Figure 3, shows the kernel distribution of the grades for all students before (1986-1989), during (1990) and after (1991-1994) the additional information treatment, separated by Level. From the figures, it is clear that the strongest positive effects appear in Levels 1 and 4. We see that both the tails and the mean shift to the right during the treatment year. With regard to Levels 2 and 3, the differences appear to be more spurious.[22]

The rest of our analysis will quantify these results and test their robustness.

[21] Since we are following students throughout their high school years, those students who received treatment in Levels 1, 2 or 3, remained in the school in their following year. Having had the information about their position in the class might also affect their performance in subsequent years (despite the information being removed). We refer to this effect as being the lasting effect studied in Section 5.4.

[22] Note that in Level 1 there is no lasting effect from previous years but in Levels 2, 3 and 4 any difference that we observe in the post-treatment years might be due to the lasting effect.

5.2. Quantifying the Effect: Estimation

In this section we proceed to quantify the effect of the relative performance feedback information on performance. We start with a simple estimation of this treatment effect on the average grade (across all subjects) at the individual student level.[23] We compare various estimators. Then, we check for the robustness of this effect using controls and placebo treatments.

We begin with a simple estimation to quantify the effect of the additional information on the average grade (across all subjects) of student *i* in year *t*, $Grade_{it}$. We pool all years between 1986 and 1994, identifying separately the treatment year 1990. We also include a linear trend to capture the general evolution of grades over the years.

$$Grade_{it} = \beta_0 + \beta_1 Trend + \beta_2 Year1990 + \varepsilon_{it} \qquad (5.1)$$

In Table 4 columns 1 to 3 we show the results from this estimation using three different estimators: ordinary least squares (OLS), random effects (RE) and fixed effects (FE), respectively. The random and fixed effects are at the student level. According to the three estimators the additional information that allowed for social comparison clearly had a positive and highly significant effect on students' average grades. Overall, the marginal effect is between 0.275 and 0.296, which at the average grade corresponds to approximately 4.5% increase in performance. This is a remarkable effect with significant policy implications. Our results are comparable to other factors that have an impact on increasing performance in schooling, such as reduced class sizes, or increased school expenditure (see Krueger, 1999). However, many papers have found small or no effect of expenditure on students' performance (Hanushek (1996)). It is important to note that, while these other measures have shown to be quite costly, providing information involves almost no cost.

We extend our analysis to include additional control variables, *X*, where *X* includes gender, level of study (Level 1-4) and whether the students are repeating the level.

$$Grade_{it} = \beta_0 + \beta_1 Trend + \beta_2 Year1990 + X_{it}'\delta + \varepsilon_{it} \qquad (5.2)$$

From columns 4 to 6 in Table 4, we can see that the effect of the treatment year remains positive and significant, although the coefficient falls very slightly for each of the three estimators. With respect to the control variables, the results go in the expected direction. Female students outperform male students significantly. Students repeating a level do significantly worse compared to others when all students are pooled together (in the OLS

[23] The information was given at class level. However, since students are randomly assigned into classes in Level 1, there should be no difference across classes in the average grade. We check for this and find no difference.

specification) but once unobserved ability differences (i.e., individual fixed effects) are taken into account, they improve on their own previous performance. Regarding different levels, the students in their final levels, Level 3 and Level 4, do on average worse than in the first two levels. We can see that the students peak in their second year, Level 2, and do the worst in their final year, Level 4. This is plausible since the final years are more demanding than the first years but at the same time, the first year involves adjustment to the new environment (the transition from middle to high school). In the specifications that control for individual fixed effects, the coefficient on trend is sensitive to whether or not we control for levels of high school. This is so because the trend is also at an individual level, rather than at an aggregate level.[24]

Using only the OLS, one could argue that the observed effects are a result of the students in year 1990 being intrinsically different from other years. For example, if there was a complete replacement of students with higher ability students in the treatment year, we would expect to observe exactly the same effect. Since we have individual level data over each of the high school years (panel data at student level), we are able to rule out this possibility using the panel data estimations and by using a dynamic OLS specification (see columns 7 and 8). The inclusion of students' past grades enables us to control for students' unobserved characteristics such as ability. We can only do this for Levels 2 to Level 4. As we would expect, previous grades are usually highly correlated with current grade. If there was something special about the 1990 students it should be captured by the coefficient on previous grades, making the *Year 1990* variable insignificant. However, we see that this is not the case because by using RE, FE and the dynamic OLS specification, the treatment years' effect remains positive and significant.[25]

Our preferred estimator is the RE, since we do not lose variables that are fixed over time (as we do with FE) nor do we lose information by lagging grades (as we do with the dynamic

[24] In the OLS regressions, the interpretation of the trend is the grade inflation over time and the controls for Levels 1 to 4 provide a comparison of the way in which grades change over the four high school years. When controlling for individual fixed effects, the trend variable is individual specific, such that it provides an estimate for how well the student does over his/her own time at the school. Without controls for levels, the trend can not be differentiated from the effect of being in the different levels - it is negative as courses become more difficult over time for the student. When we include the controls for levels in these regressions, the trend interpretation is again similar to the OLS interpretation. Note that this will not affect the coefficient on the treatment.

[25] When we restrict our sample to the sample used for dynamic OLS estimation (i.e., not including Level 1 students) and re-estimate with OLS, our results are in-line with those found in Table 4. We find the effect is 0.237 (0.083) and 0.172 (0.080) without and with controls, respectively.

OLS).[26] In the tables that follow we will use RE whenever we use the panel element and OLS when using repeated cross sectional data.

We carry out two further robustness checks. First, we cluster the average grades at the group and year level. In addition, to rule out the concern that the results are being driven by significant changes in the pool of teachers during the treatment year, we repeat the analysis controlling for the teacher fixed effects. In both cases, the treatment coefficient remains positive and significant at the 1% level. These results can be found in Table A.6 of the web-appendix.

In order to check whether the positive and significant effect that we have found is particular to the treatment year, we estimate equation (5.1) for the other years in our data. We perform this placebo treatment in two ways. First, we use only the years prior to 1990. Since there are potential lasting effects of the treatment in the subsequent years, we avoid this by only including the prior years to the treatment. Second, we use only Level 1 students' data for *all* years (except 1990), since there is obviously no concern for there being any lasting effect of the treatment for these students.

Figure 4 shows the placebo treatments for the years prior to 1990. Here, we can see that for all years the *treatment* is insignificant at both the 1% and 5% level. Moreover, there is a clear spike in year 1990, with no increasing pattern in the average grades over the years. To ensure that this effect is neither a new state nor the beginning of a new increasing pattern, we want to be able to check the post treatment years. The only way we can cleanly do this, is by using data for Level 1 students only. This is shown in Figure 5. We can see here that the spike remains in the treatment year and that all other years are insignificant with the exception of 1994. In 1994, we see a large drop in the average grade which we cannot fully explain. This may be due to a smaller sample size that we have for Level 1 in 1994 or some other change in the school that we are unaware of. To ensure this is not affecting our main result, we replicated all of our analysis (equations (5.1) and (5.2)) by removing 1994. By doing so our main results hold and the coefficients are unchanged.

5.3. Quantifying the Effect: Levels, Gender and Subjects.

In this section we look more closely at the impact of the treatment effect across levels, students' gender and subjects.

[26] A Hausman test, based on a contrast between the FE and RE estimators gives a chi-squared statistic of 2.5. This is not significant at the 5% level and so we do not reject the null hypothesis of no correlation between the individual effects and explanatory variables.

Since one might expect important differences across levels in reacting to such a policy, we begin by disaggregating the effect of additional information on average performance across Levels 1 to 4. In Table 5 we estimate equations (5.1) and (5.2) from the previous section by level. What is striking from this table is that while the effect is insignificant for Levels 2 and 3, there is a strong and positive effect for the first and final levels, Levels 1 and 4. Moreover, the coefficients on both Levels 1 and 4 are twice as large as the quantified effect at the aggregated level (shown in column 1). These estimates imply that the additional information led to an increase of 8% and 9% in the grades of students in Level 1 and Level 4, respectively. A plausible explanation for such a difference across levels may be related to how much prior knowledge students have about their position within the class. One might expect that the first year students have very little information about the ability of their classmates and therefore, whether they are above or below the average. Students in the other levels, on the other hand, might have a clearer picture of their position within the class. This is in line with the ability perception theory. Although we do not find a discouragement effect on the low performing students when the relative performance feedback is provided, we do find that the effect on grades is strongest on Level 1 students. These are the students for whom the relative performance information is likely to be more informative. However, this does not explain the strong and positive effect in Level 4. One explanation for such an effect might be due to the importance grades attain in this final year. The grades in Level 4 strongly determine the entry grade for university, making them very prominent during this year. Students might therefore put a greater emphasis on social comparison during this crucial year.

Next, we turn into the analysis of gender. We test whether girls react differently to the information about relative performance. We estimate the following equation.

$$Grade_{it} = \beta_0 + \beta_1 Trend + \beta_2 Year1990 + \beta_3 Girl + \beta_4 Girl * Year1990 + \varepsilon_{it} \qquad (5.3)$$

In Table 6 we can see that, although girls do better than boys throughout high school, there does not appear to be any gender differences in reaction to the additional information. This is consistent with Niederle and Yestrumskas (2008).

Finally, we disaggregate the analysis at the subject level. We group subjects into four: Languages (Basque, Spanish and Foreign Language), Sciences (Maths, Biology, Chemistry, Physics, Geology and Technical Drawing), Arts (History, Latin, Philosophy, Literature, Greek and History of Art) and Others (Technical and Professional Studies (TPS), Physical

Education, Religion/Ethics, Music and Drawing).[27] Table 7 includes the estimates for all levels and for each level separately. In column 1 we can see that students improve their performance in all subject groups. Moreover, the strongest effect is found in the Science group. From this analysis there are important policy implications. We can see that students are improving in subjects considered very relevant such as Math and Languages rather than in subjects such as Physical Education. In recent years the poor test scores in technical subjects, such as Physics and Math, in many western countries have hit the headlines and the improvement of which has been regarded as being high priority (See PISA reports, 2006). In columns 2 to 5 in Table 7, the estimates are presented for each of the different levels. Language and Science subjects show similar pattern to the aggregate results, with regard to the different levels. There appears to be a positive and strong effect on Levels 1 and 4, while the intermediate levels are unaffected. The exception to this is the Science subjects, which appear also significant (at the 5% level) during Level 2. Also, although we have seen a positive and significant effect on Arts, the level analysis shows that this is solely driven by changes in Level 3 that we cannot explain.

5.4. Quantifying the Effect: Distributional Analysis

The estimation analysis has so far focused on the mean effect of the treatment. However, it is important to understand how students with different levels of ability reacted to the treatment. Here, we analyze the impact of treatment along the ability distribution. This is particularly important as it allows us to discriminate between the two proposed theoretical frameworks, the competitiveness and the self-perception models. We showed that the competitive preferences hypothesis predicted similar reaction by high ability and low ability students, while the self-perception hypothesis predicted opposite reactions by low ability and high ability students. In this section we find evidence in support of the competitive preferences hypothesis, rather than the self-perception theory.

We address this analysis in four different ways. Firstly, to understand which part of the grade distribution was most affected by the treatment, we estimate quantile regression using equation (5.1), and we plot the coefficients of the treatment year for each quantile in Figure 6. Although the coefficients are significant for most parts of the distributions, in line with what we observed in our kernel distributions, we can see that the students at the tails of the distributions are affected the most. This is a very interesting result, since it rejects the

[27] TPS in Spanish is called Enseñanzas y Actividades Técnico Profesionales (EATP). This subject covers topics such as, an introduction to information technology.

hypothesis that students at the lower end of the distributions might be discouraged by this kind of social comparison.

In Table 8 we show the estimated coefficient on the treatment year for each quantile, separated by level following equation (5.2). We observe that the treatment was significant and positive for Levels 1 and 4 but not so for Levels 2 and 3, consistent with the mean analysis. In Levels 1 and 4, we can see that the effect is strongest in the left tail although it is significant for most parts of the distribution.

We extend the quantile analysis to test for gender effects. We have observed that although girls overall obtain higher grades they do not react differently to the treatment. However, the absence of the differential treatment effect for girls at the mean level could be hiding the existence of differential gender effects along the distribution of grades. We estimate quantile regression using equation (5.3) but we find no significant gender effect in the reaction to the treatment along the distribution of grades for any level. Finally, we also test for the gender reaction to the treatment at the subject level, that is, separately for Science, Language, Arts and Other subject categories. We do not see a clear and consistent pattern, except in the subject group of Others (TPS, Physical Education, Religion/Ethics, Music, Drawing), where girls react significantly less than boys.[28]

Second, using cross-sectional data, it is very difficult to disentangle ability from the treatment effect in a given year. The inclusion of students' past grades enables us to control for students' unobserved characteristics such as ability. Using the panel element of the data, we can control for students' ability by the previous year's grades. We define the dummy variable $Above_{it-1}$ for those students whose grades are above the average of their level in the previous year. We interpret this as being high ability students.

$$Grade_{it} = \beta_0 + \beta_1 Trend + \beta_2 Above_{it-1} + \beta_3 Year1990 + \beta_4 Above90_{it-1} + \varepsilon_{it} \qquad (5.4)$$

In Table 9 we can see that the year 1990 did not affect differently those students who are high or low ability. This is in line with what we observed in the quantile regression. This also has a desirable policy implication. One important concern (criticism of this policy) may be that the information that facilitates social comparison might discourage those students who are performing below the average. The results in Table 9 clearly suggest that this is not the case, since there is no differential effect. Although we see a strong positive relationship between current and past grades, the treatment did not affect differently those who are

[28] These estimations are available on request.

performing above and below the average. Moreover, we see that the treatment year effect remains positive and significant with the inclusion of these additional variables.

Third, we complement the analysis above with a more refined students' grade distribution. We define four grade groups: students whose average grades are between (a) 8 and 10, (b) 7 and 7.9, (c) 5 and 6.9 and (d) 1.5 and 4.9.[29] We compute the hazard rates for each student moving across the grade groups and we analyze if there is a differential effect in the treatment year.[30] In Table A.7 of the web appendix, we show the transition rates across grade groups for each year and in Table 10 we show the differential effect between 1990 and the average across all other years. We see that a student previously in group (a) is more likely to remain in this same group in 1990, compared with other years. Overall, students in a high grade group ((a) and (b)) are less likely to move to a lower grade group in 1990 compared to other years. Moreover, the students in the lowest grade group (d) are more likely to move to a higher grade group in 1990. These results suggest that the results are positive for all student grade (ability) groups.

Finally, we turn our attention to the effect of the relative performance feedback information on the dispersion among students' grades. So far, we have observed that (overall) there has been a positive shift in the mean grades during the treatment. However, one may also be interested in understanding whether the treatment affected the spread of grades among students. The dispersion analysis offers a new angle from which we can evaluate the relevance of the competitiveness and self-perception theory, as well as the desirability of the social comparison policy. While increasing grades may be seen as a desirable outcome, increasing the dispersion may have negative connotations. In particular, increasing the gap between bad and good students could be seen as a drawback of the policy.

We measure dispersion using the variance. Our outcome variable is therefore, the squared difference between student *i*'s average grade across subjects in year *t* and the mean grade of that year and level, given by $(Grade_{it} - Mean_t)^2$.

$$(Grade_{it} - Mean_t)^2 = \beta_0 + \beta_1 Trend + \beta_2 Year1990 + \varepsilon_{it} \qquad (5.5)$$

[29] Note that here we use four grade groups rather than the six that we use in the rest of the paper. The main reason for this is that the top (9-10) and bottom (0-2.9) grade groups have few observations. Overall, the results remain the same with six grade groups.

[30] We estimate a simple transition rate (h_{ab}) using: $\Pr(S_t = a | S_0 = a, S_t \neq c, S_t \neq d) = e^{-h_{ab}t}$. We take the negative of the log to compute the transition rate. The transition rates in Table A.7 are multiplied by 100 so that they can be interpreted as the percentage of students in one grade group moving to another in the course of a year.

Table 11 shows the estimates for (5.5). From column 1 we can see that the dispersion among students was not affected by the treatment. In addition, we see that dispersion increases with the levels of high school. From our mean analysis we observed that grades fell in the later years in high school. Here, we see that the further the student progresses through high school, not only do subjects get more difficult, such that grades become lower, but there is also a greater separation between good and bad students. In column 2 we identify separately the dispersion among students who are above and below the mean (of their level) and interact it with the treatment year. We observe that there is no differential effect above and below the mean. From columns 3 to 8, this is separately done by Levels and the same results hold.

Our analysis shows that the provision of relative performance feedback information had a positive effect along the distribution of students' performance and that it did not have a significantly different effect among the high ability and low ability students. This supports the competitive preferences hypothesis rather than the self-perception hypothesis.

5.5. Lasting Effect: Did the Treatment have a Lasting Effect?

Our analysis so far has consistently shown that there is a positive and significant effect of the additional information on grades. We may also pose the question of whether there is a lasting effect of the treatment once it has been removed, that is, whether the effect persists on those students who received the treatment. Given the panel element, we are able to track students over time to investigate whether there is a lasting effect of the treatment.

In Table 12, we allow for the possibility of lasting effects and we see that once the treatment was removed there was no further effect.[31] To understand the lasting effect, one should read the table diagonally. For example, a student who was treated in Level 1 in 1990, where the treatment effect on grades (0.594) is significant, will be in Level 2 in 1991, in Level 3 in 1992 and finally in Level 4 in 1993. The corresponding lasting effect after one year is given by -0.0545 (not significant). The lasting effect two years after treatment is given by -0.234 (not significant) and finally after three years is given by 0.222 (not significant).

When the relative performance information is removed, the grades of those students who did receive the information in the previous year are not distinguishable from the grades obtained in the academic years in which no information was provided. This may be the consequence of the students' lack of awareness about the relative performance information or, alternatively, it may be that the past information that students received is no longer relevant;

[31] Since we are including many years in the regressions, we do not control for a linear trend. However, the results remain unchanged when we control for the trend.

since students moved to a new level or they have new classmates. In turn, the kernel distribution of Figure 2 implies that the difference in grades that we observe before and after treatment is the result of grade inflation, and not due to a lasting effect. The exception comes from Level 4 students who got the additional information in Level 2 and experienced a positive effect in their grades but then those who receive it in Level 1 experience a negative effect. Notice that these effects are only significant at the 10%. We overall conclude that there is no a clear and consistent lasting effect.

5.6. External Validity: Effect on National Level Exams (Selectividad)

In this section we address the concern that the effect may be "artificially" driven by agents within the school such as teachers, and provide evidence that rules out this possibility.

An important concern is that the effect was entirely driven by teachers who reacted to the treatment by altering the grades of the students in that year. There are at least two potential reasons for why teachers might alter the way in which they grade. First, teachers might anticipate the complaints from those parents whose children are performing below the average. Second, teachers may anticipate bad behavior from some children. Notice that this would imply that teachers artificially increase bad performing students' grades, compressing the grade distribution, which is not what we observe in the data. However, there are also a number of reasons for why teachers should not alter the way in which they grade. Firstly, teachers always have this information since they know the distribution of grades in their class every year. Secondly, they have no monetary incentives to react, since they do not receive a "bonus" for good performance. Finally, schools in Spain are very careful to ensure that the Level 4 grades are representative of the Selectividad grade and many schools use this as a marketing tool to attract students to the school.

To address this issue, we use students' grades from "Selectividad" exams and repeat our previous analysis. The Selectividad exams are national level exams (similar to SATs in the USA) taken by students after the completion of Level 4. These exams are written and graded by external bodies on national standards, accounting for 50% of the overall grade to enter University. The other 50% is determined by grades in Level 1 to Level 4. However, the Selectividad exams are based on the material covered only in Level 4, making this an important level. The drawback of this external validity check is that we can only perform it with Level 4 students. However, since the observed effect is strongest for Level 1 and Level 4 students, it is yet a very relevant test. Moreover, it is reasonable to assume that we can extrapolate our finding from the proposed test to rule out teachers' effect and apply this to the other years.

Replicating our earlier analysis, we find that in the treatment year (1990) there is a strong and positive effect of additional information that allows for social comparison on students' Selectividad grades. In Table 13 we see in the first two columns our earlier results for the effect of treatment for Level 4 and in the last two columns the analysis is replicated using only Selectividad grades (on the same students). Here we can see that, like for Level 4, the treatment had a strong effect (around 0.63) on the Selectividad grades. It is reassuring to see that our main findings hold with the Selectividad data, where the school teachers had no influence on the grading, which suggests that the effect is coming from students putting in more effort when the additional information is provided. We replicate the analysis using placebo treatment for all other years and find a very similar pattern to our school results (see Figure 7).

Finally, it is also interesting to note that since we observe a positive trend in the Selectividad grades, we can take this as evidence that the overall grade inflation is not driven by the teachers within the school but that it was part of an overall trend. Notice also that girls have been consistently doing better than boys during the four years of High School. However, in Selectividad, the two day national level exam, girls no longer show a significantly higher performance.

5.7. Robustness: Repeaters and Leavers

In this section, we conduct some robustness exercises to complement our main analysis. In particular, we focus on those who repeat levels and those who leave school before graduating. We may be concerned that the policy affected them differentially and this affects, or even partly drives, our main results.

First, we focus on the students who repeat a level. We proceed in two ways. First, we look to see if the treatment had any effect on the probability of a student repeating a level. We find that this is not the case. Second, we re-estimate our main regression by including an interaction between the (treatment) year 1990 and the dummy for those who repeat a level. We do this for all levels, as well as for each level separately and conclude that there is no differential effect. See Table A.8 in the web appendix.

Second, we turn our attention to those students who leave school. In order for us to see if the treatment had any effect on leaving school, we look at the probability of staying, leaving after the first, second and third year respectively. Overall, we find that there is no significant effect on staying in school, neither is there any effect in leaving school after the second or third year. However, we do see that students in Level 1 during treatment are less likely to leave. It may be that the pool of student changes as a result of this, however, one would expect that the worst students are the ones most likely to leave and we therefore would expect to find the negative effect on the grades in year 1990. Since our main results show that there is a positive effect, this source of selection would bias down our results. See Table A.9 in the web appendix.

6. Conclusions

We have found that the provision of relative performance feedback information had a strong and positive effect on high school students' performance; it increased overall grades by 5%. This is a remarkable finding for two reasons. First, students did not receive any explicit reward (penalty) from performing above (below) the average, but still reacted to the information. Second, this effect is comparable with the effects found by the education literature on improving students' attainment by investing in school inputs. However, unlike investing in school inputs providing relative performance information involves no cost. Furthermore, this effect was significant for students of all ability types but it had no lasting effect once the treatment effect was removed.

We outlined two potential explanations for why students would react to the provision of relative performance feedback information and we find support for the competitive preferences hypothesis. We do not find evidence that low performing students were discouraged by this information. However, we also find that this information becomes more relevant when students are less familiar about how their ability relates to the ability of other students (Level 1 students compared to students in subsequent years). This implies that the self-perception of ability may be an important driving force.

This paper has shown the potential positive effect that the provision of relative performance feedback information can have in motivating high school students. Further research should be directed towards understanding how the provision of relative performance feedback information affects individuals in ways other than through increasing their performance.

References

Aoyagi, M. 2007. “Information Feedback in a Dynamic Tournament.” Mimeo.

Azmat, G. and N. Iriberri. 2010. “The Provision of Relative Performance Feedback Information and Happiness”. Mimeo.

Bandiera, O., V. Larcinese and I. Rasul. 2008. “Blissful Ignorance? Evidence from a Natural Experiment on the Effect of Individual Feedback on Performance”. Mimeo.

Bandiera, O., Barankay, I. and I. Rasul. 2009. “Team incentives: Evidence from a Field Experiment”. Mimeo.

Blanes i Vidal, J. and M. Nossol. 2009. “Tournaments without Prizes: Evidence from Personnel Records”. Mimeo.

Bolton, G. and A. Ockenfels. 2000. “ERC: A Theory of Equity, Reciprocity and Competition”. The American Economic Review 90(1), 166-193.

Card, D. and A. B. Krueger. 1992. “Does School Quality Matter? Returns to Education and the characteristics of Public Schools in the United States”, Journal of Political Economy, 100, 1-40.

Casas-Arce, P., and F. Martinez-Jerez. 2009. “Relative Performance Compensation, Contests, and Dynamic Incentives”. Management Science, 55: 1306-1320.

Charness, G. and M. Rabin. 2002. “Understanding Social Preferences with Simple Tests.” The Quarterly Journal of Economics, 117(3), 817-869.

Clark, A. E., and A. J. Oswald. 1996. “Satisfaction and Comparison Income.” Journal of Public Economics, 61(3), 359-381.

Delfgaauw, J., Dur, R., Sol J. and W. Verbeke. 2009. “Tournament Incentives in the field: Gender Differences in the Workplace”. Mimeo.

Dubey, P. and J. Geanakoplos. 2004. “Grading Exams: 100, 99, ..., 1 or A, B, C? Incentives in Games of Status.” Cowles Foundation Discussion Paper No. 1467, Yale University.

Dubey, P. and J. Geanakoplos. 2005. “Grading in Games of Status: Marking Exams and Setting Wages.” Cowles Foundation Discussion Paper No. 1544, Yale University.

Duesenberry, J.S. 1949. “Income, Saving and the Theory of Consumer Behaviour”. Harvard University Press, Cambridge.

Easterlin, R. A. 1974. “Does economic growth improve the human lot? Some empirical evidence”. David PA, Reder MW (eds.) Nations and households in economic growth.

Ederer, F. 2010. “Feedback and Motivation in Dynamic Tournaments.” Journal of Economics & Management Strategy.

Eriksson, T., Poulsen, A. and M. Villeval. 2009. “Feedback and Incentives: Experimental Evidence”, Labour Economics, 16, pp. 679-688.

Ertac, S. 2006. “Social Comparisons and Optimal Information Revelation: Theory and Experiments.” Mimeo.

Falk, A. and Ichino, A. 2006. “Clean Evidence on Peer Pressure”. Journal of Labor Economics, 24 (1), 39-57.

Fehr, E. and F. Ederer. 2007. “Deception and Incentives: How Dishonesty Undermines Effort Provision”. IZA Discussion Paper 3200.

Fehr, E. and K. M. Schmidt. 1999. "A theory of fairness, competition, and cooperation". The Quarterly Journal of Economics 114: 817–868.

Festinger, L. 1954. “A theory of social comparison processes”. Human Relations 7: 117-140.

Frank, R. H. 1984. “Interdependent preferences and the competitive wage structure.” Rand Journal of Economics 15: 510-520.

Frank, R. H. 1985. “Choosing the Right Pond: Human Behavior and the Quest for Status”. Oxford University Press.

Gershkov, A. and M. Perry. 2009. “Tournaments with Midterm Reviews”. Games and Economic Behavior, 66: 162-190.

Hannan, R. L., R. Krishnan and D. Newman. 2008. “The Effects of disseminating Relative Performance Feedback in Tournament Versus Individual Performance Compensation Plans”. The Accounting Review, 83-4.

Hopkins, E. and T. Kornienko. 2004 . “Running to Keep the Same Place: Consumer Choice as a Game of Status”. American Economic Review 94(4): 1085-1107.

Hanushek, E. A. 1996. “Measuring Investment in Education”, Journal of Economic Perspectives 10-4, 9-30.

Hanushek, E. A. 2003. "The Failure of Input-Based Schooling Policies." Economic Journal 113, F64-F98.

Kandel, E. and E.P. Lazear. 1992. “Peer Pressure and Partnerships”. Journal of Political Economy, 100(4), p.801-17.

Kuhnen, C. M. and A. Tymula. 2008. “Rank Expectations, Feedback and Social Hierarchies”. Mimeo.

Kräkel, M. 2007. “Emotions in Tournaments”. Journal of Economic Behavior and Organization 67, 204-214.

Krueger, A.B. 1999. “Experimental Estimates of Education Production Functions”. Quarterly Journal of Economics 114, 497-532.

Lai, E. K. and A. Matros. 2007. “Sequential Contests with Ability Revelation”. Mimeo.

Layard, R. 1980. “Human satisfactions and public policy.” Economic Journal 90: 737-350.

Lizzeri, A., M. Meyer and N. Persico. 2002. “The Incentive Effects of Interim Performance Evaluations”. CARESS Working Paper 02-09.

Locke, E. A., and G. P. Latham. 1990. “A Theory of Goal Setting and Task Performance.” Englewood Cliffs, NJ: Prentice Hall.

Mas, A. and Moretti, E. 2009. “Peers at Work”. American Economic Review. Forthcoming.

Moldovanu, B., A. Sela and X. Shi. 2007. “Contests for Status.” Journal of Political Economy 115.

Müller, W. and A. Schotter. 2003. “Workaholics and Drop Outs in Optimal Organizations,” Mimeo, New York University.

Niederle, M. and A. H. Yestrumskas. 2008. “Gender Differences in Seeking Challenges: The Role of Institutions”. Working Paper 13922 http://www.nber.org/papers/w13922.

Ok, E.A. and L. Kockesen. 2000. “Negatively Interdependent Preferences”, Social Choice and Welfare, 17, 533-558.

Prendergast, C. 1999. “The Provision of Incentives in Firms”. Journal of Economic Literature, Vol. 37, No. 1, pp. 7-63.

Programme for International Student Assessment (PISA), OECD, 2006.

Suls J. and L. Wheeler (Eds). 2000. Handbook of Social Comparison: Theory and Research (pp. 271-293). New York. Kluwer Academic /Plenum Publishers.

Young, S. M., J. Fisher and T. M. Lindquist. 1993. “The effect of intergroup competition and intragroup cooperation on slack and output in a manufacturing setting”. The Accounting Review 68 (3): 466-483.

Tables and Figures

Table 1. Number of Students by Year and Level

	Level 1	Level 2	Level 3	Level 4	Total
1986-1987					341
1987-1988	Cohort 1				342
1988-1989	Cohort 2	Cohort 1			388
1989-1990	Cohort 3	Cohort 2	Cohort 1		435
Treatment 1990-1991	**Cohort 4**	**Cohort 3**	**Cohort 2**	**Cohort 1**	**426**
1991-1992		Cohort 4	Cohort 3	Cohort 2	411
1992-1993			Cohort 4	Cohort 3	396
1993-1994				Cohort 4	376
1994-1995					299*
Total	943	879	857	735	3414

Notes: The data on students' grades are provided by Oiartzo Ikastola (school) for academic years 1986-1995. The dataset contains 3,414 students but an unbalanced panel of 1,313.

*We have one class group missing from our dataset in 1994-1995. We believe that this academic year was very similar to the other years (in terms of the number of students). We have replicated our analysis without including this year and our results remain unchanged.

Table 2. Subjects' Descriptive Statistics

	Level 1			Level 2			Level 3			Level 4		
Subjects	**Obs.**	**Mean**	**S. D.**	**Obs.**	**Mean**	**S. D.**	**Obs.**	**Mean**	**S. D.**	**Obs.**	**Mean**	**S. D.**
Basque	943	6.554	2.084	879	6.688	1.579	857	6.425	1.606	700	6.723	1.770
Spanish	943	6.437	1.951	879	6.601	1.703	124	5.964	1.810	718	5.919	1.722
Latin				879	6.848	2.068						
Foreign Language	943	6.654	1.901	879	6.128	1.902	855	6.022	1.940	714	5.937	1.699
Technical Drawing	943	6.413	1.754									
Philosophy							857	6.474	1.759	707	6.385	1.827
Geography				879	7.130	1.718						
Music	943	7.277	1.501									
History	943	6.486	1.906				857	6.461	1.989			
Religion	943	7.642	1.155	879	7.912	1.220	733	7.636	1.202			
Mathematics	943	6.215	1.953	879	6.211	1.979						
Physics/Chemistry				879	6.763	2.028						
Biology	943	6.599	1.713			.						
Physical Education	935	7.505	1.243	872	7.416	1.198	835	7.328	1.107			
TPS				879	7.888	1.063	857	7.477	1.291			
Third Level Options (Arts Track):												
Spanish Literature							306	6.165	1.704			
Latin							306	5.786	2.024			
Greek							110	5.805	1.941			
Mathematics							244	5.309	2.284			
Third Level Options (Science Track):												
Spanish Literature							13	5.885	2.033			
Biology							539	6.396	2.007			
Physics/Chemistry							550	5.880	2.349			
Mathematics							503	5.902	2.275			
Fourth Level Options (Arts Track):												
Spanish Literature										254	6.053	1.738
History										264	6.443	1.964
Latin										134	5.940	1.632
Greek										0		
History of Art										274	6.591	1.717
Mathematics										146	5.983	1.781
Fourth Level Options (Science Track):												
Mathematics										434	6.114	2.031
Physics										350	5.563	2.052
Chemistry										387	5.939	2.136
Biology										192	6.922	1.898
Geology										105	6.600	1.775
Technical Drawing										206	6.238	1.876

Notes: For each level, the mean and standard deviation for grades are reported. Students study 10 subjects per level. In Levels 1 and 2 all subjects are compulsory. In Levels 3 and 4 students can choose between the Arts and the Science track and can then choose options within each track. There are, however, four subjects (Basque, Spanish Mathematics and Foreign Language) that are taught in each level.

Table 3. Descriptive Statistics

Variable	**Level 1**	**Level 2**	**Level 3**	**Level 4**
Years 1986-1994				
Prop. of Girls	0.54	0.56	0.56	0.6
Prop. of Repeaters	0.04	0.04	0.07	0.08
Number of Groups	3.38	3.14	3.00	2.85
Group (Average size)	32.61	30.64	32.4	29.33
Attrition Number (by cohort)	--	-10.17	0	-16.17
Prop. of Science Route	--	--	0.64	0.61
Number of teachers	14	15	14	14
Year 1990				
Prop. of Girls	0.54	0.62	0.52	0.6
Prop. of Repeaters	0.03	0.02	0.07	0.04
Number of Groups	3.00	4.00	4.00	3.00
Group (Average size)	30.67	32.5	31	27.67
Attrition Number (by cohort)	--	0	0	1
Prop. of Science Route	--	--	0.7	0.56
Number of teachers	12	14	13	15

Notes: For each level, the means are reported for the treatment year (1990) and all other years in the data. In Levels 1 and 2 all subjects are compulsory. Attrition is the average number of students (in a given cohort) that leave (negative sign) or that arrive (positive sign) high school.

Table 4: Aggregate Effect on Performance (Grades): Different Estimators

	OLS	RE	FE	OLS	RE	FE	OLS (LAG)	OLS (LAG)
Constant	6.36	6.877	7.632	5.709	5.337	5.542	0.045	-0.797
	[0.046]***	[0.054]***	[0.048]***	[0.072]***	[0.087]***	[0.075]***	[0.110]	[0.114]***
Trend	0.06	-0.092	-0.26	0.079	0.07	0.065	-0.042	-0.034
	[0.010]***	[0.010]***	[0.012]***	[0.009]***	[0.012]***	[0.012]***	[0.008]***	[0.008]***
Year 1990	**0.296**	**0.286**	**0.275**	**0.255**	**0.272**	**0.273**	**0.187**	**0.204**
	[0.072]***	[0.041]***	[0.039]***	[0.069]***	[0.037]***	[0.037]***	[0.051]***	[0.047]***
Girl				0.279	0.323			0.13
				[0.046]***	[0.070]***			[0.034]***
Level 1				0.649	1.069	1.054		
				[0.066]***	[0.048]***	[0.048]***		
Level 2				0.787	1.068	1.059		0.601
				[0.067]***	[0.042]***	[0.042]***		[0.043]***
Level 3				0.376	0.59	0.584		0.178
				[0.067]***	[0.037]***	[0.037]***		[0.042]***
Repeater				-0.965	0.379	0.381		0.945
				[0.100]***	[0.059]***	[0.059]***		[0.075]***
Ave*(t-1)*							0.976	1.032
							[0.015]***	[0.015]***
Observations	3414	3414	3414	3414	3414	3414	2156	2156

Notes: * denotes significance at the 10% level, ** denotes significance at the 5% and *** denotes significance at the 1% level. The number of observations falls when using OLS (LAG) as we must restrict analysis to students in Levels 2 to 4.

Table 5: Year 1990 Effect on Performance (Grades) by Levels

	All Levels	Level 1	Level 2	Level 3	Level 4
Equation (5.1)	**0.286**	**0.531**	**0.018**	**0.099**	**0.585**
	[0.041]***	[0.141]***	**[0.117]**	**[0.130]**	[0.183]***
Equation (5.2)	**0.272**	**0.511**	**-0.027**	**0.112**	**0.558**
	[0.037]***	[0.138]***	**[0.116]**	**[0.127]**	[0.181]***
Observations	3414	943	879	857	735

Notes: * denotes significance at the 10% level, ** denotes significance at the 5% and *** denotes significance at the 1% level. We report the coefficients and standard errors for only the *Year 1990* variable but in the regression we include all the variables in equations (5.1) and (5.2), respectively. Estimations using All Levels are done using random effects (RE) estimation and OLS estimation is used when we estimate each level separately.

Table 6: Effect on Performance (Grades) by Gender

	All Levels	Level 1	Level 2	Level 3	Level 4
Constant	6.746	6.127	6.389	6.105	5.876
	[0.070]***	[0.090]***	[0.097]***	[0.116]***	[0.139]***
Trend	-0.09	0.13	0.114	0.052	-0.002
	[0.010]***	[0.017]***	[0.017]***	[0.019]***	[0.022]
Year 1990	**0.341**	**0.489**	**0.083**	**0.138**	**0.676**
	[0.063]* **	**[0.207]** **	**[0.188]**	**[0.188]**	**[0.288]** **
Girl	0.216	0.257	0.222	0.355	0.39
	[0.077]***	[0.088]***	[0.091]**	[0.100]***	[0.123]***
Girl*Year 1990	**-0.097**	**0.072**	**-0.134**	**-0.035**	**-0.143**
	[0.083]	**[0.282]**	**[0.241]**	**[0.259]**	**[0.372]**
Observations	3414	943	879	857	735

Notes: * denotes significance at the 10% level, ** denotes significance at the 5% and *** denotes significance at the 1% level. Estimations using All Levels are done using random effects (RE) estimation and OLS estimation is used when we estimate each level separately.

Table 7: Effect on Performance (Grades) by Subject Groups

	All Levels	Level 1	Level 2	Level 3	Level 4
Languages	**0.165**	**0.728**	**-0.144**	**0.195**	**0.587**
	[0.044]* **	**[0.191]*** **	**[0.145]**	**[0.158]**	**[0.182]*** **
Sciences	**0.385**	**1**	**0.361**	**0.1**	**0.65**
	[0.065]* **	**[0.182]*** **	**[0.177]** **	**[0.222]**	**[0.223]*** **
Arts	**0.336**	**0.146**	**0.186**	**0.685**	**0.265**
	[0.063]* **	**[0.209]**	**[0.191]**	**[0.162]*** **	**[0.205]**
Others	**0.157**	**0.245**	**-0.066**	**-0.098**	**NA**
	[0.041]* **	**[0.104]** **	**[0.078]**	**[0.086]**	**NA**
Observations	3404	943	879	857	725

Notes: * denotes significance at the 10% level, ** denotes significance at the 5% and *** denotes significance at the 1% level. Estimations using All Levels are done using random effects (RE) estimation and OLS estimation is used when we estimate each level separately. Language subjects include Basque, Spanish, Foreign Language; Science subjects include Mathematics, Biology, Chemistry, Physics, Geology, Technical Drawing; Art subjects include History, Latin, Philosophy, Literature, Greek, History of Art; Other subjects include TPS, Physical Education, Religion/Ethics, Music, Drawing.

Table 8: Quantile Estimation using Equation (5.2): Coefficients of (Year_1990) by Levels

	Percentiles								
	10	20	30	40	50	60	70	80	90
All Levels	0.375 [0.104]***	0.3 [0.081]***	0.231 [0.111]**	0.19 [0.082]**	0.2 [0.081]**	0.178 [0.088]**	0.254 [0.084]***	0.218 [0.081]***	0.33 [0.090]***
Level 1	0.75 [0.300]**	0.56 [0.182]***	0.36 [0.184]*	0.35 [0.222]	0.36 [0.201]*	0.338 [0.200]*	0.333 [0.166]**	0.675 [0.200]***	0.57 [0.156]***
Level 2	-0.2 [0.156]	-0.025 [0.151]	0.05 [0.140]	0.14 [0.142]	0.0375 [0.170]	-0.08 [0.171]	0.02 [0.169]	0.075 [0.126]	-0.107 [0.136]
Level 3	0.336 [0.134]**	0.06 [0.182]	0.117 [0.162]	0.0333 [0.164]	-0.01 [0.178]	0.225 [0.161]	0.0333 [0.178]	0.225 [0.181]	0.14 [0.235]
Level 4	0.688 [0.316]**	0.638 [0.276]**	0.464 [0.268]*	0.525 [0.227]**	0.463 [0.130]***	0.455 [0.207]**	0.313 [0.252]	0.545 [0.281]*	0.337 [0.290]

Notes: * denotes significance at the 10% level, ** denotes significance at the 5% and *** denotes significance at the 1% level.

Table 9: Effect on Performance (Grades) by Ability

	All Levels		Level 2		Level 3		Level 4	
Constant	6.044	4.988	5.739	5.721	5.393	5.247	5.373	5.183
	[0.071]***	[0.092]***	[0.075]***	[0.083]***	[0.087]***	[0.099]***	[0.140]***	[0.153]***
Trend	-0.042	0.043	0.091	0.092	0.034	0.04	-0.066	-0.066
	[0.012]***	[0.013]***	[0.014]***	[0.014]***	[0.016]**	[0.016]**	[0.023]***	[0.023]***
Year 1990	**0.194**	**0.254**	**-0.178**	**-0.188**	**0.144**	**0.169**	**0.603**	**0.65**
	[0.077]**	**[0.069]*****	**[0.118]**	**[0.119]**	**[0.125]**	**[0.124]**	**[0.197]*****	**[0.196]*****
Above*(t-1)*	1.504	1.441	1.833	1.819	1.992	1.978	2.131	2.182
	[0.057]***	[0.054]***	[0.066]***	[0.068]***	[0.072]***	[0.074]***	[0.103]***	[0.108]***
Above*(t-1)Year 90**	**0.048**	**-0.081**	**0.123**	**0.128**	**0.066**	**0.04**	**-0.01**	**-0.058**
	[0.111]	**[0.099]**	**[0.163]**	**[0.163]**	**[0.181]**	**[0.180]**	**[0.284]**	**[0.283]**
Level 2		0.975						
		[0.047]***						
Level 3		0.489						
		[0.043]***						
Girl		0.24		0.051		0.231		0.211
		[0.062]***		[0.061]		[0.066]**		[0.098]**
Repeater		0.445		-0.108		-0.057		0.464
		[0.076]***		[0.153]		[0.129]		[0.188]**
Observations	2152	2152	777	777	771	771	604	604

Notes: * denotes significance at the 10% level, ** denotes significance at the 5% and *** denotes significance at the 1% level. Estimations using All Levels are done using random effects (RE) estimation and OLS estimation is used when we estimate each level separately. The number of observations falls when using All Levels as we must restrict analysis to students in Levels 2 to 4.

Table 10: Differential in Grade Group Transition between 1990 and all other years					
		Grade Group*(t)*			
		A	**b**	**c**	**D**
Grade Group*(t-1)*	**a**	22.25	-16.76	-5.26	-0.23
	b	8.37	3.23	-12.54	0.94
	c	-0.50	0.16	11.68	-11.34
	d	0.00	-1.30	17.46	-16.16

Notes: The Grade Groups are classified as: (a) 8 and 10, (b) 7 and 7.9, (c) 5 and 6.9 and (d) 1.5 and 4.9. Each cell computes the difference between Year 1990 grade group and the grade group of the rest of the years. This table is derived using the transition rate table in Table A2 in the Appendix.

Table 11: Effect on Dispersion of Performance (Variance of Grades)										
	All Levels		Level 1		Level 2		Level 3		Level 4	
Constant	2.267	2.344	2.073	2.206	1.667	1.679	1.561	1.563	1.477	1.671
	[0.145]***	[0.151]***	[0.169]***	[0.186]***	[0.194]***	[0.211]***	[0.210]***	[0.223]***	[0.295]***	[0.315]***
Trend	0.03	0.031	-0.14	-0.14	-0.022	-0.021	0.087	0.086	0.231	0.235
	[0.019]	[0.019]*	[0.032]***	[0.032]***	[0.035]	[0.035]	[0.036]**	[0.036]**	[0.048]***	[0.048]***
Year 90	**-0.067**	**-0.103**	**-0.331**	**-0.647**	**0.319**	**0.725**	**0.124**	**0.019**	**-0.479**	**-0.625**
	[0.140]	**[0.195]**	**[0.264]**	**[0.354]**	**[0.240]**	**[0.345]***	**[0.239]**	**[0.327]**	**[0.393]**	**[0.566]**
Level 1	-0.679	-0.678								
	[0.134]***	[0.134]***								
Level 2	-0.782	-0.783								
	[0.135]***	[0.135]***								
Level 3	-0.544	-0.543								
	[0.135]***	[0.135]***								
Girl	-0.124	-0.109	0.071	0.087	-0.192	-0.18	-0.331	-0.335	-0.044	0.003
	[0.093]	[0.094]	[0.157]	[0.158]	[0.171]	[0.171]	[0.171]*	[0.172]*	[0.250]	[0.251]
Repeater	-0.304	-0.371	0.852	0.762	-0.309	-0.353	-0.108	-0.106	-1.405	-1.57
	[0.202]	[0.205]*	[0.420]**	[0.425]	[0.434]	[0.444]	[0.330]	[0.335]	[0.444]***	[0.454]***
Above(*t*)		**-0.178**		**-0.273**		**-0.038**		**0.001**		**-0.466**
		[0.100]		**[0.167]**		**[0.188]**		**[0.186]**		**[0.265]**
Above(*t*) *Yr90		**0.068**		**0.676**		**-0.778**		**0.229**		**0.302**
		[0.279]		**[0.531]**		**[0.479]**		**[0.480]**		**[0.784]**
Observations	3414	3414	943	943	879	879	857	857	735	735

Notes: * denotes significance at the 10% level, ** denotes significance at the 5% and *** denotes significance at the 1% level. Estimations using All Levels are done using random effects (RE) estimation and OLS estimation is used when we estimate each level separately. The dependent variable is the variance of the average grade.

Table 12: Lasting Effect on Performance (Grades)

	Level 1	Level 2	Level 3	Level 4
Constant	6.720	6.960	6.545	6.079
	[0.0452]	[0.0496]***	[0.0594]***	[0.0812]***
Year 1990	0.594***	0.0234	0.0722	0.614
	[0.145]	[0.122]	[0.134]	[0.190]***
Year 1991		**-0.0546**	**0.0911**	**0.300**
		[0.138]	**[0.135]**	**[0.179]***
Year 1992			**-0.234**	**-0.300**
			[0.146]	**[0.170]***
Year 1993				**0.222**
				[0.176]
Observations	943	879	857	735

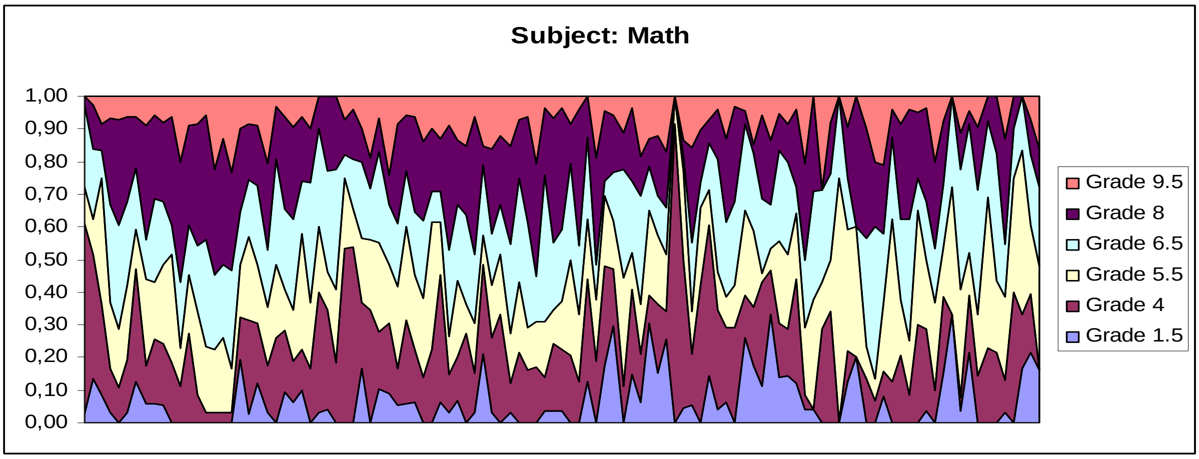
Notes: * denotes significance at the 10% level, ** denotes significance at the 5% and *** denotes significance at the 1% level. All regressions are estimated using OLS. The dummy variables *Year1991*, *Year1992* and *Year1993* denote lasting effect. To understand the lasting effect one should read the table diagonally. First, a student who was treated in Level 1 in 1990 will be in Level 2 in 1991, in Level 3 in 1992 and in Level 4 in 1993. Second, a student who was treated in Level 2 in 1990 will be in Level 3 in 1991 and in Level 4 in 1992. Third, a student who was treated in Level 3 in 1990 will be in Level 4 in 1991.

Table 13: Aggregate Effect on Performance (Grades): Selectividad Grades

	L4 Grade	L4 Grade	Select.	Select.
Constant	6.127	5.924	4.921***	4.990
	[0.115]***	[0.136]**	[0.0816]***	[0.0961]***
Trend	-0.005	0.004	0.143	0.147
	[0.022]	[0.022]	[0.0160]***	[0.0156]***
Year 1990	**0.585**	**0.558**	**0.694**	**0.635**
	[0.183]*	**[0.181]***	**[0.118]***	**[0.114]***
Girl		0.368		-0.025
		[0.115]***		[0.0788]
Repeater		-0.655		-0.947
		[0.205]***		[0.152]***
Observations	735	735	583	583

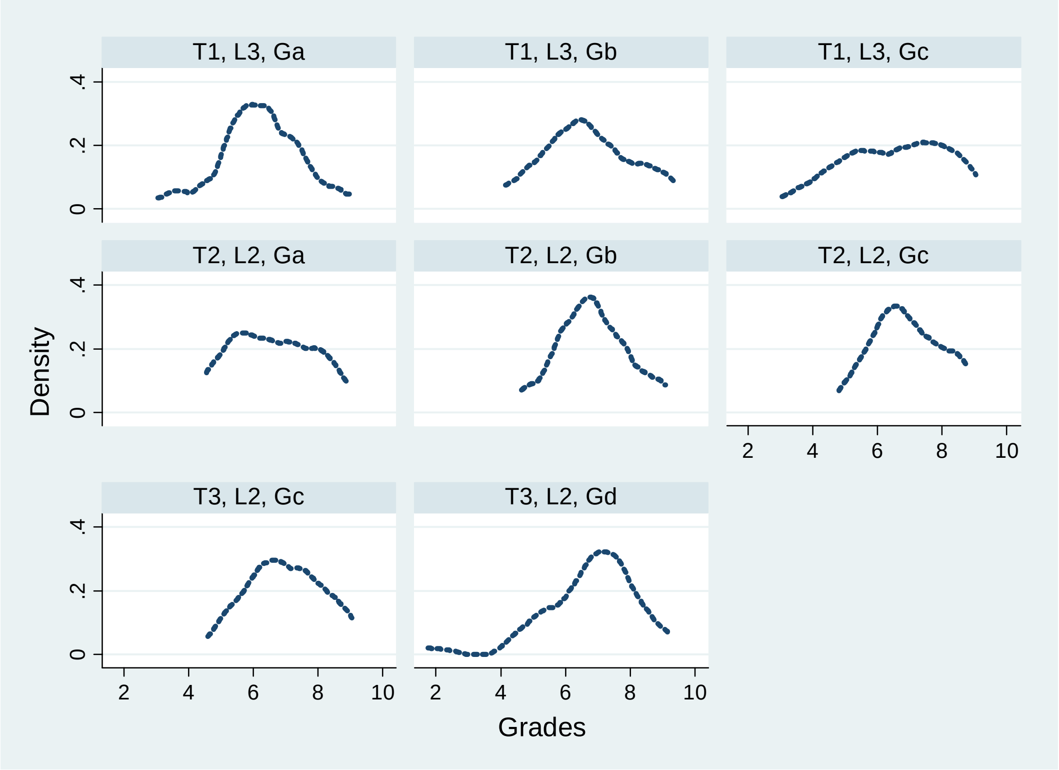
Notes: Data on students' grades are provided by Oiartzo Ikastola (school), for academic years 1986-1995. Data on students' Selectividad grades (for the same students) are provided by the University of Basque Country, for academic years 1986-1995. * denotes significance at the 10% level, ** denotes significance at the 5% and *** denotes significance at the 1% level. All regressions are estimated using OLS. The analysis is restricted to Level 4 students.

Figure 1a: Distribution of Grades for Math by Group-Year



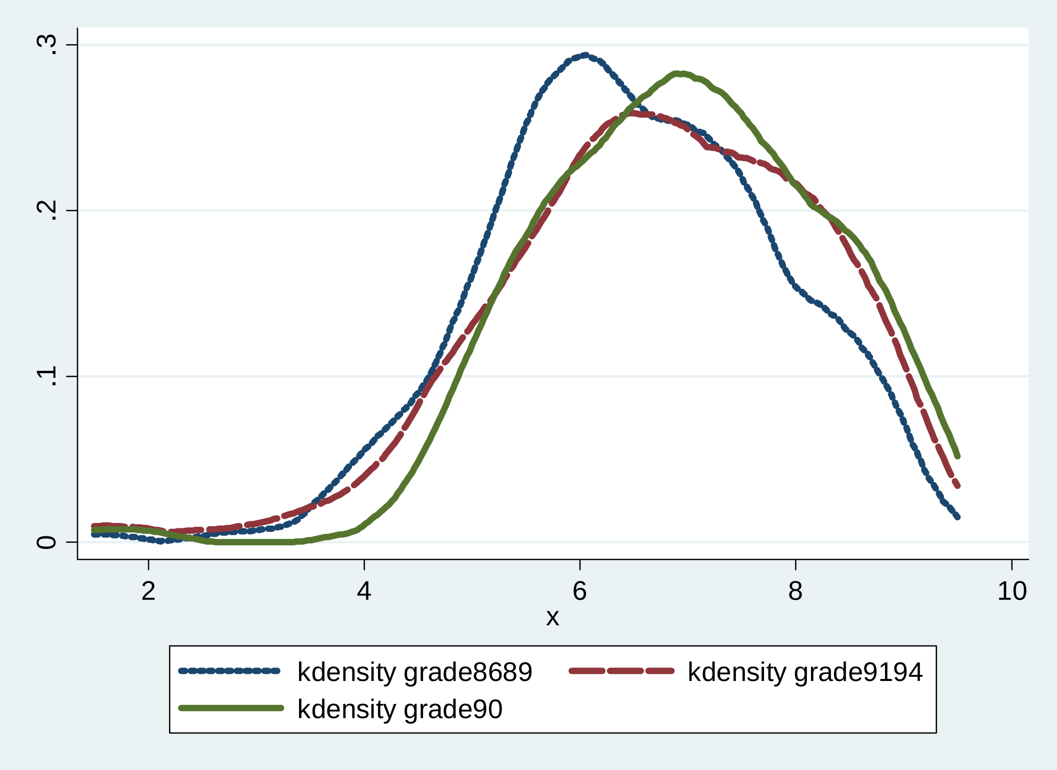
Notes: Each vertical bar represents the distribution of grades in a particular group and academic year, and the shaded regions show the proportion of students that obtain each possible grade (1.5, 4, 5.5, 6.5, 7 and 9.5).

Figure 1b: Examples of Distribution of Grades for Math by Teacher, Level and Group



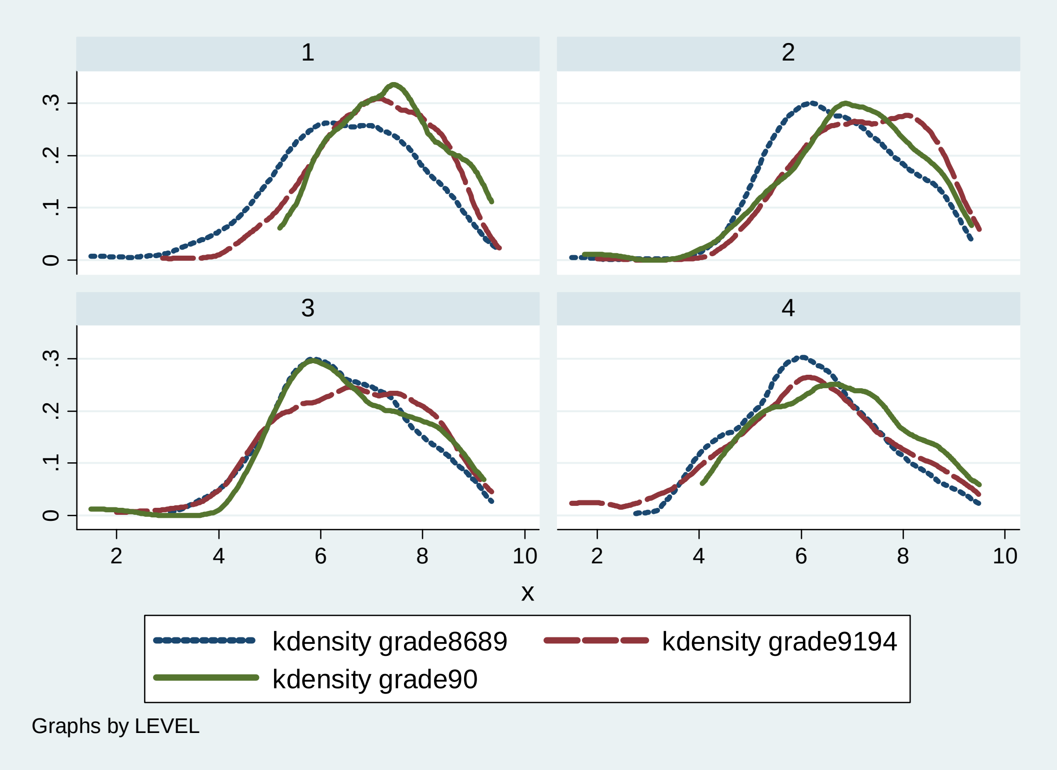
Notes: Each distribution represents the grades given by a one teacher (T) at a certain level (L) and group (G). The first distribution represents teacher “T1”, teaching Maths to Level 3 “L3” and Group A “Ga”. The other two figures in the row represent the same teacher and level but different groups. The other rows give two further examples.

Figure 2: Kernel Distribution Before, During and After the Treatment



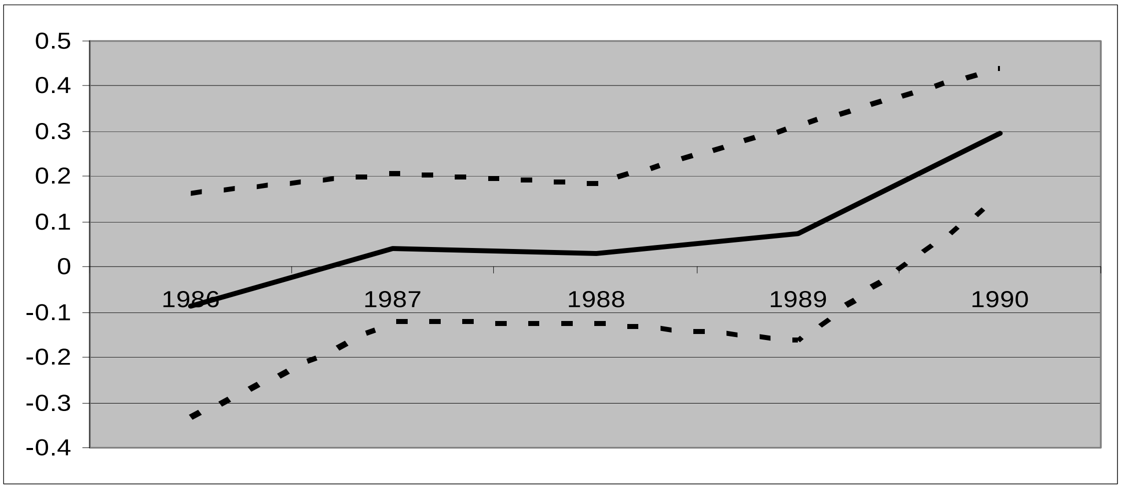
Notes: Kernel distribution of students' grades, over all students. The filled line represents the treatment year (1990)

Figure 3: Kernel Distribution Before, During and After the Treatment (By Level)



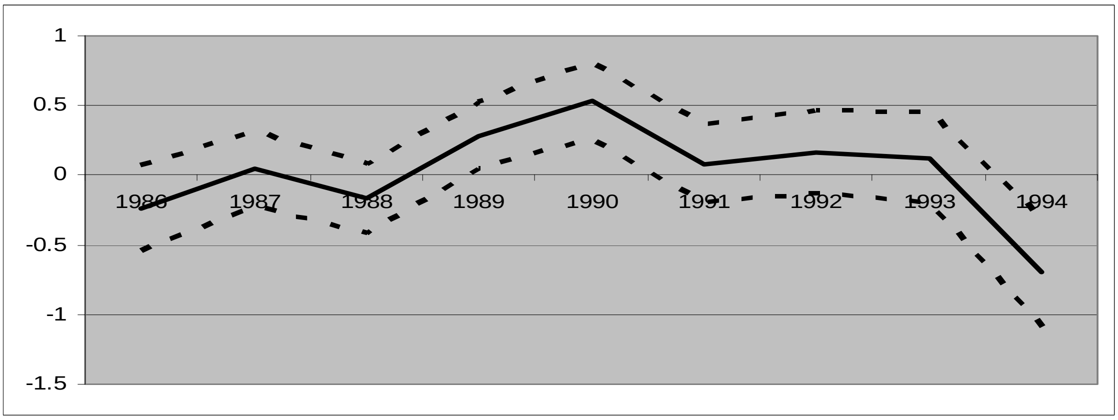
Notes: Kernel distribution of students' grades, by level. The filled line represents the treatment year (1990)

Figure 4: Placebo Treatment (Years Before)



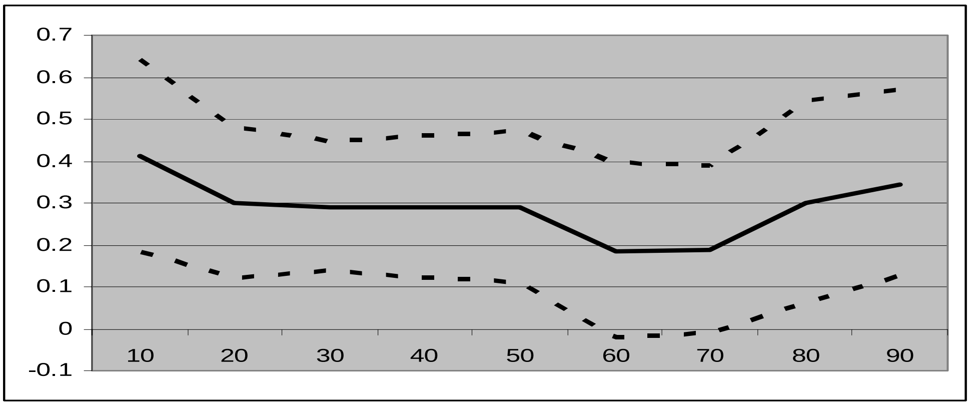
Notes: Equation (5.1) using data for years 1986-1989 and for comparison the figure includes the effect in the treatment year (1990). The dotted line represents a 95% confidence interval.

Figure 5: Placebo Treatment Using Only Level-1 (All Years)



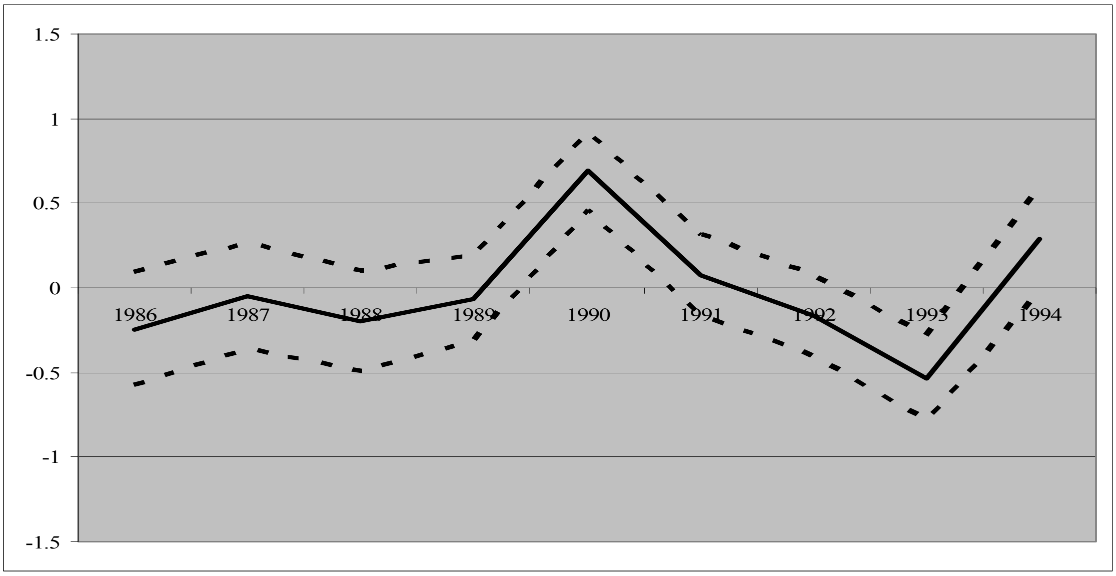
Notes: 5 (1) using data Level 1 data for all years (excluding treatment year (1990)) and for comparison the figure includes the effect in the treatment year (1990). The dotted line represents a 95% confidence interval.

Figure 6: Quantile Regression for Equation (5.1).



Notes: 10% to 90% quantile. The dotted line represents a 95% confidence interval. At the 60th quantile the average grade is 6.3 and this is close to the average.

Figure 7: Placebo Treatment Using Selectividad Grades



Notes: Equation (5.1) using data Selectividad data for all years (excluding treatment year (1990)) and for comparison the figure includes the effect in the treatment year (1990). The dotted line represents a 95% confidence interval.