Discussion Paper Series

Graduate School of Economics and School of Economics

Meisei University

Discussion Paper Series,
No.40
September, 2019

Information Feedback in Relative Grading: Evidence from a Field Experiment

Shinya Kajitani
(Kyoto Sangyo University)
Keiichi Morimoto
(Meisei University)
Shiba Suzuki
(Seikei University)

Hodokubo 2-1-1, Hino, Tokyo 191-8506
School of Economics, Meisei University

Phone: +81-(0)42-591-9479 Fax: +81-(0)42-599-3024
URL: https://keizai.meisei-u.ac.jp/econ/
E-mail: keizai@econ.meisei-u.ac.jp

Information Feedback in Relative Grading: Evidence from a Field Experiment*

Shinya Kajitani† Keiichi Morimoto‡ Shiba Suzuki§

September 26, 2019

Abstract

Previous studies have revealed the role of relative performance information feedback on providing agent incentives under a relative rewarding scheme through laboratory experiments. This study examines the impact of relative performance information feedback of students' performance on their examination score under the relative grading scheme in an actual educational environment. Conducting a randomized controlled trial in a compulsory subject at a Japanese university, we show that the relative performance information feedback has a significantly positive impact on the students' examination score *on average*, but that the average positive impact is derived by the improvement of low-performing students.

Keywords: Education; experiments; intention-to-treat effects; relative performance information feedback; relative grading.

JEL Classification: D81, I21

*This paper is a revised version of our paper, "Information Feedback in Relative Grading: Evidence from a Field Experiment," Discussion Paper Series no.40, Graduate School of Economics and School of Economics, Meisei University, July 2018. This experiment was approved by Meisei University's research ethics committee on the Use of Human Subjects (Application No. H26-002). Before conducting the studies, we obtained informed consent from all subjects. We have not received financial support for this research from any outside agency or foundation. We all contributed to the design of the experiments, and the writing of the paper. Kajitani and Suzuki conducted the actual experiments.

†Corresponding author. Associate Professor, Faculty of Economics, Kyoto Sangyo University, Motoyama, Kamigamo, Kita-ku, Kyoto-city 603-8555, Japan. Phone number: +81-75-705-1739. E-mail address: kajitani@cc.kyoto-su.ac.jp

‡Associate Professor, Department of Economics, Meisei University, 2-1-1, Hodokubo, Hino-city, Tokyo 191-8506, Japan. Phone number: +81-42-591-6964. E-mail address: keiichi.morimoto@meisei-u.ac.jp

§Professor, Faculty of Economics, Seikei University, 3-3-1, Kichijoji-kitamachi, Musashino-city, Tokyo 180-8633, Japan. Phone number: +81-422-37-3571. E-mail address: shiba.suzuki@econ.seikei.ac.jp

I Introduction

Does relative performance information feedback improve a student's incentive to study under a relative grading scheme? Many consider information feedback associated with a reward environment as an efficient way of eliciting the incentives of students to study. "Relative grading" or "grading on a curve" is widely used in grading students even though its use remains controversial.[1] Conducting a randomized controlled trial in a compulsory subject required for university graduation, we examine the impact of relative performance information feedback on students' examination scores.

In relative grading, a student's grade depends on her position in the class score distribution. To understand student incentives in a relative grading scheme, Becker and Rosen (1992) extend the rank-order tournament model of Lazear and Rosen (1981) and emphasize that the student's learning effort depends on her position in the distribution of academic attainment. This suggests that relative performance information feedback affects student decision-making in providing effort.[2] Besides, in actual schooling environments, multiple examinations typically grade students. It is then worth considering the relationship between information on a student's relative position in the distribution of earlier examination scores and her incentive to provide study effort for the following examination.[3]

How then does this relative performance information feedback affect the students' incentive to study under a relative grading scheme in a multiple examinations environment? In a relative grading scheme, to obtain a better grade a student needs to receive a higher score than her opponents do. That is, an opponent's score serves as a threshold she must exceed. In this grading environment, the relative performance information feedback is then a signal of the effort she should provide. For example, when relative performance information feedback tells the student that her current score is relatively low, she understands that she has to provide a higher level of effort to rise above the

[1] Volokh (2015) argues that teachers value relative grading as a means to control grade inflation and to ensure students have an opportunity to receive a higher grade, even if examinations make it difficult to obtain high marks. Conversely, Grant (2016) points out that relative grading can be problematic in that it prevents students from collaborating owing to the overly competitive environment.

[2] From a theoretical viewpoint, Paredes (2016) and Andreoni and Brownback (2017) construct a theoretical model of relative grading employing an all-pay auction.

[3] Aoyagi (2010) and Ederer (2010) theoretically analyze information feedback in a dynamic tournament context.

threshold. Conversely, she may give up, saving the cost of effort.

Revealing the role of relative performance information feedback in providing agent incentives in the relative rewarding environment is often through laboratory experiments. These studies generally suggest that there is no guarantee that the relative performance information feedback has a positive impact on the students' incentive to study. For example, Eriksson et al. (2009) and Freeman and Gelber (2010) conclude that relative performance information feedback lowers the performance of subjects whose interim performance is relatively low. However, those subjects whose midterm performance is relatively high do not slacken off. In contrast, Ludwig and Lünser (2012) examine the effects of effort information in a two-stage rank-order tournament. They demonstrate that laboratory subjects who lead tend to lower their effort, but those who lag increase it relative to the first stage, while the subjects who lead exert a greater effort than those who lag. Thus, the impact of relative performance information feedback may vary according to the initial level of attainment.[4]

In an actual educational environment, previous studies focus on the impact of relative performance information feedback on student incentives under absolute grading. For example, Azmat and Iriberri (2010), using data from Spanish high schools, and Tran and Zeckhauser (2012), in a field experiment of Vietnamese university students, demonstrate that relative performance information feedback raises the performance of students when rewarded absolutely. Both these studies argue that if students have competitive preferences, which means that they inherently prefer receiving a higher rank than others, relative performance information has a positive impact on their incentive to study on average.

By contrast, no existing study examines the impact of relative performance information feedback on student incentives under relative grading in an actual educational environment. The question is whether relative performance information feedback improves student examination scores in an actual relative grading environment where students sit for examinations on multiple occasions.

To examine this issue, we conduct a field randomized controlled trial employing the

[4]Azmat and Iriberri (2016) and Gill et al. (2019) conduct a laboratory experiment to examine the impact of relative performance information feedback on subject performance when rewards are absolute. In particular, Gill et al. (2019) find that the rank–response function is U-shaped, that is, subjects increase their effort most in response to relative performance information feedback when they are ranked either first or last.

compulsory subject of economics at a Japanese university. In this course, after students sat two examinations (the midterm examination and the final examination), instructors calculated the students' final raw scores mainly by taking the weighted average of their two examination scores. However, the students' grades were evaluated by grading on a curve, with the instructors adjusting the students' final raw scores subject to the entire final raw score distribution to obtain the reasonable pass rate. In our experiment, we allocated more than 200 students into a control group and a treatment group immediately following the midterm examination. We only provided students in the treatment group with feedback on their midterm examination relative performance and explored the impact of this feedback on student performance in the final examination.

This study is, to our knowledge, the first attempt to investigate the impact of relative performance information feedback on student incentives to study in an actual educational environment encompassing relative grading. We show the significant positive impact of relative performance information feedback on the students' final examination scores *on average*. Note that because students cannot graduate from the university unless they receive credit in this subject, they care about whether they can receive credit. In other words, the threshold between a pass and a fail in the course is significant for students. Moreover, the threshold depends not only on the students' ranks in the distribution of the final raw scores but also on their final raw scores *per se*. By considering the threshold, we demonstrate that the average positive impact on the final examination scores is through the improvement of low-performing students in the midterm examination.

The remainder of the paper is organized as follows. Section II describes the experimental design. Section III presents the empirical framework and reports the estimation results and Section IV concludes.

II Experimental Design

Description of the randomized trial This section provides details of the randomized trial, performed using first-year students in an economics department at a Japanese private university. We begin by describing the flow of interventions in the experiments, which are displayed in Figure 1. The academic year comprised first and second semesters: the first semester began in April 2012 and ended in July 2012; the second semester began in September 2012 and ended in January 2013. We conducted a mathematical achieve-

ment test (referred to as the Pretest of Mathematics) immediately following university entrance. Students enrolled in two compulsory introductory economics courses in their first year: Economics I in the first semester and Economics II in the second semester. In Economics I and II, we administered midterm and final examinations to grade students. While the midterm and final examinations in Economics I were in May and July 2012, those in Economics II were in November 2012 and January 2013. We note that the score for the Pretest of Mathematics was independent of the grades for Economics I and II. The dotted vertical lines in Figure 1 represent the timing of the examinations.

We evaluated the students in both Economics I and II using the same grading scheme, namely, grading on a curve. The instructors explained this grading scheme in detail to the students in Economics II at the beginning of the second semester. We provide details of the grading on a curve scheme later.

We divided students into four classes. According to their score in the Pretest of Mathematics, we placed all students with a top-40 score in one small class. Hereafter, we refer to this as Classroom 1. The designation of Classroom 1 is for purely educational purposes. For example, in teaching economics, we used a different level of mathematics in Classroom 1 and the other classrooms. We then randomly allocated the remaining students to the other three classes. Hereafter, we refer to these as Classrooms 2, 3 and 4. We fixed all class enrollments and instructors across both semesters. While each class had its own instructors, such that all classes were held in the third period (12:55 p.m.–14:25 p.m.) on Wednesday, all students took the same examination using a multiple-choice computer-scored answer sheet at the same time.

[Insert Figure 1 here]

The randomized controlled trial was conducted immediately after the midterm examination in Economics II, which then randomly assigned all students to the treatment or control group. In the first class time after the midterm examination, we handed students letters revealing their score for the midterm examination. In addition, the letters given to students in the treatment group also reported their ranks in the midterm examination. We did not include this information in the letters to the students in the control group. The student letter content is similar to that used by Ashraf et al. (2014). Figures O1 and O2 in the online supplementary material reproduce the information provided to the students in the treatment and control groups. On this basis, while students in the

treatment group knew their precise rank, students in the control group would only have a vague awareness.

There are two points to note in our randomized control trial. First, we exclude some students who did not receive the letter regarding the midterm examination from our sample. Because some students were absent from the class time just after the midterm examination, we could not hand letters to these students. Therefore, we do not consider these students as subjects in our experiment. Second, our experimental design cannot exclude the possibility that some students may have exchanged their rank information. Because our experimental design is similar to that of Tran and Zeckhauser (2012) save the grading scheme, we share the same problem that students in both the control and treatment groups sit in the same classroom, making the exchange of rankings a genuine possibility. However, it would be generally difficult for a student in the control group to identify a student in the treatment group with exactly the same score; students in the treatment group know their exact rank, while students in the control group do not. Therefore, the impact of relative performance information feedback, if any, is captured by our experimental design. We discuss this further in Section III.1.

The grading scheme In Economics II, a final raw score was calculated as follows: perfect raw scores were 110 points, in which 100 points were for the two examinations (the midterm and final examinations) and the remaining 10 points for the number of homework submissions. The examination score of 100 is divided into "40% of the midterm examination score" *and* "60% of the final examination score." "The number of homework submissions" is based on 10 homework assignments, each worth one point.

It should be noted that the instructors can adjust students' final raw score upward considering the entire final raw score distribution. This upward adjustment introduces uncertainty into threshold scores between one grade and another, especially the threshold score between pass and fail. The official university's guidelines recommend that instructors should assign the first grade (S) to scores 90 and over, the second grade (A) to scores between 80 and 89, the third grade (B) to scores between 70 and 79, the fourth grade (C) to scores between 60 and 69, and fail (F) to scores below 60. A student who marks F fails the course. However, when the instructors give grades to the students afterward, they can adjust students' final raw scores upward depending on the entire final raw score distribution. Meanwhile, none of the students ever knows whether the instructors will

adjust their final raw scores in advance of taking the final examination. For example, because the average final raw score in Economics I was quite low (52.7 points out of the perfect score of 100), the instructors decided to add 9 points to all the students' final raw scores. This upward adjustment made the student whose final raw score was above 50 receive a credit for Economics I. In contrast, because the instructors did not finally adjust the students' final raw score in Economics II, only the students whose final raw score was 60 and over got credit for Economics II.

There are three points to note in our grading scheme. First, students in Economics II knew the grading scheme described above. This is because the instructors already explained this grading scheme in detail at the beginning of the second semester and this grading scheme had already been employed in Economics I. Second, instructors grade all students in the four classrooms using the same grading criteria in Economics II. Because the grading criteria are common to the four classes, students compete not only with students in their classroom but also the other classrooms; whether students pass or fail will depend on their relative position in the entire score distribution of more than 200 students. The students in Economics II were exposed to the uncertainty of the threshold they must exceed to pass the course. Finally, in our experiment, when the instructors gave grades to the students afterward, they may well adjust the students' final raw scores upward but never adjusted their final raw scores downward. Under these circumstances, a student whose final raw score was over 60 points got credit for Economics II. If the student who already got 60 points in the midterm examination gets at least 60 points in the final examination, she can get credit for Economics II. That is, whether students pass or fail depends not only on the students' rank in the distribution of the final raw scores but also on their final raw scores *per se*.

Balance between the control and treatment groups Table 1 provides the total number of students and the means and standard deviations of the midterm examination scores in Economics II for the control and treatment groups. Table 1 also shows how we randomly divided these students into the control and treatment groups. In total, 284 students took midterm examinations, and their mean score was 49.57. We randomly divided these students into control and treatment groups. However, some students failed to receive the letter. Consequently, in our experiment, there are 255 subjects with a mean score of 50.67. There were 130 and 125 students in the control and treatment groups,

respectively. The mean scores for the control and treatment groups are 51.48 and 49.82, respectively, and there is no significant difference in the mean scores between the control and treatment groups, as shown in row (a) in Panel B.

[Insert Table 1 here]

One point to note is the differences between classrooms. Table 1 also shows that we randomly divided students into the control and treatment groups if we consider these differences. The mean score in the midterm examination in Classroom 1 is much higher than that in Classrooms 2–4 because we enrolled students with a top-40 mark in the Pretest of Mathematics in Classroom 1. The number of students who received the letter in Classrooms 2–4 is 215, while that in Classroom 1 is 40. The mean for students who received the letter in Classroom 1 is 65.48 and that in Classrooms 2–4 is 47.92. There is no significant difference in the mean scores between the control and treatment groups across Classrooms 2–4. While the number of students and the mean for the control group is 106 and 48.21, respectively, those for the treatment group are 109 and 47.63, respectively. We do not reject the null hypothesis that "the mean values of the two groups are not different," as shown in row (b) in Panel B. In addition, as for Classroom 1, the number of students for the control and treatment groups are 24 and 16, respectively, and the mean scores for the control and treatment groups are 65.96 and 64.75, respectively. We again do not reject the null hypothesis that "the mean values of the two groups are not different," as shown in row (c) in Panel B.

III Empirical design and results

III.1 Intention-to-treat effects on the final examination scores

The randomized controlled trial means that we obtain two groups that are statistically equivalent to each other. However, we note that 11 students who took the midterm examination did not take the final examination. By taking these students into account, we capture the intention-to-treat (ITT) effects of relative performance information feedback on the final examination scores, that is, the average effects of assignment to the treatment group versus assignment to the control group. We employ the following empirical

framework:

$$Y_{Fi}^{*} = \alpha D_i + X_i\beta + \epsilon_i, \tag{1}$$

where Y_{Fi}^{*} denotes the latent scores in the final examination for student i. D_i is a dummy variable equal to one if student i is given information on her relative rank in the midterm examination (i.e., the student is in the treatment group), and zero if student i is not given this information (i.e., the student is in the control group). $\boldsymbol{X}_i$ denotes the covariates including a constant term. ϵ_i are disturbances, which we assume are distributed $N(0, \sigma^2)$. The observed scores in the final examination Y_{Fi} are related to the latent scores Y_{Fi}^{*} through the observation rule: we treat $Y_{Fi}^{*} = Y_{Fi}$ when student i took both the midterm examination and the final examination, and $Y_{Fi} = 0$ when student i took the midterm examination but did not take the final examination.[5]

[Insert Table 2 here]

Table 2 provides descriptive statistics for all the variables used in this estimation model.[6] In our experiment, we randomly assigned all students to the treatment or the control group using a random number generator. While associated covariates, the midterm examination scores and the classrooms are also randomly selected, we are concerned about ex post differences in the values of the midterm examination scores. We include the midterm examination scores for the student i and dummy variables for students in different classrooms (the classroom fixed effects) in the vector $\boldsymbol{X}_i$.

[Insert Table 3 here]

Columns (1a) and (1b) in Table 3 report the results of estimating equation (1).[7] The marginal effects are for the left-censored mean. The coefficient for D is significantly

[5] As Barnett et al. (2005) argue, randomized experiments can reduce the effect of the regression to the mean (RTM). As the responses from both the control and treatment groups are equally affected by the RTM, the differences between the treatment group and the control group, that is, the coefficients for D_i, comprise the treatment effect after adjusting for the RTM.

[6] We exclude a student whose midterm score was revised from the sample.

[7] We do not report robust standard errors which are clustered at classroom levels. In our experiment, there is no cluster in the population of interest not represented in the sample. Moreover, clustered standard errors with only a few clusters (e.g., just four classrooms in our experiment), could not be reliable (Angrist and Pischke, 2008). There is no need to adjust standard errors for clustering once fixed effects are included (Abadie et al., 2017).

positive. The magnitude of the marginal effect, evaluated at the left-censored mean, is 3.768. This indicates that the scores in the final examination for students who received information on their relative rank in the midterm examination are 3.768 points higher on average than the scores for students who did not get this information.

While the students in Classroom 1 are those with a top-40 score in the Pretest of Mathematics, the students in Classrooms 2 through 4 were randomly assigned to the classrooms. In Columns (2a) and (2b), we exclude students in Classroom 1 from our baseline sample. The coefficient for D_i is significant (Column (2a)) and the magnitude of the marginal effect is 4.079 (Column (2b)). These values are slightly larger than shown in Columns (1a) and (1b).

In terms of other research considerations, such as the experimental design employed by Tran and Zeckhauser (2012), we divided students into control and treatment groups within each classroom. Because D_i was randomly assigned, ITT effects have a causal interpretation. However, the ITT effects tell us the causal effect of the offer of treatment, including the fact that some of those offered have shared their ranks with their classmates, even if it is difficult for a student in the control group to identify a student in the treatment group with precisely the same score. This leads to the suggestion that the ITT effects may be small relative to the average causal effects on those in fact treated.

III.2 The heterogeneity in the ITT effects due to high- or low-performance in the midterm examination

In our experiment, the higher a student's midterm examination score, the lower the required score in the final examination for her to receive credit in Economics II. These indicate that the rank for students with a higher (lower) midterm examination score has less (more) tangible benefits. If so, the relative performance information feedback could affect high- or low-performing students differentially.

[Insert Figure 2 here]

We can visually observe the heterogeneity in the ITT effects when depicting a violin plot of the final examination scores by group in Figure 2. When based on the median scores in the final examination, the kernel density for the treatment group is narrower than that for the control group. This suggests that the relative performance information

feedback may raise effort for students whose midterm examination scores were relatively lower, while the feedback may decrease effort for students whose midterm examination score were relatively higher. That is, the relative performance information feedback could affect high- or low-performing students differentially.

To examine the heterogeneous ITT effects, we consider the following equation:

$$Y^*_{Fi} = \alpha_1 D_i + \alpha_2 (D_i \times H_i) + \alpha_3 H_i + X_i \beta + \epsilon_i, \tag{2}$$

where H_i is a dummy variable equal to one if student i's performance in the midterm examination was relatively high. We set the threshold between high and low performances on the midterm examination to 60 points, where 60 is the *ex ante* threshold score whether the students pass or fail the course.

[Insert Table 4 here]

Table 4 reports the estimation results for the heterogeneity in the ITT effects on the final examination scores. The relative performance information feedback for low-performing students on the midterm examination indeed has a positive impact on their performance. As shown in Columns (1a) and (1b), the coefficient for D_i is significantly positive and the magnitude of the marginal effect evaluated at the left-censored mean is 5.782. Even when excluding students in Classroom 1, the coefficient for D_i is significant (Column (2a)). The final exam scores for low-performing students who received information on their relative rank in the midterm examination are 5.504 points higher on average than the scores for low-performing students who did not receive this same information (Column (2b)).

In contrast, the coefficient for the interaction term $D_i \times H_i$ in Column (1a) is significantly negative, while that in Column (2a) (where the students in Classroom 1 are excluded from the baseline sample) is insignificant. Interestingly, we do not reject the null hypothesis that "the sum of the coefficients for $D_i + (D_i \times H_i)$ equals the coefficients for H_i" in Columns (1a) and (2a). This indicates that, as for high-performing students, there is no significant difference between the final examination scores for students who *received* their rank information for the midterm examination and the scores for students who *did not receive* their rank information. That is, the relative performance information feedback does not have significant impact on high-performing students.[8]

[8]When students also have an inherent preference for a high rank, as Tran and Zeckhauser (2012)

In our experiment, we confirm the following key finding: relative performance information feedback indeed exploits the incentive to study of low-performing students in the midterm examination, but it has no significant impact on high-performing students.

IV Concluding Remarks

Our experimental results demonstrate that relative performance information feedback has a positive impact on a student's examination score in a relative grading environment where she takes examinations multiple times. In particular, this positive impact is indeed significant for low-performing students in the previous examination. As emphasized in Becker and Rosen (1992), a student's position in the distribution of academic attainment is crucial in relative grading. Because this subject was compulsory for graduation from the university, the threshold between a pass and a fail in the course would be significant for students.

Moreover, the threshold between a pass and a fail in the course depended not only on the students' rank in the distribution of the final raw scores but also on their final raw scores *per se*. The higher midterm examination score decreased the required final examination score for which she was able to get the credit. Under these circumstances, the rank for students with higher (lower) midterm examination scores would have less (more) tangible benefits.

Our results suggest that the relative performance information is beneficial in incentivizing students to study in binary grade environments (e.g., where a pass or a fail matters). In particular, for students with intermediate midterm examination scores, it is much more beneficial to inform them of their performance ranking. Therefore, from an examiner or policymaker perspective, relative performance information feedback under a relative grading scheme is favorable.

---

discuss, the relative performance information feedback could positively affect high-performing students more than low-performing students, even if the rank for the high-performing students has less tangible benefits. To examine whether being informed of ranks motivates students with higher midterm examination scores, using Tran and Zeckhauser (2012)'s model specifications, we additionally include the interaction term $D_i \times Y_{Mi}$ in equation (1). As shown in Table O1 in the online supplementary material, the estimated coefficient for the interaction term $D_i \times Y_{Mi}$ is significantly negative in Column (1a) and insignificant in Column (2a). Given the situation where the rank for high-performing students has a less tangible benefit, it is not confirmed that the relative performance information feedback positively affects the high-performing students more than the low-performing students, even if students have an inherent preference for a high rank.

Acknowledgements

We are very grateful to Akira Yamazaki, Kentaro Kobayashi, Hayato Nakata, and Masahiro Watabe for invaluable advice. We also thank Naohito Abe, Kosuke Aoki, David Gill, Tao Gu, Shigeki Kano, Vu Tuan Kai, Nobuyoshi Kikuchi, Jun-Hyung Ko, Colin McKenzie, Akira Miyaoka, Koyo Miyoshi, Tomoharu Mori, Chang-Min Lee, Masao Nagatsuka, Kengo Nutahara, Fumio Ohtake, Daniela Puzzello, David Reiley, Masaru Sasaki, Michio Suzuki, Kan Takeuchi, Ryuichi Tanaka, Tomoaki Yamada and participants at annual meeting on Japanese Economic Association 2013 Spring Meeting (Toyama, Japan), Economic Science Association European meeting 2015 (Heidelberg, Germany), and seminar participants at Meiji University, Osaka University, Seikei University, and the University of Tokyo for their helpful comments.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. (2017). When should you adjust standard errors for clustering? *National Bureau of Economic Research Working Paper* 24003.

Andreoni, J., and Brownback, A. (2017). All pay auctions and group size: Grading on a curve and other applications. *Journal of Economic Behavior and Organization*, 137: 361-373. https://doi.org/10.1016/j.jebo.2017.03.017

Angrist, J. D., and Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*, Princeton university press.

Aoyagi, M. (2010). Information feedback in a dynamic tournament. *Games and Economic Behavior* 70(2): 242–260. https://doi.org/10.1016/j.geb.2010.01.013

Ashraf, N., Bandiera, O., and Lee, S. S. (2014). Awards unbundled: Evidence from a natural field experiment. *Journal of Economic Behavior and Organization* 100: 44–63. https://doi.org/10.1016/j.jebo.2014.01.001

Azmat, G., and Iriberri, N. (2010). The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics* 94(7): 435–452. https://doi.org/10.1016/j.jpubeco.2010.04.001

Azmat, G., and Iriberri, N. (2016). The provision of relative performance feedback: An analysis of performance and satisfaction. *Journal of Economics and Management Strategy* 25(1): 77–110. https://doi.org/10.1111/jems.12151

Barnett, A. G., Van Der Pols, J. C., and Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology* 34(1): 215–220. https://doi.org/10.1093/ije/dyh299

Becker, W. E., and Rosen, S. (1992). The learning effect of assessment and evaluation in high school. *Economics of Education Review* 11(2): 107–118. https://doi.org/10.1016/0272-7757(92)90002-K

Ederer, F. (2010). Feedback and motivation in dynamic tournaments. *Journal of Economics and Management Strategy* 19(3): 733–769. https://doi.org/10.1111/j.1530-9134.2010.00268.x

Eriksson, T., Poulsen, A., and Villeval, M. C. (2009). Feedback and incentives: Experimental evidence. *Labour Economics* 16(6): 679–688. https://doi.org/10.1016/j.labeco.2009.08.006

Freeman, R. B., and Gelber, A. M. (2010). Prize structure and information in tournaments: Experimental evidence. *American Economic Journal: Applied Economics* 2(1): 149–164. https://doi.org/10.1257/app.2.1.149

Gill, D.,Z. Kissováv, J. Lee, and V. Prowse (2019) "First-place loving and last-place loathing: How rank in the distribution of performance affects effort provision." *Management Science* 65(2): 494–507. https://doi.org/10.1287/mnsc.2017.2907

Grant, A. (2016). Why we should stop grading students on a curve. *The New York Times*, accessed January 6, 2017. Available from URL: https://www.nytimes.com/2016/09/11/opinion/sunday/why-we-should-stop-grading-students-on-a-curve.html?smid=tw-shareand_r=0

Lazear, E. P., and Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy* 89(5): 841–864. https://doi.org/10.1086/261010

Ludwig, S., and Lünser, G. K. (2012). Observing your competitor: The role of effort information in two-stage tournaments. *Journal of Economic Psychology* 33: 166–182. https://doi.org/10.1016/j.joep.2011.09.011

Paredes, V. (2016). Grading system and student effort. *Education, Finance and Policy.* 12(1): 107–128. https://doi.org/10.1162/EDFP_a_00195

Tran, A., and Zeckhauser, R. (2012). Rank as an inherent incentive: Evidence from a field experiment. *Journal of Public Economics* 96(9): 645–650. https://doi.org/10.1016/j.jpubeco.2012.05.004

Volokh, E. (2015). In praise of grading on a curve. *The Washington Post*, accessed March 6, 2017. Available from URL: https://www.washingtonpost.com/news/volokh-conspiracy/wp/2015/02/09/in-praise-of-grading-on-a-curve/?utm_term=.63bfc5beb916

Table 1: Confirmation of randomness

Panel A. Descriptive statistics of the midterm examination scores

		1 Classrooms 1–4	2 Classrooms 2–4	3 Classroom 1
I. All	Obs.	284	244	40
	Mean	49.57	46.96	65.48
	S.D.	17.36	16.24	15.54
II. Receive a letter	Obs.	255	215	40
	Mean	50.67	47.92	65.48
	S.D.	17.02	15.85	15.54
-i. Control	Obs.	130	106	24
	Mean	51.48	48.21	65.96
	S.D.	18.64	17.70	15.84
-ii. Treatment	Obs.	125	109	16
	Mean	49.82	47.63	64.75
	S.D.	15.18	13.90	15.56

Panel B. Mean-comparison test (Welch t-test)

		t-value	P-value
(a)	Comparison (1, II-i) with (1, II-ii)	0.781	0.435
(b)	Comparison (2, II-i) with (2, II-ii)	0.264	0.792
(c)	Comparison (3, II-i) with (3, II-ii)	0.239	0.813

Table 2: Descriptive statistics

Variable	Definition	Mean	Std. Dev.	Min	Max
Total (Obs.=254)					
Y_F	Score in the final examination	63.807	20.341	0	100
D	= 1 if student is given information on her relative rank in the midterm examination, =0 if elsewhere	0.488	0.501	0	1
Y_M	Score in the midterm examination	50.602	17.020	12	102
H	= 1 if $Y_{Mi} \geq 60$, = 0 if elsewhere	0.291	0.455	0	1
$Class1$	= 1 if in the classroom 1 (math class), = 0 elsewhere	0.157	0.365	0	1
$Class2$	= 1 if in the classroom 2, = 0 elsewhere	0.295	0.457	0	1
$Class3$	= 1 if in the classroom 3, = 0 elsewhere	0.264	0.442	0	1
Treatment (Obs.=124)					
Y_F		64.927	17.372	0	92
D		1	0	1	1
Y_M		49.677	15.154	12	95
H		0.250	0.435	0	1
$Class1$		0.129	0.337	0	1
$Class2$		0.306	0.463	0	1
$Class3$		0.266	0.444	0	1
Control (Obs.=130)					
Y_F		62.738	22.833	0	100
D		0	0	0	0
Y_M		51.485	18.642	18	102
H		0.331	0.472	0	1
$Class1$		0.185	0.389	0	1
$Class2$		0.285	0.453	0	1
$Class3$		0.262	0.441	0	1

Note: Because we exclude a student whose midterm score was revised from the sample, our final sample comprised 254 students.

Table 3: Estimation results: the ITT effects of relative performance information feedback on the final examination scores

Dependent variables	Y_M			
Students in $Class1$	Included		Excluded	
	(1a) Coeff.	(1b) M.E.	(2a) Coeff.	(2b) M.E.
D	3.772*	3.768*	4.088*	4.079*
	(2.079)	[2.077]	(2.345)	[2.339]
Y_M	0.773***	0.772***	0.800***	0.798***
	(0.067)	[0.067]	(0.075)	[0.075]
$Class1$	-3.116	-3.113		
	(3.521)	[3.516]		
$Class2$	-4.795*	-4.790*	-4.840*	-4.827*
	(2.726)	[2.722]	(2.831)	[2.822]
$Class3$	-4.793*	-4.788*	-4.951*	-4.937*
	(2.830)	[2.825]	(2.943)	[2.933]
$Constant$	25.791***		24.381***	
	(3.787)		(4.185)	
Obs.	254		214	
Left-censored obs.	11		11	
Mcfadden's R^2	0.057		0.051	

Notes:
1) *, ** and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively.
2) Figures reported in brackets are standard errors.
3) Figures reported in square brackets are delta-method standard errors for the marginal effects.
4) The marginal effects in Columns (1b) and (2b) are computed for the left-censored mean.
5) Because we exclude a student whose midterm score was revised from the sample, our final sample comprised 254 students.

Table 4: Estimation results: the heterogeneity in the ITT effects

Dependent variables	Y_M			
Students in $Class1$	Included		Excluded	
	(1a) Coeff.	(1b) M.E.	(2a) Coeff.	(2b) M.E.
D	5.786**	5.782**	5.515**	5.504**
	(2.450)	[2.448]	(2.673)	[2.668]
$D \times H$	-8.058*	-8.045*	-8.348	-8.315
	(4.589)	[4.575]	(5.585)	[5.546]
H	-3.425	-3.422	-4.139	-4.129
	(4.459)	[4.454]	(5.280)	[5.263]
Y_M	0.906***	0.905***	0.944***	0.942***
	(0.103)	[0.102]	(0.115)	[0.115]
$Class1$	-2.141	-2.139		
	(3.489)	[3.486]		
$Class2$	-3.879	-3.876	-3.805	-3.797
	(2.736)	[2.733]	(2.855)	[2.848]
$Class3$	-3.803	-3.800	-3.829	-3.820
	(2.828)	[2.825]	(2.955)	[2.947]
$Constant$	19.373***		17.849***	
	(4.834)		(5.324)	
F test H_0: the sum of the coeff. for D and $D \times H$ equals the coeff. for H	0.03		0.02	
Obs.	254		214	
Left-censored obs.	11		11	
Mcfadden's R^2	0.060		0.054	

Notes:
1) *, ** and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively.
2) Figures reported in brackets are standard errors.
3) Figures reported in square brackets are delta-method standard errors for the marginal effects.
4) The marginal effects in Columns (1b) and (2b) are computed for the left-censored mean.
5) Because we exclude a student whose midterm score was revised from the sample, our final sample comprised 254 students.

Figure 1: The flow of interventions in the experiment

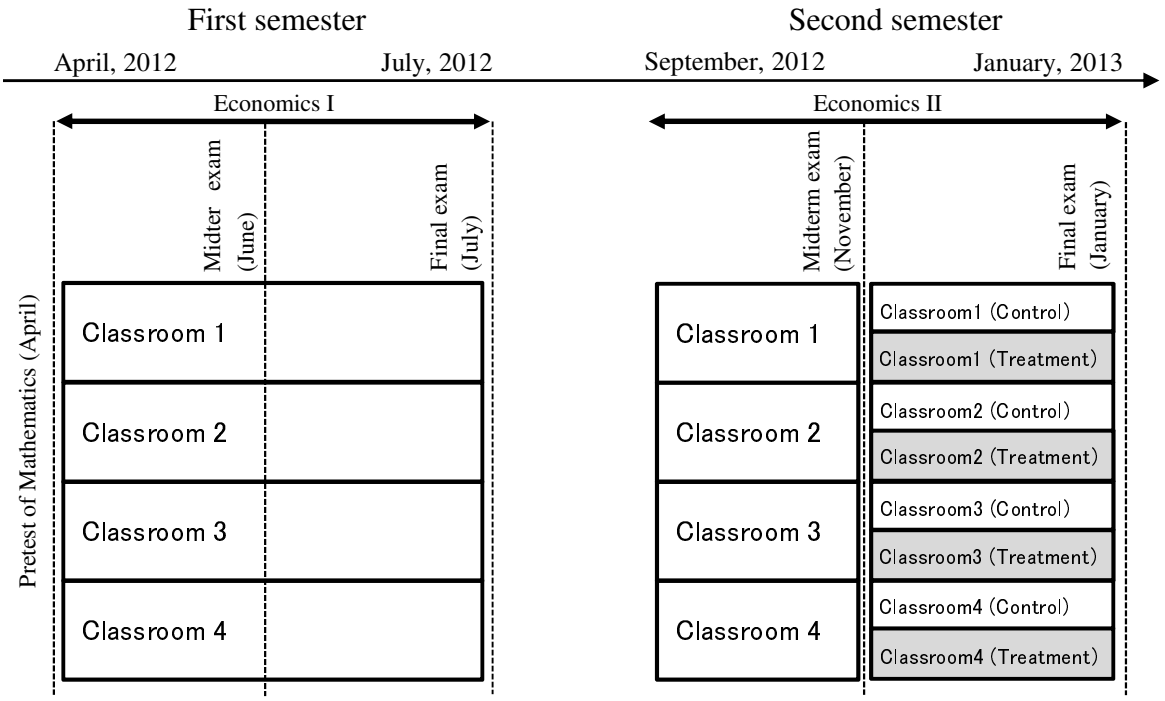
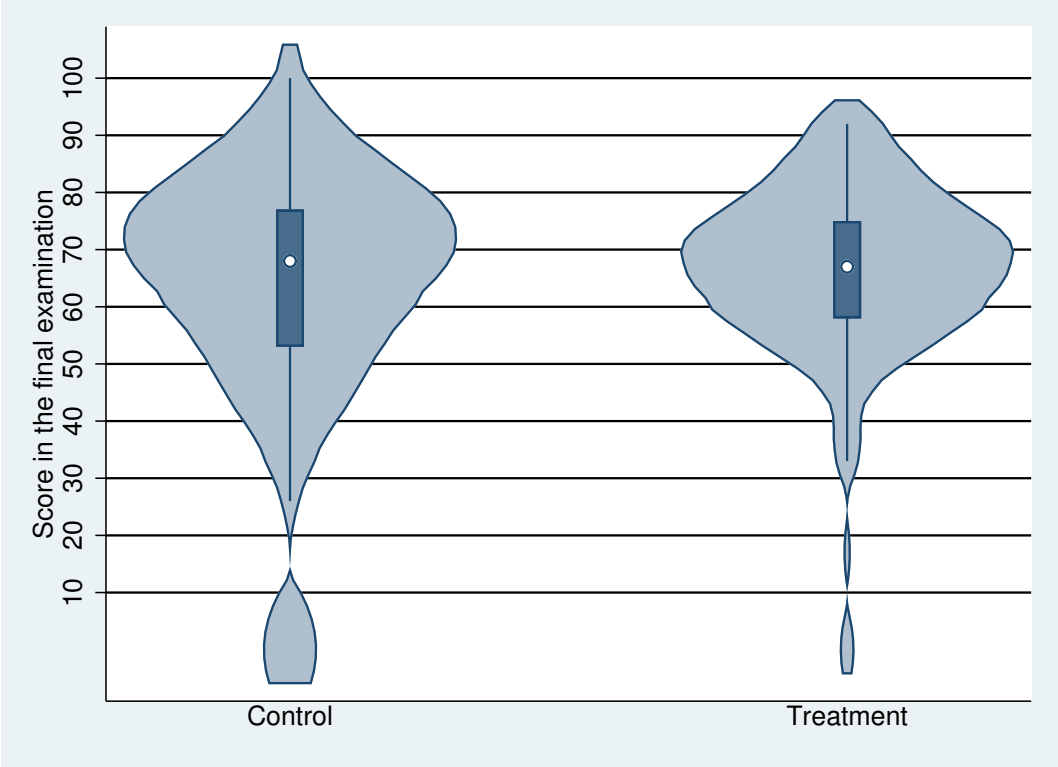


Figure 2: The violin plot of the final examination scores



Note: The violin plot comprises a combination the box plot and the density trace. This includes a marker for the median scores of the final examination, a box indicating the interquartile range and spikes extending to the upper- and lower-adjacent values, and plots of the estimated kernel density.

Online Supplementary Figure O1: The letter to students (treatment group)

Introductory Economics, Second Semester, 2012
A report on the result of your midterm examination

Student ID **7 digit ID** Name **Name of the student**
Classroom **Instructor's name**

Your score in the midterm examination is **49**/110
Of them, your score in problems of mathematics is **6**/10

Within four classrooms, you are **146th** out of 285 students.

Original (written in Japanese)

2012 年度経済学通論 2　中間テスト結果

学籍番号　12E1-XXX　　氏名　XXXX
経済学通論 2 クラス（XXXX）

中間テストの点数　　49 点／110 点満点
うち　数学出題分　　6 点／10 点満点

学年全体でのあなたの順位　　285 人中　146 位

Online Supplementary Figure O2: The letter to students (control group)

Introductory Economics, Second Semester, 2012
A report on the result of your midterm examination

Student ID **7 digit ID** Name **Name of the student**
Classroom **Instructor's name**

Your score in the midterm examination is **49**/110
Of them, your score in problems of mathematics is **6**/10

Original (written in Japanese)

2012 年度経済学通論 2　中間テスト結果

学籍番号　12E1-XXX　　氏名　XXXX
経済学通論 2 クラス（XXXX）

中間テストの点数　　49 点／110 点満点
うち　数学出題分　　6 点／10 点満点

Online Supplementary Table O1: Estimation results: the heterogeneity in the ITT effects (Tran and and Zeckhauser (2012)'s model specifications)

Dependent variables	Y_M			
Students in $Class1$	Included		Excluded	
	(1a) Coeff.	(1b) M.E.	(2a) Coeff.	(2b) M.E.
D	14.322**	14.306**	12.628	12.593*
	(6.652)	[6.640]	(7.680)	[7.650]
$D \times Y_M$	-0.209*	-0.209*	-0.179	-0.178
	(0.125)	[0.125]	(0.153)	[0.153]
Y_M	0.852***	0.851***	0.868***	0.866***
	(0.082)	[0.082]	(0.095)	[0.095]
$Class1$	-2.978	-2.975		
	(3.504)	[3.499]		
$Class2$	-4.909*	-4.903*	-4.941*	-4.928*
	(2.712)	[2.708]	(2.824)	[2.815]
$Class3$	-4.589	-4.584	-4.787	-4.774
	(2.818)	[2.814]	(2.938)	[2.928]
$Constant$	21.654***		21.088***	
	(4.515)		(5.042)	
Obs.	254		214	
Left-censored obs.	11		11	
Mcfadden's R^2	0.058		0.051	

Notes:
1) *, ** and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively.
2) Figures reported in brackets in Columns (1a) and (2a) are standard errors.
3) Figures reported in square brackets are delta-method standard errors for the marginal effects.
4) The marginal effects in Columns (1b) and (2b) are computed for the left-censored mean.
5) Because we exclude a student whose midterm score was revised from the sample, our final sample comprised 254 students.