Does active learning improve student performance? A randomized experiment in a Chilean university

Alcalde, Pilar and Nagel, Juan

Universidad de los Andes, Chile

6 November 2015

Does active learning improve student performance?

A randomized experiment in a Chilean university

Running head: Active learning experiment in Chile

Pilar Alcalde*

Assistant Professor, Facultad de Ciencias Económicas y Empresariales, Universidad de los Andes, Av. Monseñor Álvaro del Portillo 12455, Las Condes, Santiago, Chile.

palcalde@uandes.cl. +56 2 2618 1626

Juan Nagel

Adjunct Associate Professor, Facultad de Ciencias Económicas y Empresariales, Universidad de los Andes, Av. Monseñor Álvaro del Portillo 12455, Las Condes, Santiago, Chile.

jnagel@uandes.cl. +56 2 2618 1833.

The authors wish to thank the Universidad de los Andes' Research Fund for financial assistance. Leonardo Epstein and Katherine Strasser provided useful comments. Pablo Zamorano and Patricia Leal provided helpful research assistance. All remaining errors are our own.

* Corresponding author.

ABSTRACT

We study the causal effect of an active learning teaching method on grades. We designed a randomized experiment with students at an undergraduate business and economics program in Chile. Two groups were taught by the same professor: the control group used traditional lectures, while the treatment group used an active learning method. Treated students failed the class less but the effect was not significant. They also had significantly better grades at the end and during the semester. The treatment effect was larger for males and students with high application scores. The effect does not appear instantaneously, and appears to fade away at the end of the semester. Results suggest students allocate effort differently across both groups, and this interacts with the treatment effect.

Keywords: Classroom experiments, course performance, peer instruction, innovation in teaching. (JEL: A20, C21, C90)

Throughout the world, traditional lecture-style methods of teaching are being replaced with a variety of methods that emphasize student engagement in their learning. Despite the enthusiasm in these innovations, the new strategies are costly to implement in terms of resources and instructor time. Drastic changes such as these should be accompanied by appraisals of their effect on student performance and, ultimately, learning. This paper is an attempt to contribute to this literature.

Active learning methods, a particular brand of innovation, have aroused much enthusiasm. An active learning method is any instructional method that engages students in their own learning

process through activities and/or discussions in class, as opposed to passively listening to an expert. (Prince, 2004; Freeman et al., 2014). This broad definition includes many class strategies and activities used in economics and other fields, as flipped classrooms, classroom experiments and games, and peer instruction, among others.

Various authors from different fields have attempted to measure active learning's impact. Freeman et al. (2014) summarize these findings in a meta-analysis which examines 225 studies comparing student performance in traditional and active learning courses. They found that student performance in active learning groups was on average 6% higher than in traditional groups, and that students in traditional groups were 1.5 times more likely to fail a given course than students in active learning groups.

However, the studies used in this meta-analysis may not be sufficient to draw conclusions on the causal effect of active learning methods. The studies focused on many different forms of active learning[1] and included numerous non-experimental or quasi-experimental studies. They frequently used different instructors for treatment and control groups, and/or used students who had selected into their sections, and/or used students from different semesters, some of whom may have failed courses differently beforehand. These issues may prevent from identifying the causal effect of active learning methods on student performance.

Of the literature reviewed in the meta-analysis, we found sixteen studies where the instructor in both the control and treatment groups was the same. These studies were from a variety of fields, but none from economics: biology (Paschal 2002, Knight and Wood 2005, Armstrong et al. 2007, Walker et al. 2008, Carmichael 2009); chemistry (Williamson and Rowe 2002, Bilgin 2006, Bilgin et al. 2009); mathematics (Lovelace and McKnight 1980, Keeler and Steinhorst 1994, Giraud 1997); physics (Moelter et al. 2005); computer science (Hurley 2002); psychology

(Lawson et al. 2006); and engineering (Van Dijk et al. 2001, Pandy et al. 2004). Six of them studied large or medium groups, while ten of them looked at small groups. Although the average size of the treatment effect in the subgroup was 0.45 SDs (similar to the meta-analysis as a whole), only four of the sixteen studies found a positive and significant treatment effect.

Fourteen of these studies were classified as quasi-experiments, with the issues described above - generally, using students from different semesters for the treatment and control groups. Only two of the studies are random experiments. Lovelace and McKnight (1980) is the paper most similar to ours: they evaluate a peer tutoring method to foster mathematical reading skills by implementing an experiment using two randomly-assigned classes with the same instructor. Pandy et al. (2004) designed an experiment to evaluate a single multimedia-based learning module in a biomechanics course. Both studies find positive but non-significant treatment effects.

The economics literature on the effectiveness of different teaching methodologies has mostly focused on evaluating online learning (e.g., Brown and Liedholm 2002, Figlio et al. 2013, Green 2014, Olitsky and Cosgrove 2014, Joyce et al. 2014) and on specific classroom games (e.g., Brouhle 2011, Cartera and Emersona 2012, Valcarcel 2013, among many others). Our study is more (loosely) related to the first literature: their findings so far suggest that live and/or hybrid classes are more effective than online courses, but they leave open the question of how to make live classes more effective. Other studies (Ghosh and Renna 2009, Bergstrom 2009, Salemi 2009, Roach 2014, Emerson et al. 2015) report increased student satisfaction in economics courses using interactive methods similar to the one used in this paper, but they use non-experimental techniques. Our conclusion is that there is a need for experimental studies on the effects of active learning on student performance.

EXPERIMENTAL DESIGN

This paper aims to fill this gap by reporting on an experiment in which students were randomly assigned to sections taught by the same professor using two distinct methodologies. We implemented this experiment in two of six sections of Applied Algebra - a course specifically designed for and taught exclusively to first-semester students in the Business and Economics program at the Universidad de los Andes (Chile) - during the (Southern hemisphere) fall of 2014.[2] A subsample of students from the regular admissions process were randomly assigned into each of the two sections, stratified by their gender and application score - low or high relative to the mean application score of the incoming class.[3]

Several features of this course make it ideal for a randomized experiment. As a first semester math course, students have not yet self-selected by advancing the curriculum at different paces, they are given their schedule and assigned into sections by the School administration, and we can control for their prior math knowledge via the nationwide application exam.

The same instructor (one of the authors) taught the two sections using two different methods.[4] The control group had a "traditional" lecture-style class. In it, the instructor used a variety of visual aids - including Power Point presentations. He solved exercises on the board, and received unprompted questions from the students. Students were encouraged to ask questions during lecture, and to solve problems outside of it. Students were not required to read before lecture, and according to a survey at the end of the semester very few of them did.

The treatment group used a mix of "active learning" strategies. Students were required to read and complete a short online quiz before lecture.[5] The instructor would start lecture with a brief

summary of the main topics students had read about. Then, students would receive exercises to solve first individually and then with a classmate. Students would answer using some mobile device, which provided the instructor with real-time information on student performance.[6]

Both groups covered the same content. The final grade consisted of three quizzes, two midterm exams, and a final exam. All evaluations except for one – quiz 2 – were the same for both groups; quiz 2 was delivered in two different days so it was similar but not the same. Additionally, the midterms and final exams were the same for the six Algebra sections, including those taught by other instructors.

All evaluations except for the final exam were graded by a team of teaching assistants. A single teaching assistant was responsible for grading each question across both groups. The instructor graded the final exam for both groups.

We use two statistical tests to check that the random assignment indeed worked. First, we use a t-test on the equality of mean characteristics (gender and application score) between both groups; it is not possible to reject the null of equal means at the 10% significance level. Second, we use a logit model to test if any of the observed characteristics correlate with assignment to the treatment group; the model is not globally significant (p-value = 0.1439), and none of the variables is significant at the 10% level.

REGRESSION RESULTS

We focus on the percentage of students that failed the course and on their final grade; however, looking at the results of different evaluations during the semester allows us to observe when differences between the groups begin to appear or disappear.

[Insert Table 1 about here.]

Table 1 provides summary statistics for the outcome variables for both groups. In previous semesters, failures rates for this course have hovered between 20 and 40 percent, and grades range from 1 to 7, with 4 as the passing grade. Our sample size is 78 students, with 39 students in each group; administrative regulations prevented us from increasing the sample size. Attrition is very low: only 3 students are missing the final grade, 2 of whom quit the course and 1 who missed the final exam.

Students in the treatment group do better on average than students in the control group in all available outcomes. Table 1 shows that the proportion of students who failed the class is 12.8 percentage points lower in the treatment group, but this difference is not significant. The average final grade is 0.29 points larger in the treatment group, and significantly different. The differences between the treatment and the control group in the average grades of individual evaluations range from 0.1 points in midterm 1 and the final exam to 0.6 points in midterm 2 and 0.76 points in quiz 3; only the differences for quiz 3 and midterm 2 are significant.

These differences show a distinct pattern when looked at chronologically. The difference between the two groups is small and insignificant at the beginning of the semester, then it increases and becomes statistically significant toward the middle of the semester, and finally it decreases again and practically disappears for the final exam.

[Insert Table 2 about here.]

Table 2 presents summary statistics by individual characteristics. In general, males fail more than females, although they have similar average grades. Students with low application scores did worse than students with high application scores in every outcome: they fail more and have

lower average grades. Students in the treatment group do better on average than students in the control group in all available outcomes for all types of students, but the difference is larger for males and for students with high application scores.

The difference in failure rates between both groups is 22 percentage points for males and 20 percentage points for students with high application scores. The difference in the final grade between both groups is 0.35 points for males and 0.31 points for students with high application scores. All differences are significant at the 5% level. This larger difference is also observed in every evaluation during the semester, not reported in Table 2.

We measure the effect of the treatment on the probability of failing the course and on grades by using a logit model and a linear regression model respectively, controlling for individual characteristics in both cases (gender and application scores). Table 3 presents these results.

[Insert Table 3 about here.]

The treatment decreases the probability of failing by 13.4 percentage points, but it is not significant. It also increases the final grade significantly by 0.24 points. This effect corresponds to a 0.45 standard deviation increase in the final grade, which is quite large. The treatment has a positive effect in all evaluations. The effect ranges from 0.07 points in the final exam to 0.64 points in quiz 3. It is significant only for quiz 3 and midterm 2, both given in the second half of the semester.

We add interactions of the treatment status with each individual characteristic, to test if the treatment effect varies by student characteristics. Table 4 shows the effect of the treatment on the probability of failing the course. The treatment effect is negative, and it is larger for males (-25 percentage points) and for students with high application scores (-20.6 percentage points for a

student with 681 points, the average, and -26.3 percentage points for a student with 700 points). The effect is significant at the 5% level for students with 700 points, and at the 10% level for males and students with 681 points.

[Insert Table 4 about here.]

Table 5 shows the effect of the treatment on the final grade. The treatment effect is positive, and it is larger for males (0.31 points or 58.7% SD) and for students with high application scores (0.23 points or 43.7% SD for a student with 681 points, and 0.27 points or 51.3% SD for a student with 700 points). All these results are significant at the 5% level. The effect of the treatment for females and students with low application scores is not different from zero.

[Insert Table 5 about here.]

ROBUSTNESS CHECKS

The results highlighted above are robust to variations in the models. Failing to include individual controls, or using probit or linear probability models for the failure rate do not significantly affect the results. We also use other checks, as detailed below. All these robustness checks increase the estimated treatment effect, which suggests that our baseline case is conservative.

Our first robustness check recognizes the panel structure of the data: instead of taking the data as six independent evaluations, we consider that each student's performance is observed six times, from quiz 1 to the final exam. We employ the following panel data model:

$$Y_{ij} = \alpha + \theta T_i + \beta X_i + \varepsilon_i + \varepsilon_j + u_{ij}$$

where the grade in evaluation *j* for student *i*, Y_{ij}, depends on the student's characteristics X_i, the treatment status T_i, an individual unobserved effect ε_i, an evaluation unobserved effect ε_j, and an *iid* zero-mean error u_{ij}. We run 4 different panel data regressions: (1) fixed effects only for the evaluations, i.e., $\varepsilon_i = 0$ and ε_j a fixed-effect[7]; (2) random effects only for the individuals, i.e., ε_i follows a normal distribution and $\varepsilon_j = 0$; (3) random effects for the individuals and fixed effects for the evaluations; and (4) random effects for the individuals and fixed effects for the evaluations, with treatment effects that vary by evaluation θ_j.[8]

The conclusions from our analysis are essentially unchanged. Table 6 shows that the treatment effect is between 0.28-0.29 points when only the final grade is considered; it is significant at the 1% level and larger than before (see Table 3). When the treatment effect varies by evaluation we observe the same pattern than in the cross-sectional analysis: the treatment effect is always positive, but it is larger and significant in the middle of the semester.

[Insert Table 6 about here.]

For our second robustness check, we use as dependent variable the weighted grade before the exam to avoid any possible biases caused by the instructor having graded the exam. Because the effect of the treatment disappears toward the end of the semester, it is not surprising to find the results are in the same direction as those found when using the original final grade, but larger. These are shown in Table 7. The treatment effect increases to 0.34 points or 64% SD. The treatment effect is strongest for males and for students with high application scores, just as before.

[Insert Table 7 about here.]

For our third robustness check, we look more closely at grades in midterm 2. Students on average did far worse on midterm 2 than on midterm 1. We compare the grades from both midterms with the results from the four other sections not included in the experiment to see if the control group's performance in the second midterm was an outlier.[9]

The results are shown in Table 8. All groups did better on average in midterm 1 than in midterm 2, but while the control group and the sections not included in the experiment dropped their performance substantially, the fall in the treatment group's average grade was smaller. This suggests the treatment helped them perform better than they otherwise would have.

[Insert Table 8 about here.]

DISCUSSION AND CONCLUSIONS

Despite the small sample size, we find significant treatment effects on the final grade. The active learning method had larger effects on males and on students with high application scores, and the effect was larger during the middle of the semester.

Our estimate of the treatment effect of 0.45 standard deviations is similar to the 0.47 standard deviations reported in the Freeman et al. (2014) meta-analysis. Our estimates are conservative considering the larger results of our robustness checks.

Our estimates can also be viewed as conservative because of the effect the treatment may have had on student effort. Students in an active learning classroom were required to study constantly throughout the semester, and they may have felt they did not need as much study prior to the midterms or the exam. Likewise, students in the control group may have perceived they were

underperforming relative to their peers in the treatment group, and this could have prompted them to double their effort prior to the midterms and the exam. We do not observe effort, but we present three measures related to effort levels across both groups.

First, we look at attendance. Students in the treatment group attended class slightly less (91.8% attendance rate in the treatment group and 92.6% in the control group), but the difference was not significant. No clear pattern emerges when looking at weekly attendance: some weeks the treatment group has larger attendance, and some weeks the control group does.

Second, we have an anonymous survey the students completed at the end of the semester. Among other things, we asked them about the number of hours they studied for the course. Self-reported hours of study are surely measured with error, although a priori we cannot expect the measurement error to differ between groups. Treated students report studying less frequently in groups before an evaluation. Regarding the overall hours they studied for each evaluation on the days prior, both groups report studying the same for quizzes - 3.3 hours on average - but treated students report studying fewer hours for the midterms (7.9 hours vs. 8.8 hours on average for the control group). This difference is not statistically significant.

Third, we look more closely at the final exam. We found the treatment effect completely disappears in the final exam. One hypothesis is that student effort responds to the possibility of failing the course. We look at each student's weighted grade prior the exam,[10] and see how it correlates to the grade in the exam.

[Insert Figure 1 about here.]

In Figure 1 we plot the grade in the exam and the weighted grade before the exam. Students above the 45 degree line obtained a higher grade in the exam than their weighted grade up to that point. We draw a vertical line at 4, which is the passing grade for the class.

Most students in the treatment group tend to be concentrated in the area to the right of the vertical line and below the 45 degree line; this means that treated students have "passing" weighted grades but they perform worse in the exam. Students with weighted grade below the passing grade are mostly from the control group: most of them performed better in the exam than their weighted grade suggests. This helps explain the lack of treatment effect in the final exam, and is consistent with the hypothesis that students in the control group put in higher levels of effort in the exam in order to pass the course.

Taken together, these three measures suggest that the treatment affected effort not through attendance but because students seem to be shifting time away from Algebra into other courses or leisure. We hope future research on the relationship between active learning and effort outside of class sheds more light on this.

Finally, we check if the treatment had an effect on subsequent courses. Of the 78 students in the experiment, 60 passed Algebra and enrolled in Calculus in the spring semester of 2014: 30 were from the control group and 30 from the treatment group.[11] We regress the final grade in Calculus on the treatment dummy, plus individual covariates and a Calculus-instructor fixed effect. The effect of the treatment on the final grade in Calculus is 0.26 points, but it is not significant. This suggests gains from active learning in one course may help students in the subsequent course of the sequence, but the evidence is weak.

Taken as a whole, this experiment confirms findings from other experiments. The fact that this is one of the few experiments done in an economics course outside of the US suggests the positive impact of these methods is not constrained by discipline or culture. Future research on other aspects of learning under these methods – particularly effort and the aspects of peer instruction that seem to yield benefits – should provide useful insights.

NOTES

[1] "The active learning interventions varied widely in intensity and implementation, and included approaches as diverse as occasional group problem-solving, worksheets or tutorials completed during class, use of personal response systems with or without peer instruction, and studio or workshop course designs." (Freeman et al. 2014, 1)

[2] The experiment's design was approved by the Universidad de los Andes Ethics Committee. Students were not told they were part of an experiment, but they were aware that their instructor was teaching the other section using a different methodology.

[3]The PSU ("*Prueba de Selección Universitaria"*) is the Chilean college entry exam. It is mandatory for all universities. Together with high-school grades, they form the application score used to rank applying students to determine entry. The incoming students had application scores of 653 to 761 points, and the mean application score was 681.

[4] In an end-of-semester survey, students in the control group pointed out that the instructor was very motivated and interested in answering questions and explaining the material as thoroughly as possible.

[5]Online quizzes were not part of the students' final grades. They were a prerequisite to taking the final exam, one that every student in the treatment group met.

[6]The treatment group used the Learning Catalytics platform (www.learningcatalytics.com) for delivering content, questions, and answers.

[7]We obtain the same results using random effects for the evaluations.

[8]In practice we include evaluation dummies for regressions (1), (3) and (4), and interact the evaluation dummies with the treatment status for regression (4). It is not possible to use fixed effects at the individual level because they absorb the treatment status and individual characteristics.

[9]Some of these instructors implemented active learning methodologies to varying degrees. Students in these sections were not randomly selected. For these reasons, this comparison cannot be considered conclusive, but it can offer insights into the difficulty of the second midterm relative to the first.

[10] This corresponds to a weighted average of the evaluations taken up to midterm 2, weighted according to the weights for computing the final grade.

[11] These numbers include the students that passed Algebra after the final exam. A few students who were close to passing were allowed to pass after individually considering their cases. However, the final grade considered for this study was the grade they obtained prior to these deliberations.

REFERENCES

Armstrong, N., S.M. Chang, and M. Brickman. 2007. Cooperative learning in industrial-sized biology classes. *CBE-Life Sciences Education*, 6 (2): 163-171.

Bergstrom, T.C. 2009. Teaching Economic Principles Interactively: A Cannibal's Dinner Party. *The Journal of Economic Education,* 40 (4): 366-384.

Bilgin, I. 2006. Promoting pre-service elementary students' understanding of chemical equilibrium through discussions in small groups. *International Journal of Science and Mathematics Education*, 4 (3): 467-484.

Bilgin, I., E. Senocak, and M. Sözbilir. 2009. The effects of problem-based learning instruction on university students' performance of conceptual and quantitative problems in gas concepts. *Eurasia Journal of Mathematics, Science & Technology Education*, 5 (2): 153-164.

Brouhle, K. 2011. Exploring Strategic Behavior in an Oligopoly Market Using Classroom Clickers. *The Journal of Economic Education*, 42 (4): 395-404.

Brown, B., and C. Liedholm. 2002. Can Web Courses Replace the Classroom in Principles of Economics? *American Economic Review*, 92 (2): 444-448.

Carmichael, J. 2009. Team-based learning enhances performance in introductory biology. *Journal of College Science Teaching*, 38 (4): 54.

Cartera, L. and T. Emersona. 2012. In-Class vs. Online Experiments: Is There a Difference? *The Journal of Economic Education*, 43 (1): 4-18.

Figlio, D., M. Rush, and L. Yin. 2013. Is it Live or Is It Internet? Experimental Estimates of the Effects of Online Instruction on Student Learning, *Journal of Labor Economics*, 31 (4): 763-784.

Freeman, S., S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth. 2014. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences,* 111 (23): 8410-8415.

Ghosh, S. and F. Renna. 2009. Using Electronic Response Systems in Economics Classes. *The Journal of Economic Education*, 40 (4): 354-365.

Green, A. 2014. The case for the traditional classroom. *International Review of Economics Education*, 16: 87-99.

Giraud, G. 1997. Cooperative learning and statistics instruction. *Journal of Statistics Education*, 5 (3): 1-13.

Hurley, J. D. 2002. *Effect of extended use of systematic instruction model on student achievement and content coverage in a "C" programming class*. PhD Thesis (Texas A&M University, College Station, TX).

Joyce, T. J., S. Crockett, D. A. Jaeger, O. Altindag, and S. D. O'Connell. 2014. Does classroom time matter? A randomized field experiment of hybrid and traditional lecture formats in economics. National Bureau of Economic Research, Working Paper No. 20006.

Keeler, C. M., and R. K. Steinhorst. 1994. Cooperative learning in statistics. *Teaching Statistics*, 16 (3): 81-84.

Knight, J. K., and W. B. Wood. 2005. Teaching more by lecturing less. *Cell biology education*, 4 (4): 298-310.

Lawson, T. J., J. H. Bodle, M. A. Houlette, and R. R. Haubner. 2006. Guiding questions enhance student learning from educational videos. *Teaching of Psychology,* 33 (1): 31-33.

Lovelace, T. L., and C. K. McKnight. 1980. The effects of reading instruction on calculus students' problem solving. *Journal of Reading*: 305-308.

Moelter, M. J., R. D. Knight, and C. Hoellwarth. 2005. A direct comparison of conceptual learning and problem solving ability in traditional and studio style classrooms. *American Journal of physics*, 73 (5): 459-462.

Olitsky, N. H., and S. B. Cosgrove. 2014. The effect of blended courses on student learning: Evidence from introductory economics courses. *International Review of Economics Education*, 15: 17-31.

Pandy, M. G., A. J. Petrosino, B. A. Austin, B. A., and R. E. Barr. 2004. Assessing adaptive expertise in undergraduate biomechanics. *Journal of Engineering Education*, 93 (3): 211-222.

Paschal, C. B. 2002. Formative assessment in physiology teaching using a wireless classroom communication system. *Advances in Physiology Education*, 26 (4): 299-308.

Prince, M. 2004. Does active learning work? A review of the research. *Journal of Engineering Education*, 93 (3): 223-232.

Salemi, M. K. 2009. Clickenomics: Using a Classroom Response System to Increase Student Engagement in a Large-Enrollment Principles of Economics Course. *The Journal of Economic Education*, 40 (4): 385-404.

Roach, T. 2014. Student perceptions toward flipped learning: New methods to increase interaction and active learning in economics. *International Review of Economics Education*, 17: 74-84.

Tirosh, D., and P. Tsamir. 2004. What can mathematics education gain from the conceptual change approach? And what can the conceptual change approach gain from its application to mathematics education? *Learning and Instruction*, 14 (5): 535-540.

Valcarcel, V. 2013. Instituting a Monetary Economy in a Semester-Long Macroeconomics Course. *The Journal of Economic Education*, 44 (2): 129-141.

Van Dijk, L. A., G. C. Van Der Berg, and H. Van Keulen. 2001. Interactive lectures in engineering education. *European Journal of Engineering Education*, 26 (1): 15-28.

Vosniadou, S., and L. Verschaffel. 2004. Extending the conceptual change approach to mathematics learning and teaching. *Learning and instruction*, 14 (5): 445-451.

Walker, J. D., S. H. Cotner, P. M. Baepler, and M. D. Decker. 2008. A delicate balance: integrating active learning into a large lecture course. *CBE-Life Sciences Education*, 7 (4): 361-367.

Williamson, V.M., and M. W. Rowe. 2002. Group problem-solving versus lecture in college-level quantitative analysis: the good, the bad, and the ugly. *Journal of Chemical Education*, 79 (9): 1131.

TABLES AND FIGURES

TABLE 1: Summary Statistics for Outcomes

	All		Control Group		Treatment Group		Difference		
Variable	Obs	Mean	Obs	Mean (1)	Obs	Mean (2)	(2)-(1)	p-value	Significant
quit §	78	0.026	39	0.026	39	0.026	0.000	0.500	no
failed §	78	0.269	39	0.333	39	0.205	-0.128	0.103	no
grade	75	4.309	38	4.155	37	4.446	0.291	0.008**	yes
Individual Evaluations (in chronological order)									
quiz1	76	4.725	37	4.605	39	4.838	0.233	0.146	no
quiz2	77	4.564	38	4.463	39	4.662	0.198	0.185	no
midterm1	77	4.909	38	4.861	39	4.956	0.096	0.222	no
quiz3	75	3.588	36	3.194	39	3.951	0.757	0.003**	yes
midterm2	77	3.818	38	3.508	39	4.121	0.613	0.001**	yes
exam	75	4.256	38	4.192	37	4.319	0.127	0.272	no

§ Binary variables - the variable takes value 1 if condition is met, 0 if not
The p-value corresponds to a t-test of equality of means between the two groups, with alternative hypothesis that the mean of the treatment group is larger for the grades, and smaller for the failure rate.
*p<.05; **p<.01

TABLE 2: Averages for Outcomes According to Individual Characteristics

Variable	All	Control Group (1)	Treatment Group (2)	Difference (2)-(1)	p-value	Significant
Failed						
Male	0.308	0.421	0.200	-0.221	0.071	no
Female	0.231	0.250	0.211	-0.039	0.389	no
Low Score	0.395	0.421	0.368	-0.053	0.374	no
High Score	0.150	0.250	0.050	-0.200	0.040*	yes
Grade						
Male	4.300	4.117	4.465	0.348	0.042*	yes
Female	4.297	4.190	4.424	0.234	0.045*	yes
Low Score	4.068	3.942	4.200	0.258	0.050	no
High Score	4.524	4.368	4.679	0.311	0.021*	yes

The p-value corresponds to a t-test of equality of means between the two groups, with alternative hypothesis that the mean of the treatment group is larger for the grades, and smaller for the failure rate.
*p<.05; **p<.01

TABLE 3: Treatment Effects Over Outcomes

	Treatment Effect				
Variable	In units	In Std. Dev.	Std. Err.	P-value	Significant
failed §	-0.134		0.100	0.180	no
Grade	0.237	0.451	0.107	0.029*	yes
quiz1	0.348	0.363	0.213	0.106	no
quiz2	0.150	0.156	0.224	0.504	no
quiz3	0.642	0.533	0.245	0.011*	yes
midterm1	0.059	0.108	0.122	0.629	no
midterm2	0.487	0.576	0.165	0.004**	yes
Exam	0.070	0.078	0.207	0.738	no

§ Binary variable: logit model. Marginal effect is presented
*p<.05; **p<.01

TABLE 4: Treatment Effect Over the Probability of Failing, by Individual Characteristics

Characteristic	Effect	Std. Err.	P-value	Significant
Gender				
Males	-0.249	0.145	0.087	no
Females	-0.022	0.133	0.868	no
Application score (PSU - mean=681)				
PSU=660	0.088	0.192	0.645	no
PSU=681	-0.206	0.109	0.060	no
PSU=700	-0.263	0.103	0.010*	yes

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

TABLE 5: Treatment Effect Over the Final Grade, by Individual Characteristics

	Effect				
Characteristic	In units	In Std. Dev.	Std. Err.	P-value	Significant
Gender					
Males	0.309	0.587	0.152	0.047*	yes
Females	0.168	0.320	0.150	0.265	no
Application score (PSU - mean=681)					
PSU=660	0.186	0.353	0.166	0.267	no
PSU=681	0.230	0.437	0.109	0.038*	yes
PSU=700	0.270	0.513	0.134	0.047*	yes

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

TABLE 6: Treatment Effect from Panel Data Models

Specification		Treatment Effect	p-value	Significant
Estimation 1		0.292	0.000***	yes
Estimation 2		0.287	0.004**	yes
Estimation 3		0.289	0.003**	yes
Estimation 4	eval 1: quiz 1	0.181	0.414	no
	eval 2: quiz 2	0.154	0.476	no
	eval 3: midterm 1	0.052	0.680	no
	eval 4: quiz 3	0.705	0.004**	yes
	eval 5: midterm 2	0.568	0.000***	yes
	eval 6: exam	0.080	0.688	no

*p<.05; **p<.01; ***p<.001

TABLE 7: Treatment Effect Over the Final Grade Without the Exam, by Individual Characteristics

Characteristic	Effect In units	Effect In Std. Dev.	Std. Err.	P-value	Significant
All	0.337	0.640	0.098	0.001**	yes
Gender					
Males	0.426	0.811	0.140	0.003**	yes
Females	0.252	0.479	0.136	0.069	no
Application score (PSU - mean=681)					
PSU=660	0.252	0.480	0.148	0.094	no
PSU=681	0.327	0.622	0.100	0.002**	yes
PSU=700	0.394	0.749	0.124	0.002**	yes

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

TABLE 8: Average midterm grades for the different sections

Sections	Midterm 1	Midterm 2	% change
Control	4.861	3.508	-27.8
Treatment	4.956	4.121	-16.8
Others not in the experiment	4.719	3.587	-24.0


Figure 1: Exam with respect to Weighted Grade
Exam Grade
1
2
3
4
5
6
7
Weighted Grade before the Exam
Control
Treatment