WE JUST ANALYSED THE MIDTERM SCORES OF OUR STUDENTS

PATRIK JANKOVIČ[1] – RICHARD KALIŠ[2] – ERIKA STRACOVÁ[3]

Vyhodnotili sme priebežné výsledky testov našich študentov

Abstract: *In this paper, we examined midterm scores of our students. Results suggest that we were able to provide equal conditions in teaching, examining and assessing despite the fact that the seminar was taught by multiple teachers. While there is a significant difference in the scores of some groups, controlling their previous grades can explain some of the differences. Therefore, one can conclude that these can be explained by the different level of individual students, rather than by the level of teaching, examining or assessing. Based on our experience, we also propose a way to ensure equal conditions for all students.*

Keywords: *student score, education, test, regression*

JEL: I 21, C 13

Introduction

There has always been the question whether one can create such environment that all students have the same conditions, especially in the case when the seminar is taught by multiple teachers. There is a clear trade-off between the equality and manageability of seminars. Naturally, one cannot handle teaching several groups of students at one time. For this particular reason, more lecturers are usually needed. However, it is still possible to examine these students on the same day and with one single test. Then the question arises, whether there is a significant difference among teachers. Furthermore, there can be a difference among the groups themselves. Therefore, some controlling for such a variable is necessary. Furthermore, even students within the same group and with the same teacher could have different knowledge.

[1] Ing. Patrik Jankovič, University of Economics in Bratislava, Slovak Republic, e-mail: patrik.jankovic@euba.sk

[2] Ing. Richard Kališ, University of Economics in Bratislava, Slovak Republic, e-mail: richard.kalis@euba.sk

[3] Ing. Erika Stracová, University of Economics in Bratislava, Slovak Republic, e-mail: erika.stracova@euba.sk

Then, the research question is whether there is a significant difference in the performance of a group according to different teachers.

The rest of the paper is organized as follows: Section 1 explains the background of our examination together with the theory of human capital and its importance. In Section 2, the specification of the model and methodology used are presented. One can find some descriptive statistics on the data used in section 3. Results and conclusion are the last parts of the paper.

Literature and Background

One could connect our analysis to the importance of human capital that goes back to Barro &Lee [1]. Furthermore, there is a non-trivial relationship between human capital and the level of GDP per capita [3]. Not only a level of human capital, but also its quality plays an important role. Such can be seen in most of the work by Hanushek & Kimko [2]. Qualitative factors for education can more than double the R^2 in explaining of variance in income per capita. Furthermore, when it comes to the question of returns to education, the great effort has been made by Psacharopoulos & Patrinos [4]. The recent questions are mainly connected to differences in private and public contributions, and social benefits in education.

Some background information on the examination and organization of seminars is necessary. Altogether, there were 243 students signed to the Quantitative Methods in Economics (QME). These students were divided into ten groups. Four different teachers were teaching the seminars, while there has been a one common speaker for all groups. Different student groups were randomly chosen by teachers. In addition to this, all students had an access to the same problem sets during the whole semester, and all teachers have been teaching more or less the same problems. From that perspective, there was not very much else we could do to create more equal conditions in seminars. We were also trying to create common conditions for the midterm test. First, a single date for the exam was chosen and it took place in two different rooms. Students were divided into two equally large groups. The midterm exam contained seven equally weighted problems to solve. Exams were assessed by all teachers. Every teacher had one or a maximum of two problems to asses in all tests. Therefore, it can be expected that there was no difference in assessing itself.

Methodology and Model

Apart from the histograms showing the distribution of our data, two statistical tests have been used. Firstly, the Shapiro-Wilk test for normality. Under the null hypothesis, the test tests whether a sample containing came from a normally distributed population. We can reject the null hypothesis when p-value is less than the chosen critical alpha level. In such a case, there is evidence that the tested data are not normally distributed and vice versa, [5].

Furthermore, the well-known t-test (Student's test) for the mean comparison has been used. The null hypothesis says that the difference between the two independent means is equal to zero [6].

To control for more variables instead of a simple statistical comparison, the least squares regression has been used. The model is specified as follows:

$$y = \alpha + \sum_{i}^{n-1} \beta_i E_i + \sum_{j}^{m-1} \gamma_j MathGrade_j + \delta Over + \epsilon \tag{1}$$

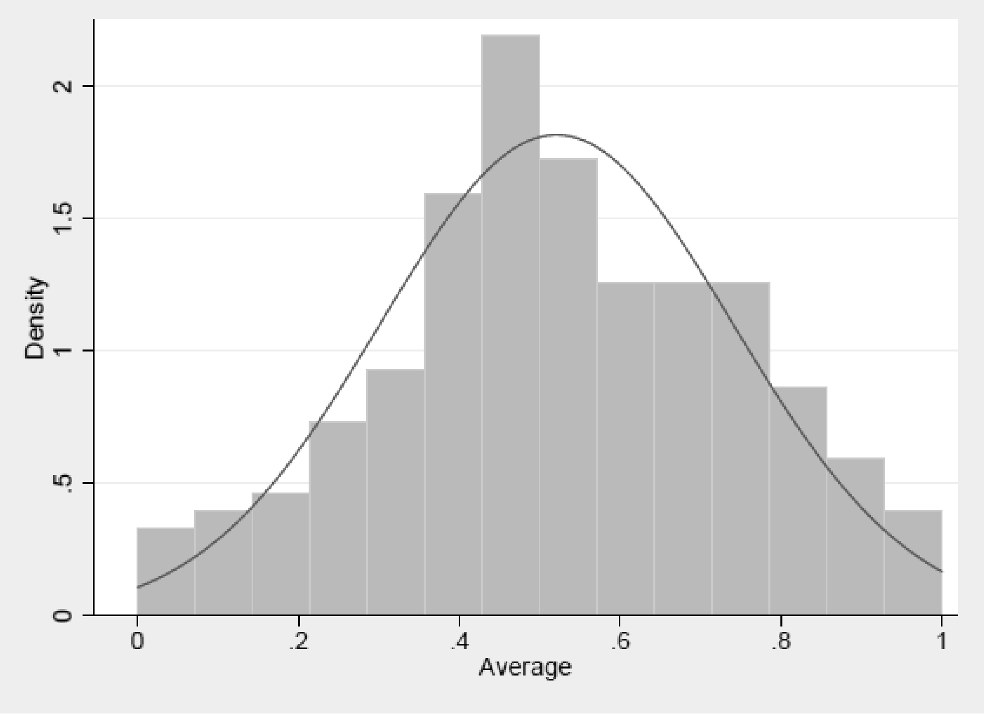
Where y represents the mean test score. E is a vector of effects specific for group or lecturer, respectively, while I is then an index of such a variable where $i = 1...n$. Furthermore, n is a sum of groups or lecturers. One group is always omitted, so actually $n - 1$ effects are examined. The vector of $MathGrade$ with the index j stands for binary variables of students' grades in Mathematics. There are m grades, but due to omitting one variable, $m - 1$ were used. Variable $Over$ is used to control for those students who repeat the course. Accordingly, β, γ, δ are coefficients and ϵ represents an error term**.**

Data

In this part of the paper, some statistics on the data are presented. As mentioned before, 243 students were signed to that particular course. However, 34 of them did not participate in the midterm examination, so the dataset is reduced to 209 students. Firstly, we can check the distribution of the average score. Every problem in the test was assessed separately on a scale from 0 to 100. Therefore, more interesting than the total score is the interpretation of the average score. However, naturally, the distribution of these two is the same. Figure 1 shows normally distributed average scores. Using the Shapiro-Wilk normality test one can conclude that at the 95 % significance level, the average score is normally distributed.

Figure 1

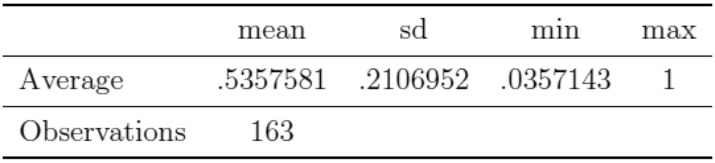
Distribution of average score



However, the dataset was further reduced. We gathered a controlling variable for the previous performance of our students, which was the final grade in Mathematics 1. For this reason, we reduced the dataset from 209 students to the final number of 163 students. Table 1 presents descriptive statistics over the sample of our students. As can be seen, there is quite a variance in the average performance of students with a mean of 0.53 and a standard deviation of 0.21.

Table 1

Summary Statistics



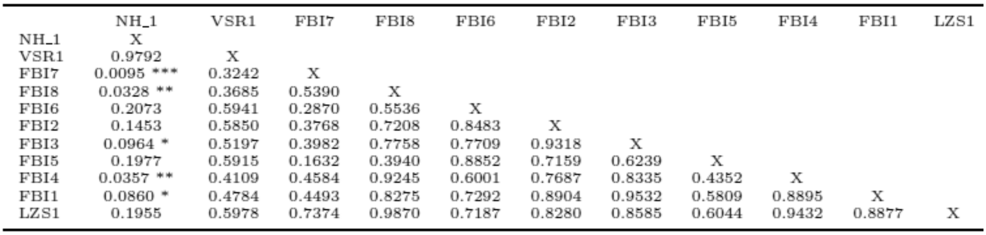
	mean	sd	min	max
Average	.5357581	.2106952	.0357143	1
Observations	163			

Results

In this section, the results of the paper are presented. Firstly, Table 2 shows p-values of the two-sample t-tests. The p-value representing the probability of rejecting the null hypothesis while it should be not rejected. The null hypothesis in this case is: *there is no difference among means of different groups*. As can be seen, there is a statistically significant difference only within the first column. In the case of the FBI7 group, there is a 0.0095 probability, that we should not reject null hypothesis, while we rejected it. Asterisks represent the confidential intervals of 0.9, 0.95 and 0.99, respectively. We cannot reject the null hypothesis within other groups.

Table 2

t-test of different groups

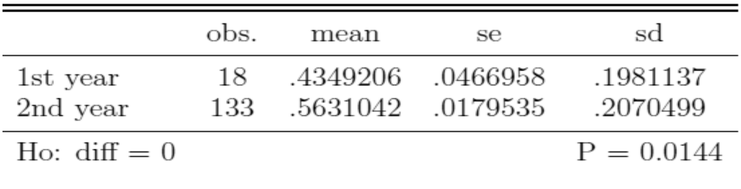


	NH_1	VSR1	FBI7	FBI8	FBI6	FBI2	FBI3	FBI5	FBI4	FBI1	LZS1
NH_1	X										
VSR1	0.9792	X									
FBI7	0.0095 ***	0.3242	X								
FBI8	0.0328 **	0.3685	0.5390	X							
FBI6	0.2073	0.5941	0.2870	0.5536	X						
FBI2	0.1453	0.5850	0.3768	0.7208	0.8483	X					
FBI3	0.0964 *	0.5197	0.3982	0.7758	0.7709	0.9318	X				
FBI5	0.1977	0.5915	0.1632	0.3940	0.8852	0.7159	0.6239	X			
FBI4	0.0357 **	0.4109	0.4584	0.9245	0.6001	0.7687	0.8335	0.4352	X		
FBI1	0.0860 *	0.4784	0.4493	0.8275	0.7292	0.8904	0.9532	0.5809	0.8895	X	
LZS1	0.1955	0.5978	0.7374	0.9870	0.7187	0.8280	0.8585	0.6044	0.9432	0.8877	X

Table 2 suggests that there is a significant difference only within the group of NH_1 students and some of other groups. This can be partly explained by the different year of their studies. While NH_1 students are only in the first year of their studies, the rest of the sample is mostly in the second year. Some of the results for this hypothesis are shown in Table 3.

Table 3

t-test of different means by years



	obs.	mean	se	sd
1st year	18	.4349206	.0466958	.1981137
2nd year	133	.5631042	.0179535	.2070499
Ho: diff = 0				P = 0.0144

As can be seen, there is a statistically significant difference between the students of the 1st and the 2nd year, respectively. Other students were excluded from this test. At the significance level of 95 %, we can reject the null hypothesis of no difference between the tested groups.

Furthermore, what could be interesting is the question whether the students that had to repeat our course were better or worse than the rest. Again, the t-test is used to compare the differences in the means. As can be seen in Table

4, there is a significant difference in the means between the students repeating the course (1) and those that took lectures for the first time (0).

Table 4

t-test of different means

	obs.	mean	se	sd
0	148	.5491795	.0173625	.2112245
1	15	.4033333	.0404354	.1566055
Ho: diff = 0				P = 0.0102

While the t-test is a useful tool for a mean comparison and can offer some explanation to this issue, we cannot control for more variables in such a simple test. For example, while there is a significant difference between the students of the first year compared to those of the second year, one cannot be really sure whether such a difference cannot be explained by different students, more than years of studies. In that case, model (1) is a much more appropriate one. Firstly, there is a -0.4884 correlation between average score of the midterm exam and the grades in Mathematics from the previous studies of our students. These grades could be a good proxy for some individual variance in skills of the students.

Altogether, three different models have been used. The overall results of the regressions can be seen in Table 5. The general idea of the model used is explained in the part two as an equation (1).

Model 1 explains the difference in average score by different groups, while controlling for whether the student had to repeat the class, as well as for the grades. In this case, the grades in Mathematics are measured on the scale from one to six. The groups are compared to the benchmark of NH_1 students. As can be seen, there is no statistically significant group. All the variance in average is explained by the variance in grades, or in other words, by the individual performance of the students. In general, we can say that a student with one degree worse grades in Mathematics had an average score worse by 7 % on average.

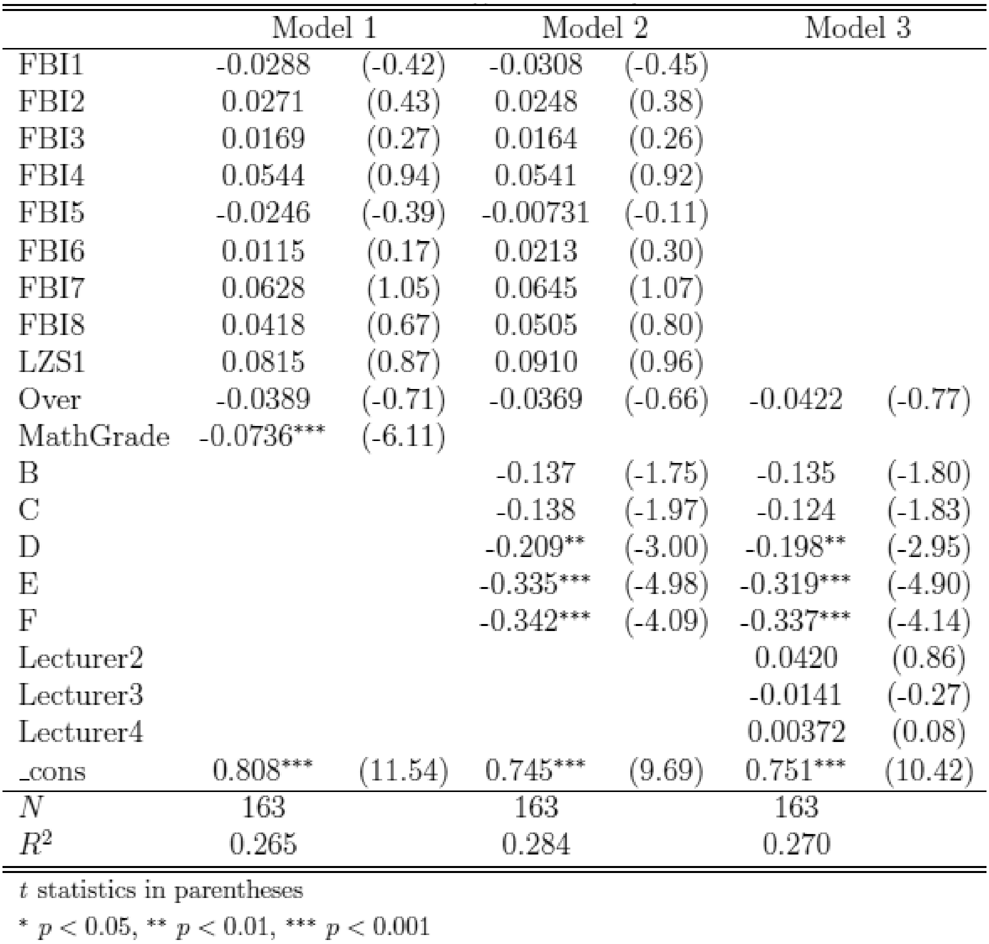
The only difference in *Model 2* is the way in the measurement of the Math degrees. A binary variable was used to capture a difference in the performance in Mathematics. In this case, the benchmark is a student from the NH_1 group with an A in Mathematics. As can be seen, there is no significant difference in the average score if student had a B or C grade. However, students that performed worse and got a D in Math had on average by 20 % less from our test. Additionally, while there is not a difference if student performed on an

E or F, both of them performed on our test by 34 % worse. There is still no significant difference among groups.

Model 3 controls rather for lecturers than for the groups. The idea behind it is that groups could be separately insignificant, while the lecturer could matter in the end. In this case, the benchmark is Lecturer1 and a student that had an A in Mathematics. Such a student had an average score of 75 % (see _cons). Again, all the variance in the average score in the midterm test can be explained exclusively by individual performance captured using previous grades in Mathematics and not by the particular teacher.

Table 5

Regression output



	Model 1		Model 2		Model 3	
FBI1	-0.0288	(-0.42)	-0.0308	(-0.45)		
FBI2	0.0271	(0.43)	0.0248	(0.38)		
FBI3	0.0169	(0.27)	0.0164	(0.26)		
FBI4	0.0544	(0.94)	0.0541	(0.92)		
FBI5	-0.0246	(-0.39)	-0.00731	(-0.11)		
FBI6	0.0115	(0.17)	0.0213	(0.30)		
FBI7	0.0628	(1.05)	0.0645	(1.07)		
FBI8	0.0418	(0.67)	0.0505	(0.80)		
LZS1	0.0815	(0.87)	0.0910	(0.96)		
Over	-0.0389	(-0.71)	-0.0369	(-0.66)	-0.0422	(-0.77)
MathGrade	-0.0736***	(-6.11)				
B			-0.137	(-1.75)	-0.135	(-1.80)
C			-0.138	(-1.97)	-0.124	(-1.83)
D			-0.209**	(-3.00)	-0.198**	(-2.95)
E			-0.335***	(-4.98)	-0.319***	(-4.90)
F			-0.342***	(-4.09)	-0.337***	(-4.14)
Lecturer2					0.0420	(0.86)
Lecturer3					-0.0141	(-0.27)
Lecturer4					0.00372	(0.08)
_cons	0.808***	(11.54)	0.745***	(9.69)	0.751***	(10.42)
N	163		163		163	
R^2	0.265		0.284		0.270	

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Based on our findings, we propose the following model of teaching, examining and assessing students. It is mostly useful for the model of teaching, when there are multiple seminar teachers leading the same course for students divided into more groups. Its aim is to provide the same conditions for all of them, including testing and evaluating. Our model is just a proposal, and it can have several modifications. It can serve as an inspiration for other teachers, mainly those teaching courses in quantitative methods.

The proposed structure of the seminar lessons and the way of examination:

- Seminar teachers meet at least once a week and decide on the structure and contents of the following lesson. They agree on the problems solved during the courses. Not all problems have to be identical; however, their degree of difficulty should be the same.
- They also present the common rules valid for all students, so everybody is aware of the way of examining and assessing.
- Teachers can also compose a common document including some of the problems solved in the classes together with explanations and results. The document can also include some additional problems for students to solve at home. These study materials can be published online, so every student can have an access to it at the same time.
- Students can use the consultation hours of any of the teachers. When they do not understand a certain topic taught by their teacher, they have a chance to consult it with the teacher of their choice.
- During the exam, the same test is provided to every student.
- The test is corrected by all teachers with each of them correcting one or two problems in every test. This ensures that every problem is corrected at the same way, with same criterion for all.

Conclusion

According to the results of this paper, it can be concluded that we were able to provide equal conditions in teaching, examining and assessing our students. All the differences in scores can mainly be explained by the individual differences among the students. Furthermore, while a single exam for all can be difficult to manage with a large group of students together with the assessing which is quite time-consuming, it can be concluded that it prevents unequal conditions.

We are fully aware that some control group should be used. Better results could be gathered by comparing different systems between these groups. However, such a test could harm some students and their performance.

In further research, we could make better use of information on previous grades of students. While a grade in Mathematics is a good proxy, the overall average would be worth examining as well.

References

[1] BARRO, R. J. – LEE, J.W. 1993. International comparisons of educational attainment. In: *Journal of Monetary Economics,* 32(3), pp. 363 – 394.

[2] HANUSHEK, E. A. – KIMKO, D. D. 2000. Schooling, labor-force quality, and the growth of nations. In: *American Economic Review,* 90(5), pp. 1184 – 1208.

[3] MANKIW, N. G. – ROMER, D. – WEIL, D. N. 1992. A contribution to the empirics of economic growth. In: *The Quarterly Journal of Economics,* 107(2), pp. 407 – 437.

[4] PSACHAROPOULOS, G. – PATRINOS, H. A. 2004. Returns to investment in education: a further update. In: *Education Economics,* 12(2), pp. 111 – 134.

[5] SHAPIRO, S. S. – WILK, M. B. 1965. An analysis of variance test for normality (complete samples). In: *Biometrika,* 52(3/4), pp. 591 – 611.

[6] STUDENT, 1908. The probable error of a mean. In: *Biometrika,* pp. 1 – 25.