e-pública
Revista electrónica sobre la enseñanza de la Economía Pública

Nº 21, septiembre, 2017
Págs. 21-30

The Influence of Non-cognitive Skills on a Multiple Choice Test Evaluation in Higher Education.

Pau Balart pau.balart@uib.cat

Universitat de les Illes Balears. Departamento de Economía de la Empresa. Cra. de Valldemossa, km 7,5. 07122–Palma de Mallorca, España.

*Este artículo recibió el **premio e-pública 2017** en las **IX Jornadas de Docencia en Economía** (Málaga, 1 y 2 de junio de 2017).*

Recibido: 1 de junio de 2017
Aceptado: 12 de julio de 2017

Abstract

I apply a recent methodology proposed by Borghans and Schils (2012) to a multiple choice test administered in an evaluation of business students. I find that on average the probability of correctly answering a question is reduced between a 8.52 and 11.2 percentage points when it occupies the last position in the test (in comparison to when it occupies the first position). In relative terms, this represents a 13.3-17.5% reduction in students' test scores. Following previous research, we can associate students' decline in performance with their non-cognitive skills.

Palabras clave: Evaluation, Non-cognitive Skills, Test Scores.

Códigos JEL: A20, A22

1. Introduction

Tests are probably the most extended way of evaluating knowledge. Despite the existence of several evaluation alternatives, different reasons may explain the extended use of tests, as its lower cost in comparison to other alternatives or its level of objectivity. There is, however, a growing attention towards the question of what is exactly measured by a test. As some authors have pointed out, exams are likely to capture some aspects that go beyond the realm of knowledge. In particular, students' personality and non-cognitive skills have been found to be relevant on test performance (Borghans et al., 2008; Segal, 2012). While these other aspects are found to be of great importance in the academic success and professional development of students, assessing knowledge with a higher degree of accuracy requires a better understanding of what is exactly captured by the evaluation tool.

Borghans and Schils (2012) proposed a methodology to decompose test scores in a cognitive and a non-cognitive component. Cognitive skills refer to students' knowledge in a given subject, such as their knowledge of Public Economics or their competence in numeracy. In contrast, non-cognitive skills positively affect student's performance in tests but are not directly related to their knowledge in the subject but to some aspects of their personality, such as perseverance or ability to work under pressure.

Borghans and Schils (2012) proposed the application of their methodology in the standardized tests of the Program of International Student Assessment (PISA). The PISA assessment includes different versions of the test (booklets). Every question (item) appears at different positions depending on which booklet is considered. Borghans and Schils (2012) observed that a larger proportion of students incorrectly answered the same item when it occupied a rearer position in the test. That is, they identify a decline in students' performance. Since student's knowledge in a given subject should not change during the course of the test, this performance decay is associated with aspects related to student's non-cognitive skills such as perseverance, intrinsic motivation or ability to maintain the attention during the test.

I apply the Borghans and Schils' (2012) methodology to a test administered in a course in Human Resource Management in the degree of Business Administration at Universitat de les Illes Balears. Carrying out this decomposition will allow to: (i) have a better understanding of what is measured by one of the most widespread evaluation tools such as exams; (ii) assist in improving the design of examinations in aspects such as their extension; (iii) be able to carry out more precise measurements of student knowledge and iv) be able to detect if any group of students is advantaged or disadvantaged by a specific design of the test.

I find that on average the probability of correctly answering a question is reduced in 8.52-11.2 percentage points when it occupies the last position in the test (in comparison to when it occupies the first position). In relative terms, this represents a 13.3-17.5% reduction in students' average score on a question. An important novelty of this work is that unlike previous studies, it applies this methodology to a high stakes exam. Previous works have found the existence of decay in low stakes environments such as the PISA test. Hence, in previous works the decline can be easily linked to intrinsic motivation or willingness to cooperate in an assessment. One may think that this decline in performance is unlikely to appear in the presence of extrinsic motivation, as it is the case in university exams. Our results show that this is not the case. We show that the performance decay also takes place in a high stakes situation. Consequently, students' grades can be potentially affected by this decay whenever we use tests as an evaluation mechanism.

The presence of this decay alerts about the importance of factors such as the duration of the exam. Duration may affect test scores for reasons that are not strictly related to students' knowledge on the evaluated topic. Hence, this should be taken into account when designing a test. In addition, there may be differences in the decay across different groups of students. If this is true, then tests might be biased in favour of the group of students that experience a lower decline. For instance, previous works showed that, in low stakes tests, girls experience a lower performance decay (Balart and Oosterveen, 2017). In the present work, I test whether there is any difference in the performance decline depending on the time slot in which the group is taking the course. In particular, I compare morning and evening groups. I observe that, students attending evening lessons experience a higher but non-statistically significant decline.

The methodology used in this work can be applied to the evaluation of any course that uses a multiple choice test with a sufficiently large number of questions and that makes permutations in question ordering. By applying this methodology, the instructor can obtain a better understanding on the type of skills that are evaluated in the course.

2. METHODOLOGY

2.1. THE PERFORMANCE DECLINE

Borghans and Schils (2012) observed a decline in performance during the PISA test. Using variation in the ordering of test items, they showed that the same question has a lower probability of being correctly answered when it occupies a rearer position in the test. By using question fixed effect, they can interpret differences in the performance decline as differences in testing behaviour.

Borghans and Schils (2012) estimate a model in which the test score of each question is explained by the sequence number of the test question:

$$y_{ij}= \eth_0+ \eth_1\, Q_{ij}+ \eth_2\, M_i+u_{ij}$$

Where y_{ij} is the score of a student (denoted by i) on a specific question (denoted by j). It can take value y_{ij} =1 if question j is correctly answered by student i or a value of y_{ij}=0, otherwise. Q_{ij} is the relative position (normalized between 0 and 1) of a specific question j in the test answered by student i and M_i accounts for controls such as gender or socioeconomic background.

The coefficient $\eth_1$ provides a measure of the average performance decline of test takers. Similarly, the intercept $\eth_0$ represents the average score of students at the beginning of the test (i.e., when Q_{ij}=0). As a consequence, $\eth_0$ can be used as a measure of knowledge net of the effects of the performance decline (Borghans and Schils, 2012). The model is estimated with question fixed effects to avoid that specific characteristics of each question, such as its level of difficulty or its clarity in the exposition, affect the estimation of the performance decline.

Borghans and Schils (2012) and Zamarro et al. (2016) validated $\eth_1$ as a measure of non-cognitive skills. In particular, they show that the performance decline is associated with higher levels of agreeableness (a Big Five personality trait) and higher motivation and ambition. Balart and Oosterveen (2017) provided additional evidence on this by showing that the performance decline is unrelated to knowledge by exploiting gender differences on the evaluated topic. In particular, it is largely documented that that, on average, females perform relatively better on reading questions, while males perform better in mathematics and science, see for instance Hyde and Linn (1988), Hyde et al. (1990), Caplan et al. (1997), Kimura (2004), Dee (2007), Fryer Jr and Levitt (2009), Cornwell et al. (2013) and Quinn and Cooc (2015). By decomposing the above equation by topic, Balart and Oosterveen (2017) show that girls experience a lower decline in performance also in the topics in which they perform worse on average (mathematics and science). This confirms that the decline in performance is not a consequence of differences in knowledge on the evaluated topic. This finding is consistent with the literature on gender differences in non-cognitive skills documenting girls to have a higher self-discipline, locus of control and conscientiousness (Duckworth and Seligman, 2006, Schmitt et al., 2008).

2.2. Decline in Performance Experienced by Students of Human Resource Management at Universitat de les Illes Balears

We implemented the previous methodology to disentangle the effect of students' knowledge in the evaluated topic and non-cognitive skills (performance decline) in their performance on a multiple-choice test used to evaluate a course on Human Resources. Several of the cognitive skills evaluated in this course are representative of the set of skills that are relevant for a degree in Business or in Economics. For instance, numeracy skills or normative analysis are critical in the analyzed course as well as in a course in Public Economics. At the same time, non-cognitive skills will be relevant not only in other courses, e.g., Public Economics, but in a broader range of situations.

The multiple-choice test has 25 questions. Two different versions of the test with two different orderings of the same set of items were administered. This is critical to the use of question fixed effects. In particular, I run the following regression:

$$y_{ij}= \eth_0+ \eth_1 Q_{ij} +u_{ij}$$

Where y_{ij} *and* Q_{ij} are defined as before. We estimated the model using question fixed effects.

To test for group differences in the performance decline, I considered the following model:

$$y_{ij}= \eth_0+ \eth_1 Q_{ij}+ \eth_2 G_i+ \eth_3 G_i * Q_{ij} + u_{ij}$$

Where G_i is a dummy variable that takes the value of one if the test taker is enrolled in the evening lessons group. The coefficient of the interacted term $G_i * Q_{ij}$ informs about group differences in the performance decline. If $\eth_3$ is negative and statistically significant, it will indicate that students in the evening group experience a higher decline in performance than students that attend morning lessons.

I estimated the previous models using OLS as well as using a Probit model. I clustered standard errors at student level. By doing so, I can answer to the following questions:

- Do higher education students experience a decline in performance when answering a high-stakes multiple-choice test?
- What is the size of such a decline?
- How does the performance decline (non-cognitive skills) contribute to their final score in the test?
- Do groups of students' differ in the performance decline?

The first question is especially interesting, as no previous study has analysed the performance decline in a high stakes exam as the present one. The second and the third questions are especially interesting from a teaching perspective. If one wants to evaluate exclusively students' knowledge in human resources, then their test scores might be corrected for the effect of the performance decline. This methodology offers the possibility of doing so. Third, it is also possible to check if there is any group of students that experiences a lower higher decline. For instance, we compare students taking their lessons in the morning with students taking their lessons in the evening. However, some other comparisons might be of especial interest. For instance, socioeconomic status has been found to play an important role on the formation of non-cognitive skills (Almlund et al., 2011). As a consequence, students proceeding form a disadvantaged environments could experience a higher decline during the test.

3. DATA

The data of this study was obtained at Universitat de les Illes Balears. Universitat de les Illes Balears has around 13.500 students. I used data from a multiple-choice test administered to evaluate students' knowledge in a course in Human Resources. Management of Human Resource is a third year course taught in the degree in Business Administration, in the degree in Labour Relationships and in the joint degree in Business and Law. I only considered students in the degree in Business Administration for this study. The test was administered in the academic year 2015-2016. Despite using data from a course in Management of Human Resources, the analysis of the present study can be extended to other courses in the degrees in Business and Economics that demand a similar set of skills such as numeracy abilities, abstract thinking, normative analysis or quantitative techniques. In particular, all the previous skills will be also critical for a course in Public Economics. At the same time, the non-cognitive skills captured in the performance decline in the test administered in the course in Human Resources will be equally relevant to succeed in a test in Public Economics or in any other discipline.

The multiple-choice test was part of the final exam of the course. It was the highest scored part of the exam with a weight of a 40%. The final exam accounts for 40% of the final score in the course.

The multiple-choice test consists in 25 questions. Every item contains four possible answers and only one was correct. A penalty of one third of the question value was implemented on incorrectly answered items. No penalty was implemented in unanswered items.

The students were asked to provide their answers in an optical answer sheet. The multiple-choice part was stopped after half an hour, thereafter the students had to continue with the rest of the exam for an additional 60 minutes. Two different versions of the test with two different orderings of the same set of items were administered. The two versions were allocated across students according to their seats in the room and preventing test takers to have the same version of the test as their immediate neighbours.

4. RESULTS

First we offer some descriptive statistics of the exam. 134 students took the test in the considered groups. The average score in the test was 6.4 points over 10.[1] We can also study the percentage of correct answers at the question level. As we can see there is substantial variation in the proportion of correct answers across questions. This indicates that there are substantial differences in question difficulty. Hence, having variation in question ordering and the use of question fixed effects is critical for being able to credibly compute the decline in performance throughout the test.

Table 1. Descriptive Statistics. Overall Test and Item Level

	Mean	St. Dev.
All Test	0.64	0.48
Item Level		
Item 1	0.60	(0.49)
Item 2	0.60	(0.49)
Item 3	0.69	(0.47)
Item 4	0.34	(0.48)
Item 5	0.71	(0.46)
Item 6	0.45	(0.50)
Item 7	0.77	(0.42)
Item 8	0.68	(0.47)
Item 9	0.50	(0.50)
Item 10	0.84	(0.36)
Item 11	0.50	(0.50)
Item 12	0.65	(0.48)
Item 13	0.60	(0.49)
Item 14	0.38	(0.49)
Item 15	0.81	(0.40)
Item 16	0.70	(0.46)
Item 17	0.86	(0.35)
Item 18	0.91	(0.29)
Item 19	0.39	(0.49)
Item 20	0.58	(0.50)
Item 21	0.84	(0.37)
Item 22	0.73	(0.44)
Item 23	0.61	(0.49)
Item 24	0.69	(0.46)
Item 25	0.57	(0.50)
Number of Students: 134		

Source: Author's computations.

Mean score of individual test items.

Management of Human Resources. Course 2015-2016. Universitat de les Illes Balears.

We compute the decline in performance using the equations described above. The results are displayed in Table 2. The first row displays the estimates of δ_1. According to the results in the first row, there is a statistically significant decline in performance. The probability of correctly answering a question is reduced in 8.52-11.2 percentage points when it occupies the last position in the test (in comparison to when it occupies the first position).

Relative to the average score of the test, this represents between a 13.3% and a 17,5% reduction in the probability of answering a question correctly.

This result is robust across all different specifications and robustness check of the model. Given that we are using question fixed effects, we can say that the very same question has a lower probability of being correctly answered when it occupies a rearer position in the test. Students experience a performance decline. As students' knowledge in the evaluated topic, i.e., human resources, should not change during the test, this implies that non-cognitive skills are affecting students' evaluation.

This result is new in the sense that it is, to the best of my knowledge, the first case in which the existence of a decline in performance is reported in a high stakes situation and at higher education. This suggests that the performance decline is not induced by a lack of effort or motivation, as it could be the case in a low stakes test such the PISA test. The performance decline also arises in the presence of extrinsic motivation. This implies that when students' knowledge is evaluated by means of tests, non-cognitive factors may potentially affect the outcome of the evaluation.

Table 2. Main Results. Performance decline Experienced by the Students of Human Resource at Universitat de les Illes Balears

	(1)	(2)	(3)	(4)
	OLS	OLS	Probit	Probit
Q	-0.112***	-0.0870**	-0.111***	-0.0852**
	(0.0370)	(0.0416)	(0.0363)	(0.0416)
G		0.0211		0.0207
(Evening group)		(0.0360)		(0.0371)
QxG		-0.0821		-0.0868
(Decline of the evening group)		(0.0543)		(0.0575)
Constant	0.642***	0.636***	0.369***	0.354***
	(0.0453)	(0.0469)	(0.118)	(0.122)
Question Fix Effects	Yes	Yes	Yes	Yes
Observations	3350	3350	3350	3350
Adj.R^2	0.096	0.096	0.081	0.083
(Pseudo-R^2 for Probit)				

When analysing differences in the decline across groups (columns 2 and 4), I observe that the sign of the coefficient $\eth_3$ is negative. Students in the evening group have a higher but non-statistically significant decline in performance. It should be emphasized that the exam was administered simultaneously to all groups (at 4:00 pm). Another interesting question would be to study whether time at which the exam was taken has any influence on the decline. However, variation in the time in which the exam was administered would be necessary to test this.

5. CONCLUSION

I have implemented a methodology to disentangle the effects of cognitive and non-cognitive skills students' test scores in a high stakes test at the university. I have found that non-cognitive skills affect students' test scores. That is, students' scores in the exam are influenced by some factors that are not strictly related to their knowledge on the evaluated topic. The size of this effect is substantial; the probability of correctly answering a question is reduced by a 13.3-17.5% when it is located at the last position of the test (with respect to when it is located at the beginning of the test).

It should be emphasized that the type of non-cognitive skills that are captured by the performance decline may be valuable in the labour market and should not be ignored. We may want to use multiple-choice tests precisely because they are also able to capture students' non-cognitive skills. In any case, it is important to understand which skills are being captured. Moreover, it is very likely that by using multiple-choice tests, we are only capturing a particular set of these skills. Some other knowledge evaluation mechanism, such as writing an essay or an oral presentation, are very likely to be affected by some other valuable non-cognitive skills. The extended use of tests as an evaluation mechanism might be promoting only some specific type of skills. In light of these results, the combination of different evaluation tools may provide a more precise way to evaluate higher education students at the same time that it captures a wider set of non-cognitive skills.

An interesting application of our results could consist in following up the students during their degree. That will allow computing performance decline at the student level and may provide a way of netting out the influence of non-cognitive skills on students' scores.

Following the research started by Balart and Oosterveen (2017) an interesting question that will be addressed in future research is the existence of gender differences in the performance decline on incentivized tests. Given the abundant existing differences between the test analysed here and the PISA ones, comparisons between the two will not be suitable. A specific study comparing tests with different level of stakes will be probably the best way of addressing this question.

The proposed methodology can be applied to the evaluation of any course that uses a multiple choice test or a sufficiently large number of questions and that makes permutations in question ordering. This may include a broad scope of courses in Economics and more specifically in Public Economics. Test evaluation can be implemented in courses related to Public Economics similarly to in any other course in Economics. Using this evaluation tool or any other mainly depends on the decision of the instructor. Universitat de les Illes Balears has two main courses in Public Economics: *Welfare Economics* and *Public Sector*. The former includes multiple-choice questions in students' assessment, while the latter does not. In particular, two midterm exams accounts for 90% of final grade of the course in *Welfare Economics*. Both of these midterms contain a multiple-choice part. Similarly as we did for the course in Human Resources, the decline can be computed for that part of the exam. In case of failing any of the two midterms, the students have the opportunity to retake the exam. Interestingly, the retake of second midterm is completely based on multiple-choice questions. This difference between the design of the ordinary and the retake exam opens the possibility of further research investigating how the duration of the exam affects the performance decline.

Notes:

[1] This is the gross score computed without taking into account penalties for incorrect answers. Despite this is not the final grade of the students; it provides a more suitable measure to compute the relative size of the performance decline.

Acknowledgments

I would like to thank Immanuel Mutua for his extremely helpful research assistance. I am also thankful to Jose María Carretero and Jenny de Freitas for their cooperation. I would also like to thank the organizers and participants at the IX Jornadas de Docencia en Economía. This project have been possible thanks to the support of Universitat de les Illes Balears and its project *Ajuts per a projectes d'innovació i millora de la qualitat docent*. This research has been founded by Obra Social "La Caixa" through the program *Ajuts per l'Assistència a Congressos "La Caixa" (Programa de Foment de la Recerca).*

References

Almlund, M.; A. L. Duckworth; J. Heckman and T. Kautz (2011): "Personality psychology and economics". *Handbook of the Economics of Education* 4.

Balart, P. and M. Oosterveen, (2017): "Wait and see: Gender gaps throughout cognitive tests". *JOLE Conference 2017 Working Paper*, 17679.

Borghans, L. and T. Schils (2012): "The Leaning Tower of Pisa. Decomposing achievement test scores into cognitive and noncognitive components". *JOLE 2012 Working Paper Proceedings.*

Caplan, P. J.; M. Crawford; J. S. Hyde and J. T. Richardson (1997): *Gender differences in human cognition. Counterpoints: Cognition, Memory, and Language Series.* ERIC.

Cornwell, C.; D. B. Mustard and J. Van Parys (2013): "Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school". *Journal of Human Resources* 48 (1): 236-264.

Dee, T. S. (2007): "Teachers and the gender gaps in student achievement". *Journal of Human Resource* 42 (3): 528-554.

Duckworth, A. L. and M. E. Seligman (2006): "Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores". *Journal of educational psychology* 98 (1): 198-208.

Fryer Jr, R. G. and S. D. Levitt (2009): *An empirical analysis of the gender gap in mathematics.* Technical report, National Bureau of Economic Research.

Hyde, J. S.; E. Fennema and S. J. Lamon (1990): "Gender differences in mathematics performance: a meta-analysis". *Psychological Bulletin* 107 (2): 139.

Hyde, J. S. and M. C. Linn (1988): "Gender differences in verbal ability: A meta-analysis". *Psychological bulletin* 104 (1): 53-69.

Kimura, D. (2004): "Human sex differences in cognition, fact, not predicament". *Sexualities, Evolution & Gender* 6 (1): 45-53.

Quinn, D. M. and N. Cooc (2015): "Science achievement gaps by gender and race/ethnicity in elementary and middle school trends and predictors". *Educational Researcher* 44 (6): 336-346.

Zamarro, G.; C. Hitt and I. Mendez (2016): *When Students Don't Care: Reexamining International Differences in Achievement and Non-Cognitive Skills*. EDRE Working Paper No. 2016-18.