Comparison of Computer-based and Paper-based Exams: Evidence from PISA[*]

Pelin Akyol[**]
Bilkent University

Abstract

We investigate the effects of the exam mode change - from paper-based to computer-based- on PISA on students' exam performance. This change took place in 2015 in 57 countries, while 15 countries continued to take the paper-based PISA. Using this change and the difference-in-differences estimation method, we find that the computer-based format reduced Turkish students' math, science and reading scores by 28.85, 29.52, and 39.975 points which correspond to a 5.9 percent decrease in math and science, and 8.1 percent decrease in reading scores compared to the average scores in OECD countries in the corresponding fields, respectively. We also investigate the heterogeneity of our results by gender and computer possession and show that there is no differential effect except a significantly negative impact for males in reading tests.

JEL Classification: I20, I21
Keywords: PISA, Mode Effects, Computer-based Assessment, Paper-based Assessment.

Bilgisayar Tabanlı ve Kağıt Tabanlı Sınavların Karşılaştırılması: PISA'dan Kanıtlar

Özet

Bu makalede sınav modu değişikliğinin (kağıt tabanlıdan bilgisayar tabanlıya) PISA'da öğrencilerin sınav performansı üzerindeki etkilerini araştırdık. 2015 yılında 57 ülkede PISA bilgisayar tabanlı gerçekleşirken, 15 ülke kağıt tabanlı PISA almaya devam etti. Bu değişikliği ve farklardaki fark tahmin yöntemini kullanarak, bilgisayar tabanlı formatın Türk öğrencilerin matematik, fen ve okuma puanlarını sırasıyla 28.85, 29.52 ve 39.975 puan azalttığını bulduk. Bu da OECD ülkelerinin ortalamasına göre matematik ve fen alanlarında yüzde 5.9'luk bir düşüşe, okuma alanında ise yüzde 8.1'lik bir düşüşe karşılık gelmektedir. Ayrıca cinsiyete ve bilgisayar sahipliğine göre sonuçlarımızın heterojenliğini araştırdık ve okuma testlerinde erkekler için önemli ölçüde olumsuz etki dışında farklı bir etki olmadığını gösterdik.

JEL Sınıflandırması: I20, I21
Anahtar Kelimeler: PISA, Mod Etkisi, Bilgisayar Tabanlı Değerlendirme, Kağıt Tabanlı Değerlendirme

[*] This work is funded by TUBITAK (The Scientific and Technological Research Council of Turkey) under project number 218K224. I thank Alfonso Leme and participants in the Meeting of the Economics of Education Association (AEDE) 2021 for useful comments and suggestions.

[**] Pelin Akyol. Department of Economics, Bilkent University, Çankaya, Ankara 06800 Turkey. E-mail: pelina@bilkent.edu.tr
ORCID: 0000-0002-1017-8844

Boğaziçi Journal Review of Social, Economic and Administrative Studies, Vol. 35, no. 2 (2021), pp. 137-150, doi: 10.21773/boun.35.2.2
Research Article | **Received:** 08.11.2021 - **Accepted:** 23.04.2022

Computer-based assessment methods have become widespread with the advent of technology. They have been used even more during the Covid-19 pandemic. Many high-stakes exams such as the TOEFL, SAT, and GRE are conducted in computer-based format. Furthermore, the Programme for International Student Assessment (PISA), an international test whose results are of great consequence for countries' education policies, is also currently computerized. It is argued that participants' performance in a computer-based exam may be affected by participant characteristics such as computer familiarity and anxiety and exam characteristics such as screen size and resolution or item review[2] possibility (Noyes et al. 2004). If computer-based assessment leads to distortions in scores, this may result in allocative inefficiency.

This paper aims to investigate the effect of computer-based format (mode effect) on test scores by using the change in the PISA exam mode in 2015 by focusing on Turkey.[3] OECD has conducted the PISA test every three years since 2000 to assess the reading, mathematics, and science skills of 15-years-old students. The test used to be in the paper-based format before 2015. It was conducted in computer-based format for the first time in 57 countries in 2015.[4] However, 15 countries where schools lacked sufficient technical infrastructure for a computer-based exam took the test in the paper-based format in 2015 (OECD 2017). We use this exam mode change and the difference-in-differences estimation method to investigate the effects of computer-based exams on students' performance in the PISA relative to paper-based exams. For this analysis, we use the 2006, 2009, 2012, and 2015 PISA data sets. We define countries that participated in the 2006, 2009 and 2012 PISA exams and took the paper-based PISA exam in 2015 as the control group and the students from Turkey as the treatment group.

According to the 2015 PISA exam results, Turkey's scores, which had had a rising trend in the past years, fell dramatically: Relative to 2012 scores, Turkey's score in math decreased from 447.9 to 420.5, science score decreased from 463.4 to 425.5, and reading score decreased from 475.5 to 428.3. Our findings suggest that computer-based testing decreased Turkey's math, science and reading scores by 28.85, 29.52 and 39.975 points, respectively. These effect sizes correspond to a 5.9 percent decrease in math and science, and a 8.1 percent decrease in reading scores compared to the average scores in OECD countries in the corresponding fields. That is, the mode of the exam primarily generated the decrease in Turkey's PISA scores in 2015. We also examine the heterogeneity of our results with respect to gender and computer possession, and find evidence that males are more adversely affected by the exam mode. Our results do not indicate heterogeneity in results by computer possession.

This paper contributes to the extensive literature that investigates the factors affecting the PISA scores and the outcome of various other exams, such as effort (Jacob 2005; Gneezy et al. 2019; Zamarro et al. 2019; Akyol et al. 2021), the penalty for wrong answers (Baldiga 2014 and Pekkarinen 2015), pollution (Ebenstein et al. 2016, Zhang et al. 2018), and temperature (Park 2020). As these studies show, factors other than students' level of knowledge or ability can affect test results. Therefore, it is crucial to identify these factors, especially in high-stakes exams.

[2] Item review or item flexibility refer to the ability to review, skip, and/or change items (Leeson 2006).

[3] Turkish students took computer-based PISA in 2015. Although we focus on students from Turkey in this paper, our approach is also applicable to other countries that took PISA in computer-based mode in 2015.

[4] 32 countries/regions participated in the PISA 2012 pencil-and-paper assessment were also invited to complete both a paper and a computer version of mathematics and reading test.

Similarly, the factors that influence countries' scores and rankings in international exams other than their education systems and students' quality can be misleading concerning the causes of changes in scores, resulting in resource misallocation.

Our paper also contributes to the literature that analyzes the effects of computer-based tests on student performance. Jerrim (2016) and Jerrim et al. (2018) are the closest works to ours. In 2012, when the PISA was a paper-based test, OECD ran field experiments in 32 countries to investigate the mode effect.[5] Jerrim (2016) finds large differences between computer- and paper-based test results using the data set of this experiment. High-scoring countries such as China and Hong Kong have lower scores in the computer-based test, whereas countries like Brazil and France performed better. Jerrim et al. (2018) also study the same question using a data set from a randomized control experiment run by the OECD in Germany, Sweden and Ireland and find a negative mode effect for all countries in all fields. The results of Jerrim (2016) are affected by the ordering effect, which has also been recognized in the literature: Students' performance drops as exam time passes (Akyol et al. 2021). Jerrim et al. (2018) solve this problem; however, the data used in this study does not represent the schools/students in those countries. Our paper contributes to this literature by providing evidence from Turkey by using representative PISA data.

The next section provides an overview of the data and the background information. Section 2 specifies the empirical methodology; Section 3 presents the results, and Section 4 concludes.

1. Background and Data

The PISA exams have been conducted every three years since 2000. More than half a million students took the 2015 PISA exam representing 28 million 15-year-old students from 72 countries. In 2015, for the first time, PISA was conducted as a computer-based exam in 57 countries, and the paper-based version was also available for countries that did not have the technical infrastructure needed. As a result, 57 countries and economies took PISA 2015 in computer-based assessment mode (CBA), and 15 countries took the paper-based version. Another feature of the computer-based PISA exam in 2015 was that students could not review the question they had already answered or skipped. This property of computer-based exams can be another channel that affects students' performance.

In addition to these changes, OECD changed the item response theory (IRT) model used to calibrate the response data in 2015. While a one-parameter logistic model (1PL) was used from 2000 to 2012, a two-parameter logistic (2PL) IRT model was used in 2015 (Feskens et al. 2019) which might have affected the scores of all students independent of the mode of the exam. This paper uses the diff-in-diffs estimation method to identify the effect of computer-based test mode on students' performance on math, science, and reading tests.

The PISA dataset is a publicly available dataset[6] that includes information on students' background characteristics and their performance in each field. These characteristics include month and year of birth, gender, family education level, school type and type of the school community, number of books at home, language spoken at home, immigration status, whether the student owns a computer, and student's socioeconomic status index (ESCS) and wealth index. ESCS and wealth

[5] In these countries, a randomly selected smaller group out of all students taking the PISA test were given a 45-minute computer-based mathematics test following the completion of the actual exam.

[6] It can be downloaded from https://www.oecd.org/pisa/data/.

indexes are readily available variables in the data set, which are created by considering different economic and social conditions in different countries. In PISA exams, theoretically, there is no minimum or maximum score. However, in every cycle, the results are scaled to fit approximately normal distributions, with means around 500 and standard deviations around 100 score points (OECD, 2019).

We use the 2006, 2009, 2012 and 2015 cycles of the PISA to examine the trends in the country scores. We first find out the countries that took the PISA in the paper-based mode in 2015 and participated in the previous cycles of the PISA.[7] After dropping countries that did not attend at least one cycle of the PISA exam between 2006 and 2015, we are left with three potential control group countries: Indonesia, Jordan, and Romania. In Section 3, we check the main assumption of the diff-in-diffs estimation method, parallel trend assumption, for these three control group countries. We continue our analysis with Romania, the only control group country that satisfies this assumption.

Table A.1 presents the descriptive statistics for pre-treatment and treatment periods for Turkey and control group countries. Table A.1 shows that Turkey's scores decreased in all fields. For the control group countries, while Romania and Indonesia experienced an increase in scores, Jordan's math and reading scores decreased slightly. Also, the characteristics of the students seem to differ across years and countries. An interesting point to note in Table A.1 is although these countries participated in paper-based PISA because of the lack of technical infrastructure at school, the share of students who have a computer at home is higher in Romania and Jordan relative to Turkey. Therefore, in our empirical analysis, we present results with and without controls for background characteristics.

2. Empirical Framework

We estimate the effect of computer-based testing relative to paper-based testing by estimating the following diff-in-diffs model:

$$S_{ict} = \beta_0 + \beta_1(\delta_T \times \delta_{2015}) + \beta_3' X_{ict} + \delta_T + \delta_{2015} + \varepsilon_{ict} \qquad (1)$$

where S_{ict} is the score of student i in country c who has participated in exam in period t. X_{ict} is the vector of covariates. δ_T is an indicator variable equal to 1 for students in Turkey, treatment country, 0 otherwise. δ_{2015} is an indicator variable equal to 1 for the year 2015, treatment period, 0 otherwise. The coefficient, β_1, of the interaction between δ_T and δ_{2015} gives the causal effect of computer-based assessment.

The vector of covariates, X_{ict}, include gender, month and year of birth of the student, school type, whether it is public or private, type of the school community (village, town, large town, etc.), economic, social and cultural status (ESCS) and wealth index, schooling level of mother and father, number of books at home, computer possession dummy, language spoken at home and the immigration status. We use survey weights in all estimations.

[7] Every year, the set of countries taking PISA changes. Some countries may drop from the sample while some others may join.

3. Results

3.1. Parallel Pre-Trend Assumption

For the diff-in-diffs method to give unbiased estimates, the parallel trend assumption should be satisfied in the pre-treatment period between control and treatment groups. In Figure 1, we present the performance of control and treatment countries in each field from 2006 to 2015. The visual inspection of the parallel pre-trend assumption shows that Romania's scores follows a similar trend to Turkey's scores in the pre-treatment period. Therefore, we will use Romania as the control group in our empirical analysis. Raw data patterns clearly show that although Turkey's scores decreased substantially in 2015, Romania did not experience this fall.

Table 1 Testing Common Trend Assumption

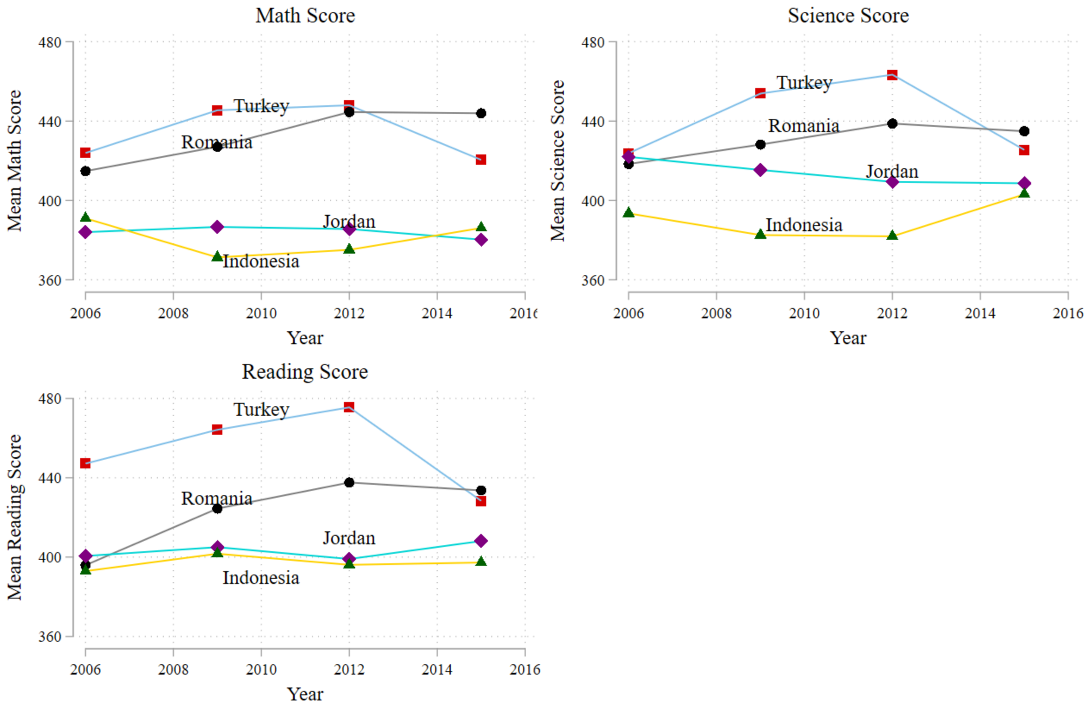
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Math	Math	Science	Science	Reading	Reading
Panel A) Periods of Observations: 2006-2015						
Year_2009	12.283*	10.198*	9.793	8.553	28.527***	27.900***
	(7.019)	(5.416)	(6.923)	(5.291)	(7.985)	(5.868)
Year_2012	29.758***	29.810***	20.382***	20.375***	41.668***	42.742***
	(6.819)	(5.127)	(6.746)	(5.116)	(7.586)	(5.582)
Year_2015	29.158***	25.046***	16.498**	13.175***	37.685***	35.507***
	(6.855)	(5.286)	(6.562)	(5.081)	(7.500)	(5.688)
Turkey	9.145	34.807***	5.446	28.881***	51.209***	76.689***
	(8.015)	(6.406)	(7.445)	(5.842)	(7.970)	(5.918)
Year_2009*Turkey	9.227	7.250	20.284**	17.772**	-11.473	-15.326**
	(10.933)	(8.115)	(9.877)	(7.260)	(10.716)	(7.480)
Year_2012*Turkey	-5.715	-3.585	19.198*	19.469***	-13.317	-14.065*
	(11.096)	(8.142)	(10.083)	(7.343)	(11.118)	(7.774)
Year_2015*Turkey	-32.646***	-27.281***	-14.842	-11.398	-56.491***	-56.020***
	(10.282)	(7.685)	(9.654)	(7.197)	(10.492)	(7.603)
Observations	40,525	40,333	40,525	40,333	40,525	40,333
Control Variables	No	Yes	No	Yes	No	Yes
Panel B) Periods of Observations: 2009-2015						
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Math	Math	Science	Science	Reading	Reading
Year_2012	17.475***	20.892***	10.589*	12.955***	13.141*	15.660***
	(6.263)	(4.806)	(6.176)	(4.653)	(7.402)	(5.277)
Year_2015	16.875***	13.904***	6.706	3.927	9.158	6.808
	(6.302)	(4.878)	(5.974)	(4.580)	(7.313)	(5.285)
Turkey	18.372**	41.119***	25.731***	45.424***	39.736***	61.459***
	(7.437)	(5.982)	(6.492)	(5.030)	(7.165)	(5.090)
Year_2012*Turkey	-14.942	-10.855	-1.086	1.577	-1.844	1.140
	(10.687)	(7.894)	(9.402)	(6.836)	(10.556)	(7.223)
Year_2015*Turkey	-41.872***	-34.193***	-35.126***	-28.745***	-45.018***	-39.414***
	(9.839)	(7.539)	(8.940)	(6.742)	(9.896)	(7.087)
Observations	30,465	30,291	30,465	30,291	30,465	30,291
Control Variables	No	Yes	No	Yes	No	Yes

Note: Data are from 2006, 2009, 2012, and 2015 PISA. * p<0.1 ** p<0.05 *** p<0.01. Standard errors are clustered at the school level. Turkey is the treatment country, and Romania is the control country. 2006, 2009, and 2012 are the pre-treatment periods, and 2015 is the treatment period. Regressions with control variables include gender and month and year of birth of the student, school type and type of the school community (village, town, large town, etc.), ESCS and wealth index, schooling level of mother and father, number of books at home, computer possession dummy, equal to one if the student has a computer, 0 otherwise, language spoken at home and the immigration status. Survey weights are used in all estimations.

We also test the common pre-trend assumption formally by running a regression of exam performance in each test on year fixed effects, treatment country dummy, and their interactions; and test the coefficients of the interaction of the year fixed effects and treatment dummy variables. In these specifications, year dummies capture factors affecting scores in both control and treatment countries in each year. Country dummies capture the factors affecting scores in each country that do not change over the years. The coefficients of interaction between year fixed effects and treatment country dummy show the differential trend between treatment and control country. If the coefficients are insignificant for the interaction of pre-treatment year fixed effects and treatment country dummy, it would mean no differential trend between control and treatment countries.

Table 1 presents the results. Panel A shows the results using the data from 2006 to 2015, and in panel B, we dropped 2006 data from our estimation sample. For the math test, the parallel pre-trend assumption holds in both panels A and B. For the science test, in panel A, before the treatment period, Turkey had a positive trend relative to the control country, and in 2015, it was reversed. So, our estimates for the effect of the computer-based exam on science performance from the 2006-2015 sample would be biased towards zero. Similarly, for the reading test, the interaction between year fixed effects and treatment dummy is negative and marginally significant (see Table 1 Column 6), which implies that there is a negative trend in the treatment country relative to the control country before the treatment period which would bias our estimates. When we examine the results in Panel B of Table 1, for all fields, the trend between control and treatment country is the same in the pre-treatment period (2009-2012). As Figure 1 shows, in reading and science tests, from 2006 to 2009, the change in the scores have different trends between Turkey and Romania. However, from 2009 to 2012, they have the same trends. So, the coefficients of interaction become insignificant when we exclude 2006 data. Therefore, the estimates in Panel B are more reliable estimates of the effect of computer-based assessment.

Figure 1 PISA Scores by Treatment and Control Countries



3.2. Main Results

In Table 2, we present our main results from estimating equation (1). Columns 1, 3 and 5 present results for math, science and reading without controlling background characteristics. Columns 2, 4, and 6 present the results by controlling background characteristics. Controlling for background characteristics decreases the size of the coefficients slightly, but they all stay large and significant. Our results in panel A show that computer-based assessment decreased Turkey's math, science and reading scores by 28.43, 22.66, and 47.11 points, respectively. However, as we explained, the estimates for science and reading tests would be biased downwards and upwards, respectively. The results in panel B show that computer-based assessment decreased Turkey's Math, Science, and Reading scores by 28.85, 29.52, and 39.975 points, respectively. We can also express the magnitude of this effect in terms of ranking. If the scores of other participating countries remained the same and had Turkey's scores not been affected by this change, Turkey's ranking in math, science, and reading would have increased by 7, 12 and 14 places, respectively. The magnitude of the effect can also be expressed according to the OECD average score of 493 in science and reading, and 490 in mathematics in 2015: Turkey's scores decreased by 5.9 percent in math and science, and 8.1 percent in reading compared to the OECD average in the corresponding field.

Table 2 Difference in Differences Estimation Results

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Math	Math	Science	Science	Reading	Reading
Panel A) Periods of Observations: 2006-2015						
Year_2015	17.432***	25.958***	8.063*	21.522***	17.948***	28.898***
	(5.246)	(4.763)	(4.859)	(4.439)	(5.853)	(5.027)
Turkey	13.636***	35.946***	21.943***	39.991***	47.846***	67.910***
	(4.606)	(4.068)	(4.131)	(3.538)	(4.528)	(3.561)
Year_2015*Turkey	-37.136***	-28.431***	-31.338***	-22.668***	-53.127***	-47.117***
	(7.918)	(5.972)	(7.405)	(5.559)	(8.189)	(5.969)
Observations	40,525	40,333	40,525	40,333	40,525	40,333
Control Variables	No	Yes	No	Yes	No	Yes
Panel B) Periods of Observations: 2009-2015						
	(1)	(2)	(3)	(4)	(5)	(6)
	Math	Math	Science	Science	Reading	Reading
Year_2015	8.443	9.523**	1.597	4.564	2.818	7.268
	(5.365)	(4.833)	(5.007)	(4.367)	(6.076)	(4.931)
Turkey	11.292**	35.785***	25.693***	46.199***	39.424***	62.019***
	(5.383)	(4.643)	(4.729)	(3.881)	(5.327)	(4.017)
Year_2015*Turkey	-34.792***	-28.851***	-35.088***	-29.521***	-44.706***	-39.975***
	(8.395)	(6.398)	(7.755)	(5.828)	(8.658)	(6.245)
Observations	30,465	30,291	30,465	30,291	30,465	30,291
Control Variables	No	Yes	No	Yes	No	Yes

Note: Data are from 2006, 2009, 2012, and 2015 PISA. * p<0.1 ** p<0.05 *** p<0.01. Standard errors are clustered at the school level. Turkey is the treatment country, and Romania is the control country. 2006, 2009, and 2012 are the pre-treatment periods, and 2015 is the treatment period. Regressions with control variables include gender and month and year of birth of the student, school type and type of the school community (village, town, large town, etc.), ESCS and wealth index, schooling level of mother and father, number of books at home, computer possession dummy, equal to one if the student has a computer, 0 otherwise, language spoken at home and the immigration status. Survey weights are used in all estimations.

We should also mention that any change made in the education system between 2012 and 2015 can affect our results. Right after the PISA 2012 exam held in April 2012, years of compulsory schooling in Turkey was extended to 12 years, and the system called 4+4+4 was introduced. This situation may have changed the group of students who could take the PISA exam by keeping students who would not continue their education after the 8th grade in the education system. If it is expected that students who would not continue their education after 8 years of schooling are from relatively more disadvantaged groups, then we expect a decrease in Turkey's PISA performance in 2015. According to OECD (2016a), the ratio of 15-year-old students to 15-year-olds in Turkey was 47 percent in 2006, 57 percent in 2009, and 68 percent in 2012. Spaull (2019) showed that this situation masked the increase in Turkey's performance between 2003 and 2012, and the actual increase was much higher. With the same reasoning, if there was an increase in the student population who could take the PISA exam in 2015, it may have contributed to the decrease in Turkey's score. However, according to the information presented in OECD (2016a), the rate of students in the 15-year-old age group who were in school and therefore could take the PISA exam was 68 percent in 2012, while it was 70 percent in 2015. So, there is only a tiny increase in the school participation rate. In addition, when we examine the results presented in Table 2, the results with and without controls are very similar to each other. Therefore, the increase of compulsory education to 12 years cannot drive our results.

In addition to the increase in the years of compulsory schooling, starting in 2010, Anatolian High Schools were converted to General High Schools gradually in Turkey. This conversion process was completed in June 2013. This change is also less likely to affect our results because this process started before the 2012 PISA exam, and no performance loss was observed in the 2012 PISA exam. Moreover, the PISA exam measures the basic knowledge that a person should have acquired by the age of 15, rather than the curriculum taught in high school. Therefore, in the short run, changes in high school education are not expected to have a major impact on the PISA exam results. However, in the long run, if having exam schools or different types of schools increase students' effort and motivation in the earlier grades, this may affect the students' accumulated knowledge by the age of 15, which will also affect their PISA performance.

3.3. Heterogeneity of the Results

We also investigate the heterogeneity of our results by gender and computer possession. We do so by estimating equation (1) and adding the triple gender (computer possession) interaction with treatment country and treatment year dummy variables.[8] The results presented in Table 3 show that the effects of computer-based assessment do not differ by computer possession.[9] These results provide evidence that the effect is not generated by the computer experience. However, it is also important to note that the question we use in the analysis only includes information about whether the person has a computer or not. Unfortunately, we do not have any information about how often and for what purpose the student uses the computer.

When we examine the effects by gender, we only observe marginally significant negative effects for males on the reading test. Therefore, we find evidence that male students were affected more negatively by this change in reading test. OECD (2016b) investigates heterogeneity in mode

[8] This specification includes double interaction of gender (computer possession) and treatment country and the interaction of gender (computer possession) and treatment year.

[9] Kroehne et al. (2019) also reached a similar conclusion for Germany that there is no differential mode effect by computer possession.

effects for all three assessment fields after the PISA 2015 computer-based exam and finds no significant effect by gender. Jerrim et al. (2018) find that men in Ireland were less affected by the computer-based test in the reading test, while in the science test, in Sweden, men were more adversely affected by the computer-based mode. Kroehne et al. (2019) could not find heterogeneity by gender in the mode effect. It seems that there is no consensus on the differential effect of computer-based mode by gender. Therefore, we also add to this literature by providing evidence from PISA.

Table 3 Heterogeneity Results

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A) Periods of Observations: 2006-2015						
	I) By Gender					
VARIABLES	Math	Math	Science	Science	Reading	Reading
Year_2015*Turkey	-38.119***	-27.344***	-35.025***	-24.101***	-50.237***	-41.076***
	(8.641)	(6.472)	(7.955)	(6.080)	(8.623)	(6.567)
Year_2015*Turkey*Male	2.340	-2.026	6.812	2.826	-8.116	-12.430**
	(6.157)	(4.617)	(6.188)	(4.671)	(6.567)	(4.974)
Observations	40,525	40,333	40,525	40,333	40,525	40,333
	II) By Computer Possession					
VARIABLES	Math	Math	Science	Science	Reading	Reading
Year_2015*Turkey	-30.977***	-21.059***	-24.965***	-13.830**	-45.206***	-38.014***
	(8.166)	(7.671)	(7.518)	(6.783)	(8.805)	(7.751)
Year_2015*Turkey*Computer Possession (Yes)	-7.305	-9.900	-6.448	-9.528	-6.243	-7.722
	(8.461)	(7.256)	(7.877)	(6.498)	(9.167)	(7.442)
Year_2015*Turkey*Computer Possession (Missing)	-16.074	-0.745	-2.800	-3.265	22.816	14.041
	(16.041)	(17.889)	(15.073)	(14.934)	(22.818)	(24.174)
Observations	40,525	40,333	40,525	40,333	40,525	40,333
Panel B) Periods of Observations: 2009-2015						
	(1)	(2)	(3)	(4)	(5)	(6)
	I) By Gender					
VARIABLES	Math	Math	Science	Science	Reading	Reading
Year_2015*Turkey	-34.420***	-27.871***	-36.927***	-30.613***	-41.941***	-35.867***
	(9.149)	(6.892)	(8.283)	(6.304)	(8.991)	(6.791)
Year_2015*Turkey*Male	-0.487	-1.925	3.327	2.130	-7.164	-8.612*
	(6.422)	(4.844)	(6.304)	(4.824)	(6.782)	(5.196)
Observations	30,465	30,291	30,465	30,291	30,465	30,291
	II) By Computer Possession					
	Math	Math	Science	Science	Reading	Reading
Year_2015*Turkey	-30.195***	-24.613***	-34.303***	-28.412***	-41.257***	-36.636***
	(8.124)	(7.666)	(7.307)	(6.631)	(8.655)	(7.545)
Year_2015*Turkey*Computer Possession (Yes)	-6.840	-5.045	-0.284	0.811	-3.163	-1.256
	(8.226)	(7.090)	(7.723)	(6.520)	(8.759)	(7.132)
Year_2015*Turkey*Computer Possession (Missing)	-16.294	2.272	3.740	10.690	19.119	18.898
	(16.362)	(17.830)	(16.200)	(15.720)	(23.835)	(24.651)
Observations	30,465	30,291	30,465	30,291	30,465	30,291

3.4. Placebo Analysis

As a robustness check of our results, we conduct a placebo analysis by focusing on the pre-treatment period, 2006-2012, and assigning a placebo treatment period to 2012. So, we estimate equation (1) by treating 2012 as the treatment period. Table 4 presents these results and shows that the coefficient of Year_2012*Turkey is insignificant except for the Science test, which is in the opposite sign and marginally significant. The results in Table 4 show that our results are not driven by the differences in pre-trends in scores between control and treatment countries.

Table 4 Placebo Difference in Differences Estimation Results with Pre-Treatment Period Data

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Math	Math	Science	Science	Reading	Reading
Year_2012	24.808***	33.296***	16.435***	26.776***	30.172***	38.962***
	(5.641)	(4.799)	(5.556)	(4.691)	(6.438)	(5.057)
Turkey	15.644***	42.433***	17.509***	38.779***	48.790***	72.596***
	(5.603)	(4.993)	(5.074)	(4.355)	(5.405)	(4.269)
Year_2012*Turkey	-12.214	-7.205	7.136	11.633*	-10.898	-7.648
	(9.503)	(6.861)	(8.486)	(6.128)	(9.451)	(6.511)
Observations	29,754	29,623	29,754	29,623	29,754	29,623
Control Variables	No	Yes	No	Yes	No	Yes

Note: Data are from 2006, 2009, and 2012 PISA. * p<0.1 ** p<0.05 *** p<0.01. Standard errors are clustered at the school level. Turkey is the treatment country, and Romania is the control country. 2006 and 2009 are the pre-treatment periods, and 2012 is the placebo treatment period. Regressions with control variables include gender and month and year of birth of the student, school type and type of the school community (village, town, large town, etc.), ESCS and wealth index, schooling level of mother and father, number of books at home, computer possession dummy, equal to one if the student has a computer, 0 otherwise, language spoken at home and the immigration status. Survey weights are used in all estimations.

4. Conclusion

Computer-based assessments have started to be used widely over the last years and this transformation has accelerated with the Covid-19 pandemic. However, we still do not have a good understanding of the effect of computer-based assessment on students' performance. In this paper, we investigate the effect of the computer-based exams on the performance of Turkish students by exploiting the PISA data set and the fact that it was conducted as a computer-based exam for the first time in 2015. However, some countries, such as Romania, continued to take the PISA exam in the paper-based mode as their schools do not have sufficient technical infrastructure. Using this change in the structure of the PISA exam and the difference-in-differences method, we find that the computer-based exam negatively affected Turkish students' performance in math, science, and reading tests by 28.85 points, 29.52 points, and 39.975 points, respectively.

We also investigate the heterogeneity of the results by gender and computer possession and find the decrease in reading score is larger for male students; however, we do not observe any heterogeneous effect by computer possession. These results imply that the negative effect of computer-based exams is less likely to be driven by computer inexperience. Therefore, further research is needed to identify the mechanisms through which computer-based testing decreases exam performance of examinees.

References

Akyol, P., Krishna, K., and Wang, J. (2021). “Taking PISA Seriously: How Accurate are Low-Stakes Exams?.” *Journal of Labor Research*, 1-60.

Baldiga, K. (2014). “Gender differences in willingness to guess.” *Management Science*, *60*(2), 434-448.

Ebenstein, A., Lavy, V., & Roth, S. (2016). “The long-run economic consequences of high-stakes examinations: Evidence from transitory variation in pollution.” *American Economic Journal: Applied Economics*, *8*(4), 36-65.

Feskens, R., Fox, J. P., and Zwitser, R. (2019). “Differential Item Functioning in PISA due to Mode Effects.” *Theoretical and Practical Advances in Computer-based Educational Measurement* (pp. 231-247). Springer, Cham.

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., and Xu, Y. (2019). “Measuring Success in Education: The Role of Effort on the Test Itself.” *American Economic Review: Insights*, 1(3), 291-308.

Jacob, B. A. (2005). “Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools.” *Journal of Public Economics*, 89(5-6), 761-796.

Jerrim, J. (2016). “PISA 2012: How do Results for the Paper and Computer Tests Compare?.” *Assessment in Education: Principles, Policy & Practice*, 23(4), 495-518.

Jerrim, J., Micklewright, J., Heine, J. H., Salzer, C., and McKeown, C. (2018). “PISA 2015: How Big is the ‘mode effect’ and What has been one about it?.” *Oxford Review of Education*, 44(4), 476-493.

Kroehne, U., Buerger, S., Hahnel, C., & Goldhammer, F. (2019). “Construct equivalence of pisa reading comprehension measured with paper-based and computer-based assessments.” *Educational Measurement: Issues and Practice*, 38(3), 97-111.

Leeson, H. V. (2006). “The Mode Effect: A literature Review of Human and Technological Issues in Computerized Testing.” *International Journal of Testing*, 6(1), 1-24.

Noyes, J., Garland, K., and Robbins, L. (2004). “Paper-based versus Computer-based Assessment: is Workload another Test Mode Effect?” *British Journal of Educational Technology*, 35(1), 111-113.

OECD (2016a). “PISA 2015 results (volume I).” *OECD Publishing*, Paris. doi: 10.1787/19963777

OECD (2016b), “Table A5 - Changes in the administration and scaling of PISA 2015 and implications for trends analyses”, in PISA 2015 Results (Volume I): Excellence and Equity in Education, PISA. *OECD Publishing*, Paris, https://doi.org/10.1787/9789264266490-table116-en. Look again here.

OECD (2017). “Programme for International Student Assessment an Overview.” *OECD Technical Report*, pp. 25

OECD (2019), “How PISA results are reported: What is a PISA score?”, in PISA 2018 Results (Volume I): What Students Know and Can Do. *OECD Publishing*, Paris. ditto

Paris.Park, R. J. (2020). “Hot temperature and high stakes performance.” *Journal of Human Resources.*

Pekkarinen, T. (2015). “Gender Differences in Behaviour under Competitive Pressure: Evidence on Omission Patterns in University Entrance Examinations.” *Journal of Economic Behavior & Organization*, 115, 94-110.

Spaull, N. (2019). “Who makes it into PISA? Understanding the impact of PISA sample eligibility using Turkey as a case study (PISA 2003–PISA 2012).” *Assessment in Education: Principles, Policy & Practice*, 26(4), 397-421.

Zamarro, G., Hitt, C., and Mendez, I. (2019). “When students don’t care: Reexamining international differences in achievement and student effort.” *Journal of Human Capital*, 13(4), 519-552.

Zhang, X., Chen, X., & Zhang, X. (2018). “The impact of exposure to air pollution on cognitive performance.” *Proceedings of the National Academy of Sciences*, 115(37), 9193-9197.

ONLINE APPENDIX

Table A.1: Descriptive Statistics

	Turkey		Romania		Jordan		Indonesia	
	Year<2015	Year=2015	Year<2015	Year=2015	Year<2015	Year=2015	Year<2015	Year=2015
	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
	(SD)	(SD)	(SD)	(SD)	(SD)	(SD)	(SD)	(SD)
Math Score	440.158	420.454	426.522	443.954	385.499	380.259	378.906	386.11
	(89.37)	(75.126)	(78.538)	(78.355)	(76.954)	(76.908)	(69.502)	(71.423)
Science Score	448.765	425.49	426.822	434.885	415.144	408.669	385.756	403.1
	(78.94)	(75.128)	(76.323)	(73.946)	(82.548)	(78.993)	(64.3)	(62.743)
Reading Score	463.514	428.335	415.669	433.617	401.527	408.102	396.882	397.259
	(83.46)	(76.658)	(88.164)	(87.047)	(87.688)	(86.054)	(67.914)	(67.896)
Gender(Male)	0.521	0.5	0.494	0.498	0.498	0.495	0.505	0.501
	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)
ESCS	-1.308	-1.428	-0.391	-0.582	-0.515	-0.424	-1.634	-1.873
	(1.145)	(1.171)	(0.941)	(0.868)	(1.059)	(1.012)	(1.098)	(1.114)
Mother Education-not complete	0.137	0.134	0.008	0.007	0.062	0.033	0.12	0.118
	(0.344)	(0.341)	(0.091)	(0.084)	(0.242)	(0.178)	(0.325)	(0.323)
Completed ISCED 3A	0.462	0.369	0.033	0.049	0.06	0.037	0.333	0.332
	(0.499)	(0.482)	(0.178)	(0.216)	(0.237)	(0.188)	(0.471)	(0.471)
Completed ISCED 3B, 3C	0.165	0.2	0.152	0.155	0.17	0.18	0.194	0.206
	(0.371)	(0.4)	(0.359)	(0.362)	(0.376)	(0.385)	(0.395)	(0.404)
Completed ISCED 2	0.013	0.138	0.162	0.123	0.061	0.083	0.039	0.034
	(0.113)	(0.345)	(0.369)	(0.328)	(0.239)	(0.276)	(0.193)	(0.181)
Completed ISCED 1	0.176	0.147	0.62	0.665	0.572	0.597	0.284	0.282
	(0.381)	(0.354)	(0.485)	(0.472)	(0.495)	(0.491)	(0.451)	(0.45)
Mother Education(missing)	0.048	0.011	0.024	0.002	0.075	0.07	0.03	0.028
	(0.213)	(0.106)	(0.153)	(0.042)	(0.264)	(0.255)	(0.171)	(0.166)
Father Education-not complete	0.048	0.059	0.007	0.007	0.037	0.031	0.088	0.098
	(0.214)	(0.235)	(0.082)	(0.086)	(0.19)	(0.173)	(0.283)	(0.297)
Completed ISCED 3A	0.34	0.31	0.034	0.044	0.069	0.041	0.291	0.298
	(0.474)	(0.463)	(0.182)	(0.205)	(0.254)	(0.198)	(0.454)	(0.457)
Completed ISCED 3B, 3C	0.231	0.27	0.11	0.125	0.162	0.171	0.187	0.191
	(0.422)	(0.444)	(0.313)	(0.331)	(0.368)	(0.376)	(0.39)	(0.393)
Completed ISCED 2	0.027	0.188	0.247	0.215	0.096	0.11	0.069	0.061
	(0.163)	(0.391)	(0.431)	(0.411)	(0.294)	(0.313)	(0.253)	(0.239)
Completed ISCED 1	0.314	0.16	0.568	0.604	0.565	0.584	0.322	0.319
	(0.464)	(0.366)	(0.495)	(0.489)	(0.496)	(0.493)	(0.467)	(0.466)
Father Education (missing)	0.039	0.012	0.034	0.005	0.071	0.064	0.043	0.033
	(0.194)	(0.109)	(0.182)	(0.069)	(0.257)	(0.244)	(0.204)	(0.179)
Language at home (another)	0.044	0.071	0.026	0.027	0.035	0.047	0.618	0.616
	(0.205)	(0.257)	(0.16)	(0.163)	(0.185)	(0.213)	(0.486)	(0.486)
Language at home (test language)	0.948	0.922	0.971	0.973	0.938	0.924	0.366	0.346
	(0.221)	(0.268)	(0.168)	(0.163)	(0.242)	(0.265)	(0.482)	(0.476)
Language at home (missing)	0.008	0.006	0.003	0	0.027	0.028	0.016	0.038
	(0.089)	(0.08)	(0.052)	(0.014)	(0.162)	(0.166)	(0.127)	(0.19)

0-10 books	0.248	0.249	0.178	0.231	0.287	0.348	0.226	0.303
	(0.432)	(0.432)	(0.383)	(0.422)	(0.453)	(0.477)	(0.418)	(0.46)
11-25 books	0.265	0.285	0.226	0.222	0.254	0.238	0.372	0.355
	(0.442)	(0.451)	(0.418)	(0.416)	(0.435)	(0.426)	(0.483)	(0.479)
26-100 books	0.281	0.275	0.301	0.284	0.244	0.216	0.264	0.223
	(0.449)	(0.447)	(0.459)	(0.451)	(0.429)	(0.412)	(0.441)	(0.416)
101-200 books	0.102	0.094	0.143	0.127	0.086	0.076	0.061	0.051
	(0.303)	(0.291)	(0.35)	(0.333)	(0.28)	(0.265)	(0.239)	(0.221)
201-500 books	0.058	0.055	0.092	0.091	0.035	0.029	0.024	0.017
	(0.233)	(0.228)	(0.289)	(0.288)	(0.185)	(0.167)	(0.155)	(0.129)
More than 500 books	0.031	0.03	0.052	0.042	0.041	0.036	0.021	0.011
	(0.172)	(0.169)	(0.222)	(0.201)	(0.199)	(0.187)	(0.143)	(0.107)
Number books (missing)	0.015	0.013	0.008	0.002	0.052	0.057	0.032	0.039
	(0.123)	(0.115)	(0.088)	(0.046)	(0.223)	(0.231)	(0.175)	(0.194)
Computer Possession (No)	0.423	0.316	0.248	0.122	0.265	0.208	0.77	0.706
	(0.494)	(0.465)	(0.432)	(0.327)	(0.441)	(0.406)	(0.421)	(0.456)
Computer Possession (Yes)	0.558	0.663	0.735	0.872	0.703	0.756	0.197	0.279
	(0.497)	(0.473)	(0.442)	(0.335)	(0.457)	(0.429)	(0.397)	(0.449)
Computer Possession (Missing)	0.018	0.021	0.017	0.007	0.033	0.036	0.034	0.015
	(0.134)	(0.144)	(0.129)	(0.082)	(0.178)	(0.186)	(0.18)	(0.12)
Immigration Status (Native)	0.966	0.968	0.981	0.984	0.819	0.833	0.984	0.975
	(0.182)	(0.175)	(0.135)	(0.124)	(0.385)	(0.373)	(0.126)	(0.155)
Second-Generation	0.006	0.005	0	0.002	0.101	0.085	0	0
	(0.079)	(0.069)	(0.019)	(0.05)	(0.301)	(0.28)	(0.02)	(0.018)
First-Generation	0.003	0.003	0.001	0.001	0.039	0.03	0.002	0.001
	(0.055)	(0.053)	(0.035)	(0.036)	(0.193)	(0.169)	(0.04)	(0.031)
Immigration Status (missing)	0.025	0.024	0.017	0.012	0.041	0.052	0.014	0.023
	(0.156)	(0.154)	(0.129)	(0.109)	(0.199)	(0.223)	(0.118)	(0.151)
School Type (Private)	0.178	0.048	0.172	0.011	0.173	0.197	0.222	0.408
	(0.383)	(0.213)	(0.377)	(0.103)	(0.378)	(0.397)	(0.416)	(0.492)
School Type (Public)	0.789	0.948	0.795	0.989	0.782	0.787	0.736	0.592
	(0.408)	(0.222)	(0.404)	(0.103)	(0.413)	(0.409)	(0.441)	(0.492)
School Type (missing)	0.033	0.005	0.033	0	0.045	0.016	0.042	0
	(0.178)	(0.068)	(0.179)	(0)	(0.208)	(0.126)	(0.2)	(0)
School Community (Village)	0.121	0.014	0.113	0.108	0.13	0.136	0.127	0.299
	(0.326)	(0.117)	(0.317)	(0.31)	(0.337)	(0.343)	(0.333)	(0.458)
Small Town	0.205	0.065	0.214	0.21	0.224	0.292	0.22	0.414
	(0.404)	(0.247)	(0.41)	(0.408)	(0.417)	(0.455)	(0.414)	(0.493)
Town	0.28	0.311	0.313	0.38	0.268	0.221	0.272	0.138
	(0.449)	(0.463)	(0.464)	(0.485)	(0.443)	(0.415)	(0.445)	(0.345)
City	0.205	0.212	0.198	0.244	0.224	0.177	0.23	0.061
	(0.403)	(0.409)	(0.399)	(0.429)	(0.417)	(0.381)	(0.421)	(0.24)
Large City	0.15	0.392	0.125	0.058	0.125	0.156	0.117	0.079
	(0.357)	(0.488)	(0.33)	(0.234)	(0.331)	(0.363)	(0.321)	(0.27)
School Community (missing)	0.038	0.005	0.037	0	0.028	0.018	0.034	0.008
	(0.192)	(0.068)	(0.188)	(0)	(0.166)	(0.132)	(0.182)	(0.091)
Wealth Index	-1.368	-1.474	-0.9	-0.937	-1.159	-0.911	-2.294	-2.673
	(1.123)	(1.017)	(0.98)	(0.978)	(1.061)	(1.267)	(1.264)	(1.344)