A Work Project, presented as part of the requirements for the Award of a Master's degree in Economics from the Nova School of Business and Economics.

Can the keyboard beat the good old pen? The effect of computer-based testing on students' performance

ANA CLÁUDIA MELO DOS REIS

Work project carried out under the supervision of:

Luís Catela Nunes

João Pereira dos Santos

Pedro Freitas

16/01/2023

Abstract (100 words maximum)

Computer-based testing is being increasingly adopted by several education institutions worldwide. Whether this transition from paper-based testing yields different impacts on students' performance is still an open question. This paper aims to assess the impact of computer-based testing exploiting a large-scale pilot-program deployed in 2022 low-stake exams in Portugal. We rely on rich cross-sectional student-level data to implement pooled OLS and difference in differences approaches. Results indicate statistically significant differences in students' score between computer-based and paper-based testing with differences ranging from -5pp to -24pp. Finally, we discuss important heterogeneity effects and implications of implementing this policy at scale.

Keywords: Computer-based testing; Paper-based testing; Low-stake exams; Student performance

I would like to express my deepest gratitude to IAVE's team: the President of the board, Luís Santos; the Teacher Training director, Rui Pires; Paula Simões; Catarina Lains and Marta Teixeira, for providing the used data and the immense support throughout the whole analysis process.

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

I. Introduction

The assessment of students' performance is an essential part of any education system. There are several types of students' assessments, that aim to reflect student proficiency at a given moment. A frequently used assessment tool are tests, either the ones graded by the students' respective teachers or national standardised ones. Since the beginning of the century, computer-based testing (CBT) mode is becoming increasingly common across private and public education institutions worldwide (Parhizgar, 2012). The COVID-19 pandemic has intensified the use of CBT, as many governments have temporarily closed education institutions in order to mitigate the spread of the virus (Wang, Jiang, Li, & Lie, 2022).

In this paper, we study the mode effect of computer-based testing on low-stake national exams for Portuguese 2nd, 5th, and 8th graders, taking advantage of a pilot-project of CBT mode conducted by the education evaluation institute of Portugal (IAVE[1]) in 2022. We exploit rich cross-sectional data, covering 172 299 students from these three grades in 4 166 schools and all five written exams conducted this year, totalling 277 390 exams. The analysis also includes socioeconomic background variables of students and schools, such as age, gender, social support from school, among others. Since this pilot-program was not implemented as a randomised control experiment, we exploit different but complementary econometric strategies to assess the CBT mode effect, namely a pooled OLS with different samples and a student-level difference in differences. We find that students who used CBT performed worse than their paper counterparts, on average. These results are consistent throughout all strategies, strengthening the conclusions reached.

This paper contributes to the literature in two ways. First, to our knowledge, this is the first large-scale study of computer-assisted testing mode effects using national-wide data for

[1] Previously named GAVE until 2013

low-stake exams in several grades, whereas previous literature mainly focused on online assessments in small-scale contexts. Secondly, we explore CBT mode effects in a wide range of exams' subjects, understanding how CBT can have varying impacts according to the evaluated subject matter domains. Since IAVE plans to have all basic and secondary education national exams, both low-stake and high-stake, as computer-based by 2025, it is crucial to understand (1) what are the effects of CBT versus PPT in terms of students' performance and (2) if those effects are heterogeneous across the population, so that appropriate policies can be undertaken to mitigate possible negative effects of transitioning to CBT.

There are several advantages of using CBT when compared with paper-and-pencil testing (PPT), namely: higher flexibility in test item design; access to a large repository of items; reduced scope for cheating (Backes & Cowan, 2019); reduced "time lag" in reporting scores; increase in the efficiency of assessment; possibility of analysing student performance that cannot be investigated from paper-based tests; reduced paper consumption and costs (Öz & Özturan, 2018); and reduce potential biases on scoring, which can have long run implications for occupational choices and earnings at adulthood (Lavy, Sand, & Shayo, 2022). In fact, many national and international institutions are planning to or already applying CBT in their assessments, for example, ACT, PISA, NAEP and PARCC, as well as at country level for high-stake exams, such as in Norway[2] and Sweden (Bagger, Norén, Boistrup, & Lundahl, 2019).

Notwithstanding, CBT also poses some challenges such as: infrastructure and software quality and compatibility, the risk of testing candidates that are not used to the tool, hence testing other skills than intended; typing problems; and possible high investment costs (Scheuermann & Björnsson, 2009). In the European Union (EU), recent data shows a disparity

[2] Leire Aranbarri, "Digitise the Norwegian National Tests at Scale", *Inspera*, last accessed December 14, 2022, https://www.inspera.com/blog/digitise-the-norwegian-national-tests-at-scale.

in digital education across Member States: in 2019 access to broadband internet varied from 74% of households for the lowest-income quartile to 97% in the highest-income quartile (European Comission, 2020). On teacher preparedness, the OECD Teaching and Learning International Survey in 2018 showed that only 39% of educators in the EU felt well or very well prepared for using digital technologies in their daily work, with significant differences between Member States (European Comission, 2020).

This paper proceeds as follows. Section II shows previous literature work on the matter. Section III provides institutional background and describes the experiment set-up and data. Section IV lays out the empirical strategy used. Section V presents results and Section VI explores heterogeneity effects. Section VII concludes and discusses policy implications. Section VIII are references and Section IX is the appendix.

II. Literature Review

Over the last two decades, several researchers tried to assess the impacts of different testing modes, namely computer-based vs. paper-based. Some studies indicate that CBT students outperform their paper peers (Wang, Kao, & Chen., 2021). Clariana & Wallace (2002) uses data from 105 first-year business undergraduates for the course of Computer Fundamentals, concluding that CBT students performed better than the PPT ones. However, most literature supports PPT over CBT (Backes & Cowan, 2019; Wuthisatian, 2020) or does not find statistically significant differences between the two testing modes (Goodwin, Cho, Reynolds, Brady, & Salas, 2020; Öz & Özturan, 2018). For example, Beatty, Esco, Curtiss & Ballen (2022) found out that students who preferred online CBT consistently underperformed the PPT students, for 305 second-year organic chemistry undergraduate students.

One big question that arises in the literature is the comparability of CBT with PPT, i.e., whether exam scores obtained from either testing mode are interchangeable. Examinees may

have different performance levels between the testing modes due to several factors, namely item presentation. Some studies concluded that CBT is more challenging than PPT (Laborda, 2010), but the majority showed that these test modes are comparable (Logan, 2015; Jeong, 2014; Retnawati, 2015; Öz & Özturan, 2018). Nevertheless, type of item and item presentation may still influence results. Wang, Kao & Chen (2021) show that CBT students had a better answering performance than PPT in multiple-choice questions with multiple-item presentation, but no statistical differences were found between CBT and PPT with single-item presentation, for 381 primary school fifth graders in northern Taiwan on the subject of Natural Sciences. Leeson (2006) finds that a single-item presentation on CBT may lead students to answer hastily, leading them to make mistakes and negatively influencing performance.

Other factors may also weigh in on performance levels of CBT students, such as screen and font size, interline spacing and number of lines, whitespace, item review, resolution of graphics, scrolling and the presence of multiscreen, graphical, or complex displays (Leeson, 2006; Wang, Jiao, Young, Brooks, & Olson, 2007; Dadey, Lyons, & DePascale, 2018). For instance, Goodwin, Cho, Reynolds, Brady & Salas (2020) show that for longer sections of text, reading comprehension is stronger when reading on paper than on a screen. These factors may increase the complexity of the computer interface and, if they are not eliminated or recalculated when grading, there is the risk of evaluating other factors than the desired ones.

Another key factor to take into consideration regards examinees' computer familiarity. If CBT examinees are familiar and comfortable with typing on a computer, they exhibit better answering performance than PPT students when being evaluated for their writing ability (Russell & Plati, 2002). Therefore, inclusion of Information and Communication Technology (ICT) in classrooms is crucial for using CBT. Several studies also show how technology-aided education can be beneficial for students' performance and development of cognitive skills (Beg, Halim, Lucas, & Saif, 2022; Machin, McNally, & Silva, 2007; Cristia, Ibarrarán, Cueto,

Santiago, & Severín, 2017).

Heterogeneity effects should also be factored in, i.e., who benefits or not from the transition to CBT mode. (Bacher-Hicks, Goodman, & Mulhern, 2021) found that COVID-19 widened academic gaps between high and low SES students and will persist in future school years. To mitigate this, the Portuguese government launched a program in 2021 to distribute free computers and internet hot-spots for all students and teachers in public schools, which totals more than one million equipment. However, a large proportion of parents (20%) are refusing to accept these equipment.[3] These type of 1-to-1 initiatives tend to lower performance levels of students from low socio-economic background, increasing education inequality (Hall, Lundin, & Sibbmark, 2021), whereas for example online tutoring programmes tend to have positive impact (Carlana & Ferrara, 2021). Different genders also react differently to CBT. Azmat, Calsamiglia & Iriberri (2016) show that male students consistently underperform compared with the female ones. Martin (2009) argues that male students tend to prefer dynamic stimuli, therefore they may underperform when they lack dynamic media such as videos or simulations.

The current paper contributes to literature on computer-based vs. paper-based testing mode where students are subject to the same set-up, whereas most studies compare online vs. face-to-face tests. It also explores heterogeneity effects across gender, socio-economic background, parents' education, and exams' subjects.

[3] Victor Ferreira, "Escolas acumulam 200 mil computadores por entregar e pagos pelo PRR", *Público*, last accessed December 11, 2022, https://www.publico.pt/2022/11/15/sociedade/noticia/escolas-acumulam-200-mil-computadores-entregar-pagos-prr-2027697

III. Background and data collection

a) Institutional Setting

The *Instituto de Avaliação Educacional* (IAVE) is the institution responsible for planning, designing, and validating external evaluation instruments for primary and secondary education in Portugal, among other objectives. The national low stake-exams for primary education (*Provas de Aferição*) are one of IAVE's external evaluation instruments, "conducted with the goal of monitoring the development of the curriculum in different areas; provide detailed information to schools, teachers, parents, and students on their performance; and promote a timely pedagogical intervention given the specific difficulties of each student".[4] Since 2016 (except 2020 due to COVID-19), the evaluated students were from the 2nd, 5th, and 8th grade, covering all three cycles of primary education. In 2022, the following exams were conducted: Portuguese, Maths, Artistic Education and Physical Education for 2nd graders; Maths and Natural Sciences and Visual and Technological Education for 5th graders; and History and Geography, Portuguese Second Language, Portuguese, and Physical Education for 8th graders. These exams do not influence the students' final score on the subject; therefore, they are considered low-stake.

b) Set-up

In 2022, the low-stake exams took place in the months of May and June and had the duration of 90 minutes. These exams were blindly graded by teachers from different schools than the origin school of the student. The majority of the students (98,6%) did this exam on paper, while 3 841 students (1,4%) from 58 schools used CBT. The schools selected to implement the CBT mode in mainland Portugal were the ones with better broadband internet

[4] IAVE, "Informação-Prova: Provas de Aferição – 2º ano de escolaridade", *IAVE*, last accessed December 15, 2022, https://iave.pt/wp-content/uploads/2022/11/IP-PA-2-2023-1.pdf.

connection, i.e., the ones considered most technologically prepared. Regarding Portuguese schools in foreign countries, these volunteered to participate in the pilot-program. Therefore, this cannot be considered a randomised control experiment. Every invited school accepted to participate. Within each school, the school board selected the classes which would be subject to CBT mode, thus it cannot be guaranteed that the students were chosen at random.

Students wrote the exam in a room in their school[5] and received the exam at the same time. Computer-based students used IAVE's platform, which students could access both online and offline in case of connection failure. Prior to the exam, students had access to a set of exercises to get familiar with the digital format. The exams' questions were presented as single-item, that is, one item per webpage, whereas PPT students had more than one item per page in some cases. Computer-based students were able to highlight and delete the highlights on the page, increase or decrease zoom and use the magnifying glass tool from 2x to 8x. The students could skip questions, go forward and backwards on the exam as they pleased and had a progress tracking tool always visible. Contrarily to PPT students, CBT ones could not have an overall view of the exam, as they could only go to the immediately before or after question at each click, e.g., they were not able to go from question 1 to question 5 directly. For math questions, students had access to a tool to enable the writing of mathematical symbols (Figure A 1 in the appendix). In some open-ended and mathematical formulas questions, students were not allowed to use the "Enter" key to make paragraphs. Students could use the copy and paste keys on their keyboards (in all questions) or use the computer mouse for the same purpose (only in open-ended questions).

[5] According to IAVE, a small set of students did the exams in a different school due to lack of appropriate infrastructure on their home school.

c) Data

The data set is composed by 277 390 exams performed in 2022 by 172 299 students from 4 166 schools in mainland and islands of Portugal as well as Portuguese schools in foreign countries (Figure A 2 to Figure A 7 in the appendix).[6] These students are distributed among three grades: 49 609 from 2nd grade, 60 071 from 5th grade, and 62 619 from 8th grade. The exams are divided into completed exam on paper (PPT) and completed exam on computer (CBT), only diverging in terms of test format (Figure A 8 in the appendix). IAVE provided information on students' performance by exam item. Non-written exams, e.g., Artistic Education, were excluded from the analysis. For each student, there is data about their age, gender, parents' education[7], if they have special education needs, and social support from the government, which is divided in three levels (ASE_1, ASE_2, ASE_3)[8]; for schools, there is data about its type (public or private) and location.

d) Descriptive Statistics

In Table 1, we can observe that the majority of students are from public schools in mainland Portugal (Lisbon, North Centre or South). On average, 30% of students' fathers and 40% of mothers have a higher education degree and 30% of parents only have secondary education. Approximately 30% of the students are under school means tested programmes (ASE). The female and male gender are balanced within the dataset.

To be able to compare students' performance across different exams and grades, we create a standardised score variable. Firstly, we calculate the score per item of the exam, that takes the

[6] The students are divided into a treatment group if they performed the exam on computer and control group if they performed it on paper.

[7] Parents' education has a large number of missing values - 72 729 students for father's and 63 146 students for mother's -, so we dropped these observations in the main analysis but included in robustness checks.

[8] The first level students are the ones that receive most support from government, such as paid meals, free textbooks, budget for school materials, among others. Students without any level do not receive any kind of social support from the Portuguese government.

value 1 if the student answered the question correctly or partially correctly and the value 0 if the student answered wrongly or did not answer the question. Then, the standardised score was computed as the sum of the score of all exam items divided by the total number of items in the exam; therefore, the standardised score takes values between 0 and 1. For the exam of Mathematics and Sciences of the 5th grade, we were able to compare the standardised score with an alternative score that accounted for competency levels for a number of subject matter domains as defined by IAVE. The correlation between the two was equal to 0.94. Looking at the distribution of the standardised score, we observe that most students had scores above 50% (Figure A 9 in the appendix).

Table 1 - Descriptive Statistics

	Mean	Stdv	Min	Max
Age (years)	11.42	2.70	7	17
Gender (1 Male, 0 Female)	0.51	0.50	0	1
Type of School (1 Public, 0 Private)	0.89	0.32	0	1
Father College	0.27	0.44	0	1
Father Secondary	0.29	0.45	0	1
Mother College	0.38	0.48	0	1
Mother Secondary	0.31	0.46	0	1
Special needs	0.01	0.10	0	1
ASE level 1	0.13	0.33	0	1
ASE level 2	0.14	0.35	0	1
ASE level 3	0.02	0.15	0	1
Lisbon	0.20	0.40	0	1
North	0.33	0.47	0	1
Centre	0.22	0.42	0	1
South	0.21	0.41	0	1
Portuguese Islands	0.03	0.17	0	1
Foreign school	0.004	0.06	0	1
N = 277 390				

Note: This table presents means, standard deviation, minimum and maximum values of the variables used in the paper, for the full sample.

IV. Empirical Strategies

a) Pooled OLS

To estimate the test mode effect, students are divided into two groups: the treatment group that includes all students that did a computer-based exam; and the control group that includes

all students that did a paper-based exam (Figure A 10 and Figure A 11 in the appendix). In order to check if both groups are comparable, we performed a balance test (Table 2).

Table 2 - Balance test for full sample by treatment status

	Control	Treatment	Difference	P-value
Age (years)	11.417 (0.486)	12.079 (0.407)	0.663	0.106
Gender (1 Male, 0 Female)	0.514 (0.001)	0.522 (0.014)	0.008	0.557
Type of School (1 Public, 0 Private)	0.887 (0.009)	0.913 (0.036)	0.026	0.480
Father College	0.268 (0.006)	0.315 (0.043)	0.047	0.272
Father Secondary	0.291 (0.003)	0.289 (0.018)	-0.08	0.933
Mother College	0.378 (0.006)	0.413 (0.046)	0.036	0.443
Mother Secondary	0.306 (0.003)	0.294 (0.018)	-0.012	0.511
Special needs	0.011 (0.0007)	0.012 (0.011)	0.001	0.901
ASE level 1	0.127 (0.002)	0.109 (0.011)	-0.018	0.106
ASE level 2	0.142 (0.002)	0.156 (0.017)	0.013	0.435
ASE level 3	0.022 (0.003)	0.048 (0.018)	0.026	0.142
Lisbon	0.205 (0.011)	0.209 (0.093)	0.005	0.961
North	0.330 (0.012)	0.238 (0.11)	-0.092	0.406
Centre	0.222 (0.010)	0.215 (0.105)	-0.006	0.953
South	0.213 (0.010)	0.179 (0.077)	-0.033	0.670
Portuguese Islands	0.028 (0.004)	0.103 (0.042)	0.075	0.069*
Foreign school	0.004 (0.002)	0.055 (0.028)	0.052	0.061*
N = 277 390				

Note: This table presents means, standard deviations in parentheses, and differences between the treatment and comparison groups. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively.

The only variables with p-values lower than the 10% significance level are Portuguese Islands and Foreign school, therefore mitigating concerns that the two groups are not

comparable. To further be sure of this conclusion, we performed an Imbens test, that returned the same results, i.e., the variables Portuguese Islands and Foreign school are the only ones that are not balanced, as the value of the test is higher than the rule of thumb of 0.3 (Table A 1 in the appendix). For robustness checks[9], we also use a set-up with only mainland Portugal students. Table A 2 in the appendix reports the balance test of this sample. We observe that our groups are fairly balanced, except for the variables age and type of school.

For student *i* in school *s* and exam *e*, to estimate δ - the mean difference in exam scores between students who used CBT and students who used PPT – we use the following statistical model:

$$(1) \quad y_{eis} = \delta treatment_{ei} + \varphi_e + \beta X_{is} + \gamma K_s + \varepsilon_{eis}$$

where *y* is the exam's score, *treatment* is the binary variable that identifies the treatment status of the student and takes value 1 if the student performed exam *e* using CBT, φ_e are exam-specific fixed effects, *X* are control variables of students (age, gender, parents' education, if they have special education needs, social support from the government, and public/private type of school), *K* is a vector of binary variables that identify the school's region (NUTs II, district and municipality), and ε is the error term.

The first estimation strategy uses all students from all grades to identify the impact of performing the exam by computer. We use a Pooled OLS with and without controls. With this method, we are taking in all the available information across several exams. To capture the within geographical location variation and control for location-specific differences, such as regional unemployment, access to infrastructures like public libraries and socio-economic

[9] Here, we also include a set-up with the missing values of parents' education imputed as 0. The balance test (Table A 3 in the appendix) shows that the variables age, father and mother secondary, ASE level 1 and 3, Portuguese islands and foreign are not balanced.

disparities, we add region-specific[10] fixed effects to the regression model, estimating a Fixed Effects of District and a Fixed Effects of Municipality. Lastly, to capture within school variation, we use a School Fixed Effects. Standard errors were clustered at the school level. For robustness check, we run the following three specifications: 1. All students from all grades and locations, with robust standard errors;[11] 2. All students from all grades in mainland Portugal, with clustered standard errors at the school level;[12] 3. All students from all grades and locations, including students with missing values for parents' education, with clustered standard errors at the school level.[13]

b) Pooled OLS for students in mixed schools

Among the schools that participated in the pilot, there are some schools where there are both treated and non-treated students – which we call mixed schools. In this identification strategy, we only focus on these mixed schools as we believe that the students from these schools are more comparable among each other than the ones we are comparing in the previous identification strategy. In this specification, our data set is reduced to 2090 students in 11 schools, with 789 treated and 1301 non-treated students (Figure A 12 and Figure A 13 in the appendix), therefore it may be less representative, which may bias results. As before, we run a balance test to ensure comparability (Table A 4 in the appendix). The variables public school, father secondary, mother college, and ASE level 2 are imbalanced.[14]

To test the hypothesis that the aforementioned sample may be less representative, we

[10] The region-specific fixed effects encompass 20 districts and 308 municipalities in both mainland and Islands of Portugal.

[11] This sample includes 2 049 treated students and 170 250 control students.

[12] This sample includes 2 792 treated students and 265 545 control students.

[13] This sample includes 6 182 treated students and 430 765 control students.

[14] For robustness checks, we run the same regressions for the students in mixed schools and imputing the missing values of parents' education as 0, making our dataset composed by 5371 in 32 schools, with 2345 treated and 3026 non-treated students. In this specification, all variables are balanced between the treatment and control groups (Table A 6 in the appendix).

performed a balance test between our full sample and the mixed school's sample. Results are reported in Table A 5 in the appendix. We infer that the variables Special needs, Lisbon and North are imbalanced, for a 1% significance level. Special needs students represent overall 1% of the Portuguese students, therefore in terms of population it is a small imbalance. Apart from the location imbalances, results show that the mixed schools are representative of the full sample.

For this strategy, we run OLS specifications, like in equation (1), with and without controls as well as School Fixed Effects. Note that the latter returns similar results as the FE of School in the previous identification strategy as it is only capturing the within school variation, that is, only compares treated and non-treated students within the same school. However, in the FE of School using the full sample, the coefficients are estimated with higher precision as they are based on more observations.

c) Difference in differences

The last identification strategy relies on a difference in differences approach, taking advantage of the fact that students in the 8th grade took two exams: some on paper only, some on computer only, and some on both. The first DiD considers the following two groups: 8th graders that did both the History and the Portuguese exams on paper (57 380 students) and the ones that did the History exam on computer and the Portuguese exam on paper (88 students); the second DiD considers: 8th graders that did both exams on computer (814 students) and the ones that did the History exam on computer and the Portuguese exam on paper (88 students). It should be noted that, even though by using a DiD identification strategy we potentially obtain more robust results, the sample size for these DiDs is very reduced (only 88 students treated in 3 schools), harming the representativity of these results. One of the schools represented in the treatment group for this set-up is an Angolan private school. As this school may have quite different characteristics from other Portuguese schools, we run another DiD only including

schools from mainland Portugal.[15]

The first difference comes from differences in mean scores of History and Portuguese exam for students who did them on paper (computer) and the second difference comes from the differences in mean scores of History exam on computer and Portuguese exam on paper. In this identification strategy, we are only considering 8th graders since they are the only grade in which there is a relevant number of students that performed one exam on computer and another on paper.

For the first specification, Table A 7 in the appendix reports the balance test between groups, where we observe that the two groups are not balanced for the variables of gender, special needs, ASE level 2 and 3, Lisbon, North, Centre and Islands of Portugal. For the second specification, that is, the DiD comparing CBT students with students who did one exam on paper and another on computer, the two groups are imbalanced for the variables of gender, ASE level 2 and 3, North and Centre (Table A 8 in the appendix).[16]

For student *i* in school *s* and exam *e*, to estimate θ – the mean difference in exam scores between one exam on computer and another on paper – we use the following statistical model:

$$(2) \quad y_{eis} = \delta treatment87_i + \mu Exam87_e + \theta(treatment87_i * Exam87_e) + \beta X_{is} + \gamma K_s + \varepsilon_{eis}$$

where *y* is the exam's score, *treatment87* is the binary variable that identifies the treatment status of the student and takes the value 1 if the student did the History exam on computer and the Portuguese exam on paper, *Exam87* is a binary variable that takes the value 1 if the exam

[15] This sample includes 22 students who did the History exam on computer and the Portuguese exam on paper and the 56 034 students who performed both exams on paper.

[16] We use the same specifications for both data sets but including the missing values of parents' education as 0. Sample size increases to 312 "mixed students", i.e., students that did the Portuguese exam on paper and the History exam on computer, 173 724 PPT students, and 1 368 CBT students. For "mixed" vs. PPT students there are several imbalanced variables, such as mother secondary, special needs, Lisbon, North and Centre (Table A 8Table A 9 in the appendix). For "mixed" vs. CBT students, we do not reject the null hypothesis that the groups are balanced at a 1% significance level (Table A 10 in the appendix).

the student did was the History one and 0 if it was the Portuguese exam, *X* are control variables of students (age, gender, parents' education, if they have special education needs, financial support from the government, and public/private type of school), *K* is a vector of binary variables that identify the school's location (NUTs II, district and municipality), and ε is the error term.

The variable *treatment87* captures the mean difference in scores between treated and non-treated students, and *Exam87* captures the mean difference in scores between the Portuguese and History exam. The coefficient of the interaction term *treatment87*Exam87* captures the average treatment effect of doing the CBT version of the History exam for the first DiD group (treatment vs. PPT) and the average treatment effect of doing the PPT version of the Portuguese exam for the second DiD group (treatment vs. CBT).

V. Results

a) Results for pooled OLS

Table 3 - Estimated CBT effect, full sample

Treatment effect on exam's score of CBT students

	(1) Score	(2) Score	(3) Score	(4) Score	(5) Score
treatment	-0.088***	-0.087***	-0.088***	-0.092***	-0.056**
	(0.0121)	(0.0103)	(0.0107)	(0.0106)	(0.0278)
Observations	277 390	277 390	277 390	277 390	277 373
R-squared	0.321	0.404	0.406	0.414	0.469
Controls	NO	YES	YES	YES	YES
District FE	NO	NO	YES	NO	NO
Municipality FE	NO	NO	NO	YES	NO
School FE	NO	NO	NO	NO	YES

Note: This table presents estimates of the effects of CBT mode effect on exam's score. Each column corresponds to a separate regression. Standard errors, reported in parenthesis, are clustered at the school level. In regression (5), we drop 17 singleton observations. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively. Controls include age, gender, parents' education, if students they have special education needs, social support from the government, type of school (public or private) and location.

Table 3 reports results for the first estimation strategy. In all five specifications, the treatment effect of doing an exam through CBT mode on scores is negative. In the first four specifications, doing an exam through the computer decreases scores, on average, by approximately 9pp on a scale of 0 to 1, and the coefficients are significant at any level. When controlling for school fixed effects, the average treatment effect is approximately -6pp and it is only significant at a 5% level.

We run the same regressions with robust standard errors (Table A 11 in the appendix). The results show the same coefficients as before but much smaller standard errors. The regressions only considering mainland Portugal students return stronger negative effects (Table A 12 in the appendix): on average, students using CBT have between 5.5 and 9.5pp lower exam scores than PPT students, being significant at any level except when adding school fixed effects, where the coefficient is not significant (p-value higher than 10%).[17]

b) Results for pooled OLS for students in mixed schools

Table 4 - Estimated CBT effect, mixed schools' sample

Treatment effect on exam's score of CBT			
	(1)	(2)	(3)
	Score	Score	Score
treatment	-0.049*	-0.052**	-0.052**
	(0.0232)	(0.0200)	(0.0182)
Observations	2 090	2 090	2 090
R-squared	0.416	0.454	0.509
Controls	NO	YES	YES
School FE	NO	NO	YES

Note: This table presents estimates of the effects of CBT mode effect on exam's score. Each column corresponds to a separate regression. The sample includes exams of students from schools that had both treated and non-treated students. Standard errors, reported in parenthesis, are clustered at the school level. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively. Controls include age, gender, parents' education, if

[17] When including more 160 000 observations, that we previously dropped due to null values of parents' education and are now imputed as 0, we observe that the average treatment effect is around -8/9 pp (Table A 13 in the appendix).

students they have special education needs, social support from the government, type of school (public or private) and location.

Regarding the second identification strategy, where we are only comparing treated and non-treated students within the same category of schools, i.e., mixed schools, results show that the treatment effect is less strong than before. On average, CBT students have between 4.9pp and 5.2pp lower exams scores than PPT students. All coefficients are significant at a 10% significance level, and we observe that standard errors decrease as we add controls and School Fixed Effects (Table 4).[18]

c) Results for difference in differences

The last identification strategy consists of a difference in differences estimator based on 8th grade students that did both exams on paper (computer) and 8th grade students that did the History exam on computer and the Portuguese exam on paper.

Table 5 - Estimated CBT effect, DiD with PPT students as control group

Treatment effect on exam's score of CBT

	(1) Score	(2) Score
1.treatment87	-0.041***	-0.012
	(0.0125)	(0.0207)
1.Exam87	-0.241***	-0.241***
	(0.0014)	(0.0014)
1.treatment87#1.Exam87	-0.214***	-0.214***
	(0.0149)	(0.0149)
Observations	114 936	114 936
R-squared	0.412	0.481
Controls	NO	YES

Note: This table presents estimates of the effects of CBT mode effect on exam's score. Each column corresponds to a separate regression. The sample includes 8th grade exams of students that did the History exam on computer and the Portuguese exam on paper and students that did both exams on paper. Standard errors, reported in parenthesis, are clustered at the school level. Significance at the 1, 5 and 10 percent levels is indicated by ***, **

[18] When imputing missing values of parent's education as 0, the results show more negative treatment effects than before, between -5.4 and -6.1pp, with statistically significant coefficient levels (Table A 14 in the appendix).

and *, respectively. Controls include age, gender, parents' education, if students they have special education needs, social support from the government, type of school (public or private) and location.

Table 5 presents the estimated results for students who did the History exam on computer and the Portuguese one on paper vs. PPT students. Adding controls to the regression does not significantly change results. The treatment effect of doing the History exam on computer, captured by the interaction between the variable *treatment87* and *Exam87*, is of -21pp, on average, a much stronger effect than before.[19]

Table A 16 in the appendix presents the results for the same DiD specification, but only including students from mainland Portugal. We observe that all coefficients are significant at a 1% significance level and the treatment effect of doing the History exam through the computer is now more negative than before with a value of -24pp, on average. It should be noted that the sample size of the treatment group in these regressions is only 22 students in 2 schools (95% of these are from the same school), therefore, in the second regression we are only comparing within-student variation from the same district, as we add District Fixed Effects.

Table 6 reports results comparing doing both 8th grade exams (History and Portuguese) on computer and doing only the Portuguese one on paper. As with the previous specification, results do not differ when adding controls to the regression, except for the variable *treatment87*. The treatment effect of performing the Portuguese exam using PPT is only -3.4pp, with a coefficient significant at a 10% significance level in the first regression.[20]

[19] When we run the same regressions with the sample including parents' education missing values as 0, the treatment effect of CBT in History diminishes to -19 pp, on average. The coefficient is statistically significant at any level (Table A 15 in the appendix).

[20] Table A 17 in the appendix reports the same regressions for the sample including students with parents' education's missing values as 0. The treatment effect given by the interaction variable shrinks to -0.9pp but it is not statistically significant at any level.

Table 6 - Estimated CBT effect, DiD with CBT students as control group

Treatment effect on exam's score of CBT

	(1) Score	(2) Score
1.treatment87	-0.009	0.049***
	(0.0194)	(0.0110)
1.Exam87	-0.421***	-0.421***
	(0.0107)	(0.0107)
1.treatment87#1.Exam87	-0.034*	-0.034
	(0.0195)	(0.0196)
Observations	1 804	1 804
R-squared	0.629	0.664
Controls	NO	YES

Note: This table presents estimates of the effects of CBT mode effect on exam's score. Each column corresponds to a separate regression. The sample includes 8th grade exams of students that did the History exam on computer and the Portuguese exam on paper and students that did both exams on computer. Standard errors, reported in parenthesis, are clustered at the school level. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively. Controls include age, gender, parents' education, if students they have special education needs, social support from the government, type of school (public or private) and location.

VI. Heterogeneity effects

In addition to assess the overall impact of CBT on performance, it is also relevant to understand the students who are most affected by this mode. To do so, we run school fixed effects regressions by adding an interaction term of a heterogeneity variable with treatment to equation (1).

Figure 1 - CBT mode effect per exam

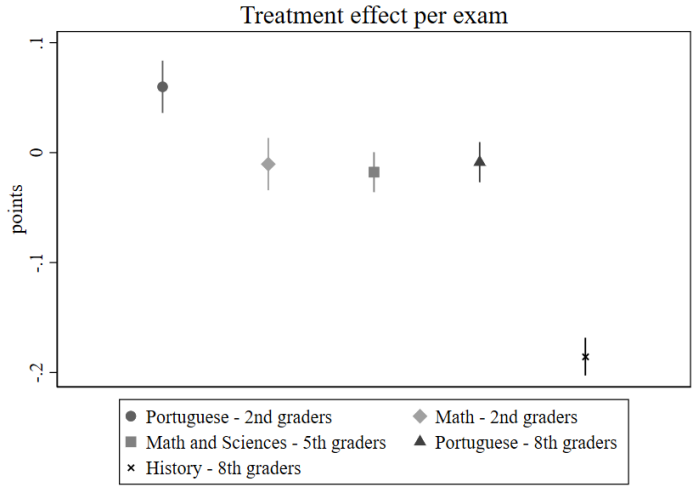
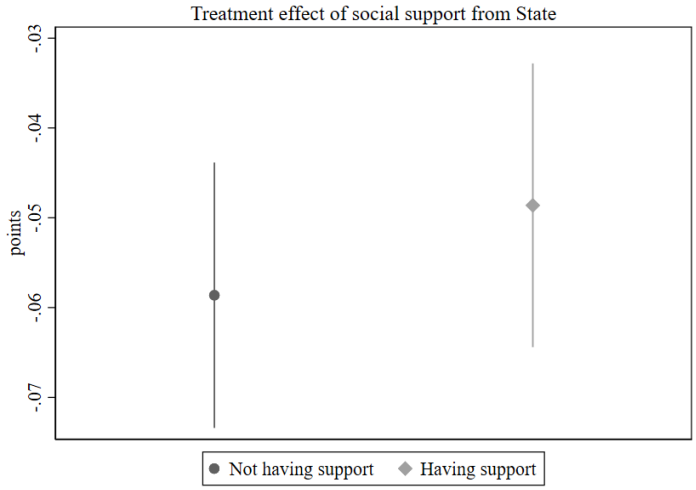


Figure 1 shows us the treatment effect per exam: two exams of 2nd grade (Portuguese, and Math), one exam of 5th grade (Math and Sciences), and two exams of 8th grade (Portuguese, and History). Surprisingly, the treatment effect is larger for older students than younger ones, contrarily to what we would expect, as 2nd graders tend to have far less computer experience than older graders. On the other hand, it may also be because exams' complexity increases as the school grade increases. 8th grade History exam CBT students are the most penalised of them all, with an average treatment effect of almost -20pp, which goes in line with the results from the DiD specification. The 2nd graders CBT students have a positive treatment effect for the Portuguese exam of approximately 6pp.

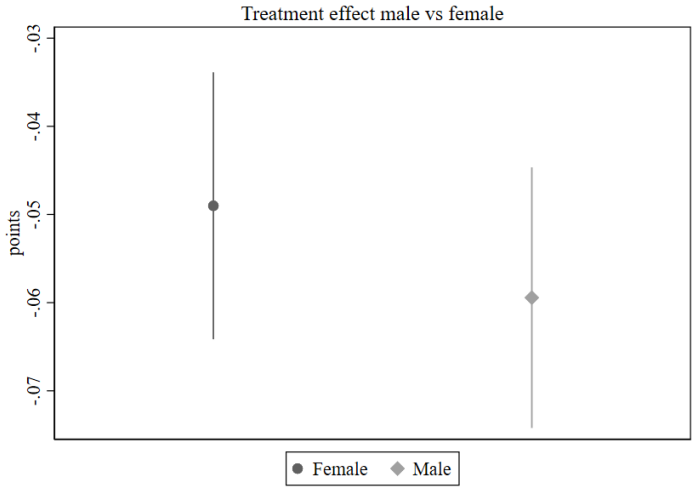
Figure 2 - CBT mode effect according to social support from State



We additionally measure the treatment effects by gender and observe that male students are 1pp more penalised than female ones when it comes to CBT (Figure 3), even though this difference is not statistically significant. These results go in line with the literature (Azmat, Calsamiglia, & Iriberri, 2016; Martin, 2009). It can also be justified with the fact that on average boys have fewer digital skills than girls. According to the European Commissions' Education and Training Monitor 2022, 30.8% of Portuguese 8th grade female students underachieve in

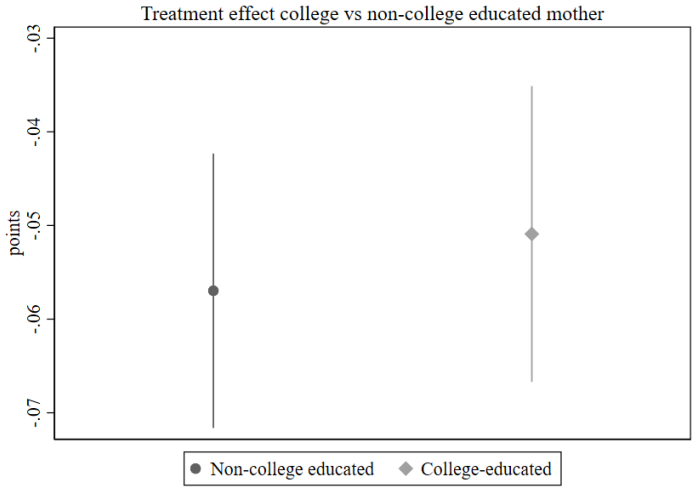
digital skills against 36% of males.[21]

Figure 3 - CBT mode effect per gender



Then, we compare the treatment effect for the students from higher and lower socio-economic background, measured by the access to social support in school. Students who receive social support seem to be marginally less penalised than those who do not, however, the difference is not statistically significant (Figure 2).

Figure 4 - CBT mode effect according to mother's education



[21] European Commission, "Digital Skills", *Education and Training Monitor 2022*, last accessed December 14, 2022, https://op.europa.eu/webpub/eac/education-and-training-monitor-2022/en/monitor-toolbox/digital-skills.html.

Lastly, we explore the CBT mode effect according to the student's mother education, specifically if the mother has a college degree, on exams' outcomes (Figure 4). As expected, having a college-educated mother penalises less students' performance than non-college educated ones. However, the difference between the treatment effect of both is small (less than 1pp).

VII. Conclusion and policy implications

The present study aimed to understand the score gap according to the test mode effect between computer-based and paper-based testing for Portuguese low-stake exams in the 2nd, 5th and 8th grade. The analysis is based on rich data including 172 299 students and 277 390 exams, as well as socioeconomic variables of students and schools. To assess the test mode effect, we use several econometric strategies. We also explore heterogeneity effects, with the purpose of understanding what are the policy implications for different segments of students. We found out that CBT students consistently underperform their PPT peers, with differences ranging from -5pp to -24pp, regardless of their grade or location and even when controlling for student-specific characteristics.

Our findings have significant policy implications that IAVE and other educational institutions should take into consideration before transitioning to computer-based testing. One of the reasons for the lower performance of CBT students may be the lack of familiarity with computers. In fact, not all Portuguese schools have the proper infrastructure, internet capacity, and human resources needed for this transition. In April 2021, it was estimated that a third of Portuguese teachers did not have the proper technological training to use computers in the context of the classroom.[22] In the beginning of the 2022 school year, there was also a lack of

[22] "Ensino à distância expôs a falta de formação tecnológica dos professores", *SIC Notícias*, last accessed

ICT professors, namely in Lisbon and Algarve.[23] Notwithstanding, the Portuguese government is highly investing on ICT on classrooms, under the digital transition goal of the Recovery and Resilience Plan. It is expected that until 2026 all students will use digital textbooks in classrooms and schools will have reliable and stable broadband internet connection, as well as new ICT classrooms adapted to the new reality of students having their own portable computer.[24] However, this timeline does not go hand in hand with IAVE's plans of implementing CBT mode for all national low-stake and high-stake exams by 2025.

Another important policy implication from this study is the exam design. Literature shows that CBT students exhibit a better performance when presented with multiple-item presentation, as it gives them the possibility of building off from other items' information, a "facilitating effect" that has a beneficial impact on performance levels (Leeson, 2006). Hence, the CBT exam structure may have penalised students.

Regarding heterogeneity effects, IAVE should consider how the implementation of this policy may affect different segments of students, specifically the gender differences. Our findings show that female students tend to outperform their male counterparts when using CBT. Therefore, this difference should be accounted for, and exam design should have type of questions that are appealing to these two genders, so that neither is at disadvantage (Martin & Binkley, 2009). This is particularly relevant if CBT is to be implemented in high-stake exams, as these strongly influence access to higher education and future career. Lastly, we should also

December 14, 2022, https://sicnoticias.pt/pais/2022-02-21-ensino-a-distancia-expos-a-falta-de-formacao-tecnologica-dos-professores.

[23] "Falta de professores: regiões de Lisboa e Algarve são as zonas mais afetadas", *SIC Notícias*, last accessed December 14, 2022, https://sicnoticias.pt/pais/2022-09-20-Falta-de-professores-regioes-de-Lisboa-e-Algarve-sao-as-zonas-mais-afetadas-d4989c0e.

[24] Cristina A. Ferreira, "Um milhão de computadores já chegaram às escolas. O que ainda falta mudar para termos uma Escola Digital?", *SAPO.pt*, last accessed December 14, 2022, https://tek.sapo.pt/noticias/computadores/artigos/um-milhao-de-computadores-ja-chegaram-as-escolas-o-que-ainda-falta-mudar-para-termos-uma-escola-digital.

keep in mind how transitioning to CBT may have differing effects on different subjects. We found that, for example, the negative CBT effect was significantly larger for the History exam compared with other exams. It may be that some subjects are more easily adaptive to CBT than others.

There are some limitations to this study. As the selection of schools and students to participate in the pilot-project was not random (selection bias), we used several estimation strategies to estimate the mode effect of CBT. However, the sample size for the pooled OLS with "mixed schools" and difference in differences approach was highly reduced, which may jeopardize the external validity of the results. The lack of data for students' previous performance levels does not allow us to check that the effects estimated are not also capturing previous differences in performance of students and schools, irrespective of the testing mode used. Furthermore, due to this caveat, we cannot guarantee the parallel trend assumption needed for the validity of the DiD specification. Further studies can focus on overcoming our limitations, namely by conducting randomised control experiments or quasi-random experiments. Increasing the sample size of treated students may also produce more accurate and generalized results.

VIII. References

Azmat, G., Calsamiglia, C., & Iriberri, N. (2016). "Gender Differences in Response to Big Stakes." *Journal of the European Economic Association*, 14(6), 1372-1400.

Bacher-Hicks, A., Goodman, J., & Mulhern, C. (2021). "Inequality in household adaptation to schooling shocks: Covid-induced online learning engagement in real time." *Journal of Public Economics*, 193, 104345–104361.

Backes, B., & Cowan, J. (2019). "Is the pen mightier than the keyboard? The effect of online testing on measured student achievement." *Economics of Education Review*, 68, 89-103.

Bagger, A., Norén, E., Boistrup, L., & Lundahl, C. (2019). "Digitized national tests in mathematics: a way of increaseing and securing equity?" *Proceedings of the Tenth International Mathematics Education and Society Conference*.

Beatty, A. E., Esco, A., Curtiss, A. B., & Ballen, C. J. (2022). "Students who prefer face-to-face tests outperform their online peers in organic chemistry." *Chemistry Education Research and Practice*(2), 464.

Beg, S., Halim, W., Lucas, A. M., & Saif, U. (2022). "Engaging Teachers with Technology Increased Achievement, Bypassing Teachers Did Not." *American Economic Journal: Economic Policy*, 14(2), 61-90.

Carlana, M., & Ferrara, E. L. (2021). "Apart but Connected: Online Tutoring and Student Outcomes during the COVID-19 Pandemic." *HKS Faculty Research Working Paper Series RWP21-001*.

Clariana, R., & Wallace, P. (2002). "Paper-based versus computer-based assessment: key factors associated with the test mode effect." *British Journal of Educational Technology*, 33(5), 593–602.

Cristia, J., Ibarrarán, P., Cueto, S., Santiago, A., & Severín, E. (2017). "Technology and Child

Development: Evidence from the One Laptop per Child Program." *American Economic Journal: Applied Economics*, 9(3), 295-320.

Dadey, N., Lyons, S., & DePascale, C. (2018). "The comparability of scores from different digital devices: A literature review and synthesis with recommendations for practice." *Applied Measurement in Education*, 31(1), 30-50.

European Comission. (2020). *Digital Education Action Plan 2021-2027: Resetting education and training for the digital age.* Brussels: COM(2020) 624 final.

Goodwin, A. P., Cho, S.-J., Reynolds, D., Brady, K., & Salas, J. (2020). "Digital Versus Paper Reading Processes and Links to Comprehension for Middle School Students." *American Educational Research Journal,* 57(4), 1837-1867.

Hall, C., Lundin, M., & Sibbmark, K. (2021). "A laptop for every child? The impact of technology on human capital formation." *Labour Economics*, 69(C), 101957.

Janke, S., Rudert, S. C., Petersena, Ä., Fritz, T. M., & Daumillerc, M. (2021). "Cheating in the wake of COVID-19: How dangerous is ad-hoc online testing for academic integrity?" *Computers and Education Open*, 2(100055).

Jeong, H. (2014). "A comparative study of scores on computer-based tests and paper-based tests." *Behaviour and Information Technology*, 33(4), 410-422.

Laborda, J. G. (2010). "Contextual clues in semi-direct interviews for computer assisted language testing." *Procedia Social and Behavioral Sciences*, 2, 3591-3595.

Lavy, V., Sand, E., & Shayo, M. (2022). "Discrimination Between Religious and Non-Religious Groups: Evidence from Marking High-Stakes Exams." *The Economic Journal*, 132(646), 2308-2324.

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.

Logan, T. (2015). "The influence of test mode and visuospatial ability on mathematics

assessment peformance." *Mathematics Education Research Journal,* 27(4), 423-441.

Machin, S., McNally, S., & Silva, O. (2007). "New Technology in Schools: Is There a Payoff?" *The Economic Journal,* 117(522), 1145-1167.

Martin, R., & Binkley, M. (2009). "Gender Differences in Cognitive Tests: a Consequence of Gender-dependent Preferences for Specific Information Presentation Formats?" *The transition to computer-based assessment,* 85-81.

Öz, H., & Özturan, T. (2018). "Computer-based and paper-based testing: Does the test administration mode influence the reliability and validity of achievement tests?" *Journal of Language and Linguistic Studies,* 14(1), 67-85.

Parhizgar, S. (2012). "Testing and Technology: Past, Present and Future." *Theory and Practice in Language Studies,* 2(1), 174-178.

Retnawati, H. (2015). "The comparision of accuracy scores on the paper and pencil testing versus computer-based testing." *TOJET,* 14(4), 135-142.

Russell, M., & Plati, T. (2002). "Does it matter with what I write? Comparing performance on paper, computer and portable writing devices." *Current Issues in Education, 5*(4).

Scheuermann, F., & Björnsson, J. (2009). *The Transition to Computer-Based Assessment: New Approaches to Skills Assessment and Implications for Large-scale Testing.* Luxembourg: European Commission - Joint Research Centre.

Wang, S., Jiao, H., Young, M., Brooks, T., & Olson, J. (2007). "A meta-analysis of testing mode effects in grade K-12 mathematics tests." *Educational and Psychological Measurement,* 67(2), 219-238.

Wang, T.-H., Kao, C.-H., & Chen., H.-C. (2021). "Factors Associated with the Equivalence of the Scores of Computer-Based Test and Paper-and-Pencil Test: Presentation Type, Item Difficulty and Administration Order." *Sustainability,* 13(17), 9548.

Wuthisatian, R. (2020). "Student exam performance in different proctored environments:

Evidence from an online economics course." *International Review of Economics Education*, 35, 100196.

IX. Appendix

Figure A 1 - Example of a mathematical question on CBT mode

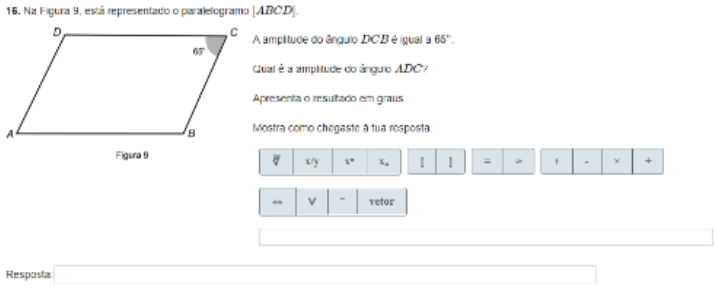


Figure A 2 - Distribution of CBT exams in mainland Portugal

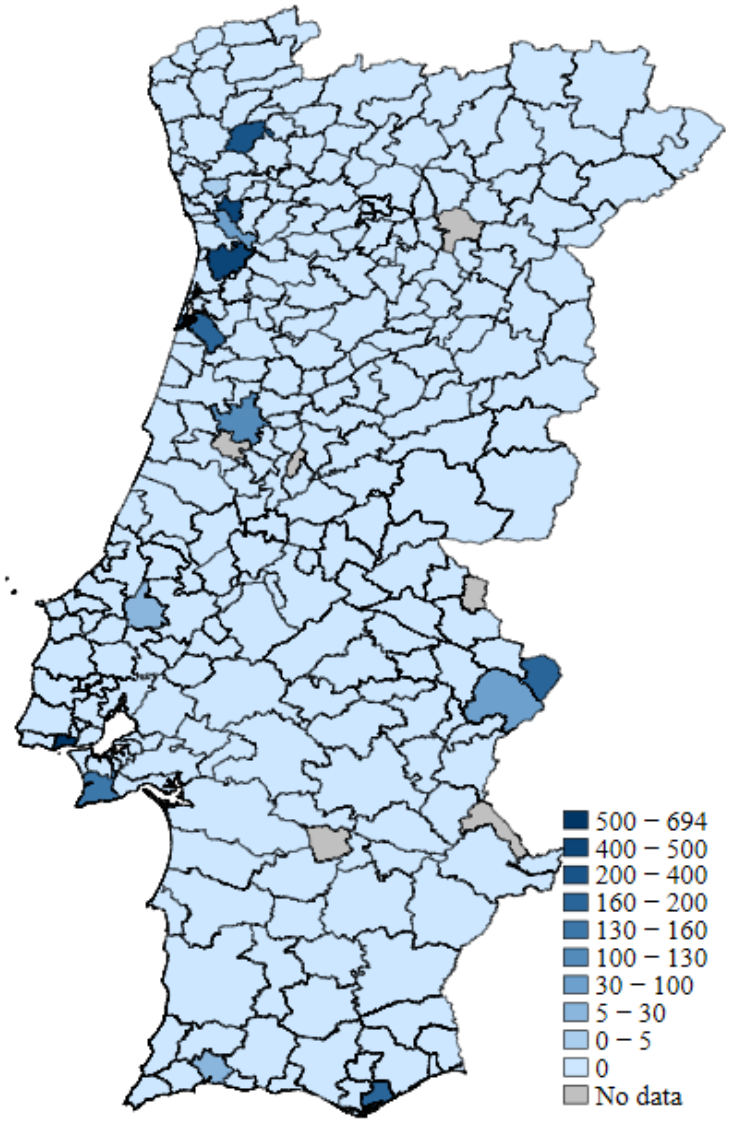


Figure A 3 - Distribution of schools with CBT exams in mainland Portugal

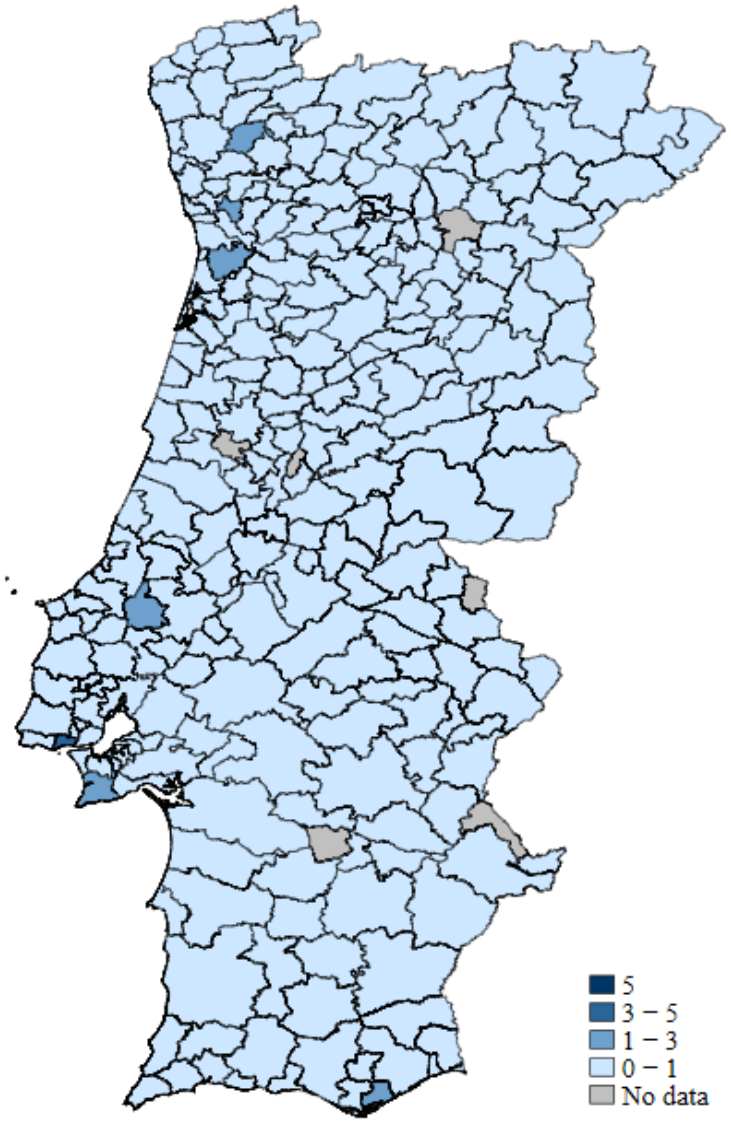


Figure A 4 - Distribution of CBT exams in Azores (Island of Portugal)

Figure A 5 - Distribution of schools with CBT exams in Azores (Island of Portugal)

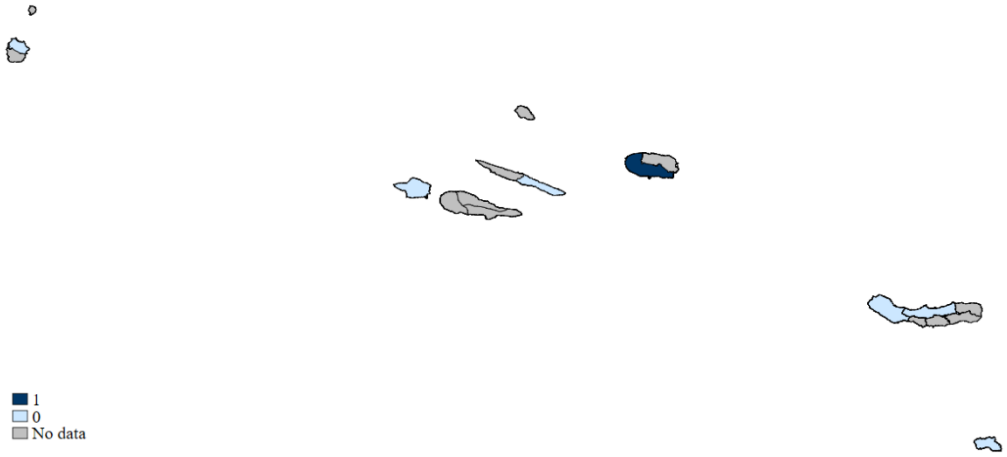


Figure A 6 - Distribution of CBT exams in Madeira (Island of Portugal)

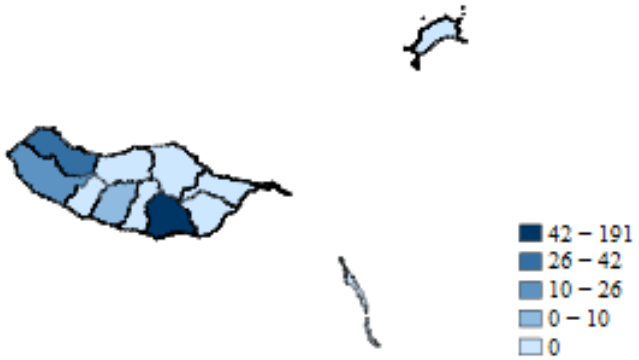


Figure A 7 - Distribution of schools with CBT exams in Madeira (Island of Portugal)

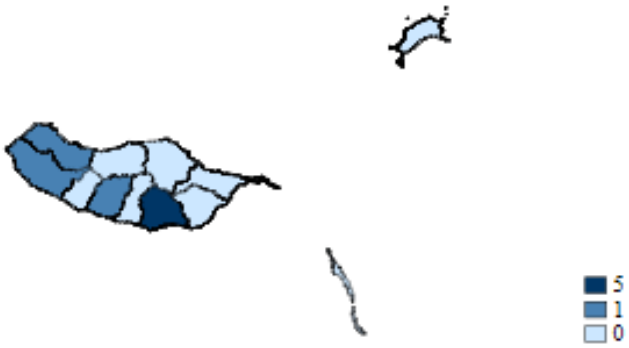
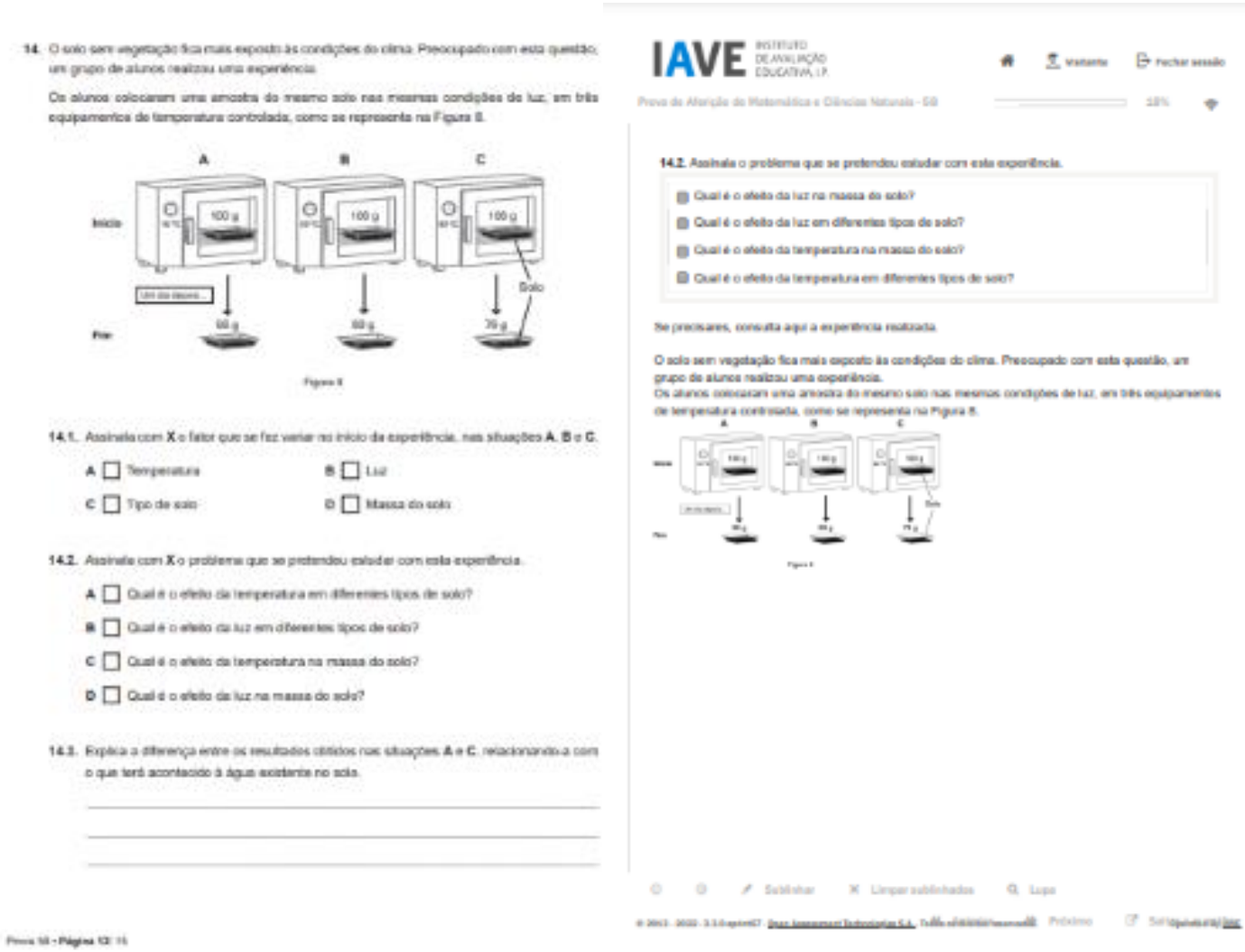


Figure A 8 - Example of PPT item (lhs) and CBT item (rhs)



14. O solo sem vegetação fica mais exposto às condições do clima. Preocupado com esta questão, um grupo de alunos realizou uma experiência.

Os alunos colocaram uma amostra do mesmo solo nas mesmas condições de luz, em três equipamentos de temperatura controlada, como se representa na Figura 8.

A B C

Início

100 g 100 g 100 g

Solo

Fim

60 g 60 g 70 g

Figura 8

14.1. Assinala com **X** o fator que se fez variar no início da experiência, nas situações **A**, **B** e **C**.

A ☐ Temperatura B ☐ Luz

C ☐ Tipo de solo D ☐ Massa do solo

14.2. Assinala com **X** o problema que se pretendeu estudar com esta experiência.

A ☐ Qual é o efeito da temperatura em diferentes tipos de solo?

B ☐ Qual é o efeito da luz em diferentes tipos de solo?

C ☐ Qual é o efeito da temperatura na massa do solo?

D ☐ Qual é o efeito da luz na massa do solo?

14.3. Explica a diferença entre os resultados obtidos nas situações **A** e **C**, relacionando-a com o que terá acontecido à água existente no solo.

Prova 10 - Página 12 | 15

IAVE INSTITUTO DE AVALIAÇÃO EDUCATIVA, I.P.

Anterior Fechar sessão

Prova de Aferição de Matemática e Ciências Naturais - 6º

18%

14.2. Assinala o problema que se pretendeu estudar com esta experiência.

- Qual é o efeito da luz na massa do solo?
- Qual é o efeito da luz em diferentes tipos de solo?
- Qual é o efeito da temperatura na massa do solo?
- Qual é o efeito da temperatura em diferentes tipos de solo?

Se precisares, consulta aqui a experiência realizada.

O solo sem vegetação fica mais exposto às condições do clima. Preocupado com esta questão, um grupo de alunos realizou uma experiência.
Os alunos colocaram uma amostra do mesmo solo nas mesmas condições de luz, em três equipamentos de temperatura controlada, como se representa na Figura 8.

Sublinhar Limpar sublinhados Lupa

Próximo

Table A 1 - Imbens Test for full sample

	Imbens Test
Age (years)	0.25
Gender (1 Male, 0 Female)	0.017
Type of School (1 Public, 0 Private)	0.087
Father College	0.104
Father Secondary	0.003
Mother College	0.073
Mother Secondary	0.0.27
Special needs	0.013
ASE level 1	0.055
ASE level 2	0.037
ASE level 3	0.144
Lisbon	0.011
North	0.205
Centre	0.015
South	0.083
Portuguese Islands	0.308*
Foreign school	0.309*
N = 277 390	

Figure A 9 - Distribution of standardised scores

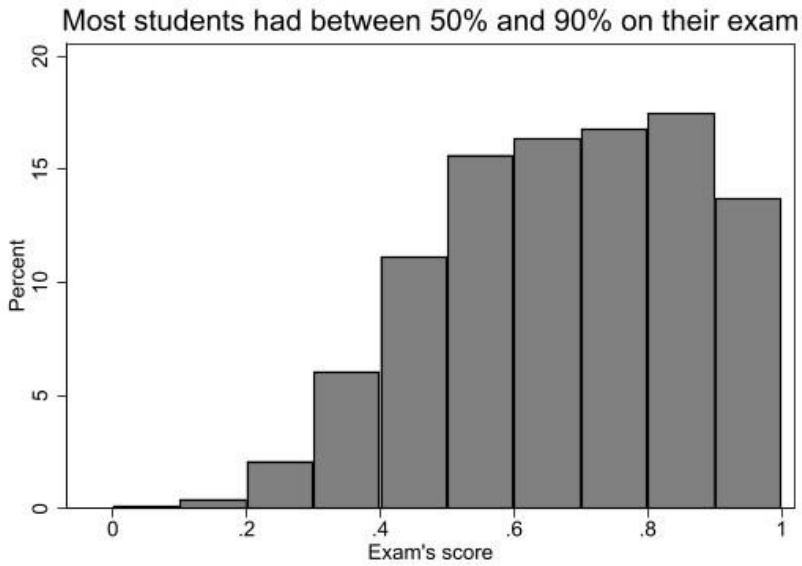


Figure A 10 - Number of students per treatment status

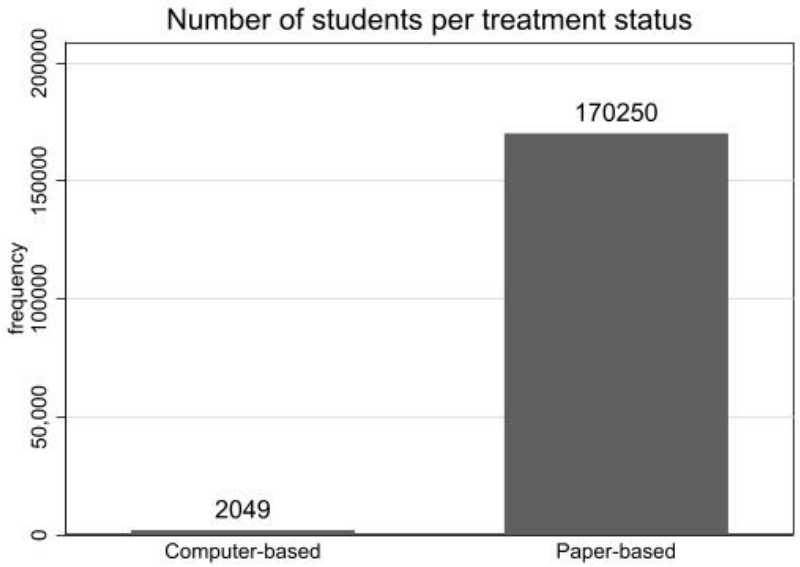


Figure A 11 - Number of students per exam and treatment status

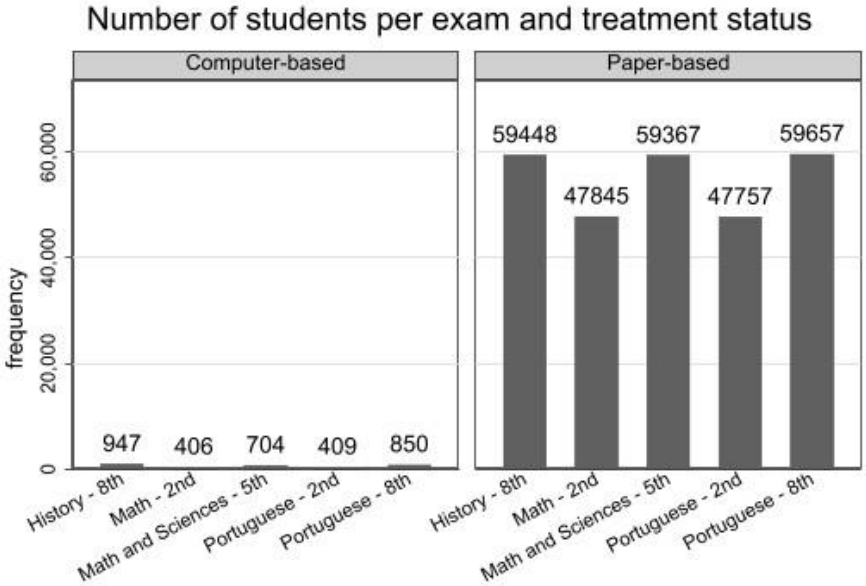


Table A 2 - Balance test, full sample only considering mainland Portugal

	Control	Treatment	Difference	P-value
Age (years)	11.437 (0.049)	12.375 (0.430)	0.937	0.030**
Gender (1 Male, 0 Female)	0.514 (0.001)	0.520 (0.016)	0.006	0.705
Type of School (1 Public, 0 Private)	0.891 (0.009)	0.970 (0.023)	0.079	0.001***
Father College	0.268 (0.006)	0.302 (0.047)	0.035	0.468
Father Secondary	0.292 (0.003)	0.293 (0.019)	0.002	0.927
Mother College	0.377 (0.006)	0.408 (0.053)	0.030	0.569
Mother Secondary	0.307 (0.003)	0.292 (0.019)	-0.015	0.434
Special needs	0.010 (0.0007)	0.013 (0.013)	0.003	0.830
ASE level 1	0.126 (0.002)	0.112 (0.012)	-0.014	0.253
ASE level 2	0.141 (0.002)	0.151 (0.017)	0.010	0.574
ASE level 3	0.015 (0.003)	0.021 (0.013)	0.006	0.641
Lisbon	0.211 (0.011)	0.249 (0.110)	0.037	0.736
North	0.341 (0.013)	0.283 (0.128)	-0.058	0.652
Centre	0.229 (0.011)	0.256 (0.123)	0.027	0.826
South	0.219 (0.011)	0.213 (0.093)	-0.006	0.947
N = 268 337				

Note: This table presents means, standard deviations in parentheses, and differences between the treatment and comparison groups. The sample includes all students from mainland Portugal. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively.

Table A 3 - Balance test for full sample, including parents' education's missing values imputed as 0

	Control	Treatment	Difference	P-value
Age (years)	11.341 (0.045)	11.956 (0.298)	0.616	0.042**
Gender (1 Male, 0 Female)	0.514 (0.001)	0.506 (0.009)	-0.007	0.445
Type of School (1 Public, 0 Private)	0.878 (0.008)	0.843 (0.046)	-0.035	0.458
Father College	0.201 (0.004)	0.193 (0.027)	-0.008	0.772
Father Secondary	0.216 (0.002)	0.178 (0.019)	-0.038	0.044**
Mother College	0.281 (0.005)	0.258 (0.034)	-0.024	0.497
Mother Secondary	0.238 (0.003)	0.182 (0.022)	-0.056	0.012**
Special needs	0.012 (0.0007)	0.016 (0.007)	0.003	0.641
ASE level 1	0.141 (0.002)	0.110 (0.011)	-0.031	0.006***
ASE level 2	0.138 (0.002)	0.132 (0.014)	-0.005	0.710
ASE level 3	0.023 (0.002)	0.085 (0.028)	0.062	0.029**
Lisbon	0.234 (0.010)	0.149 (0.064)	-0.085	0.187
North	0.296 (0.011)	0.224 (0.082)	-0.072	0.379
Centre	0.201 (0.009)	0.129 (0.066)	-0.071	0.283
South	0.215 (0.009)	0.170 (0.054)	-0.045	0.408
Portuguese Islands	0.045 (0.005)	0.259 (0.061)	0.214	0.000***
Foreign school	0.009 (0.003)	0.069 (0.025)	0.060	0.016**
N = 436 947				

Note: This table presents means, standard deviations in parentheses, and differences between the treatment and comparison groups. The sample includes students with null values for parents' education, now imputed as 0. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively.

Table A 4 - Balance test for mixed schools

	Control	Treatment	Difference	P-value
Age (years)	12.077 (0.556)	11.895 (0.917)	-0.182	0.890
Gender (1 Male, 0 Female)	0.559 (0.022)	0.556 (0.034)	-0.002	0.953
Type of School (1 Public, 0 Private)	0.664 (0.180)	0.918 (0.063)	0.254	0.069*
Father College	0.371 (0.111)	0.250 (0.052)	-0.122	0.179
Father Secondary	0.272 (0.051)	0.362 (0.036)	0.090	0.020**
Mother College	0.446 (0.095)	0.312 (0.044)	-0.134	0.081*
Mother Secondary	0.294 (0.041)	0.341 (0.025)	0.047	0.148
Special needs	0.005 (0.003)	0.000	-0.005	0.160
ASE level 1	0.118 (0.029)	0.118 (0.024)	-0.000	0.987.
ASE level 2	0.104 (0.026)	0.217 (0.030)	0.113	0.006***
ASE level 3	0.083 (0.046)	0.091 (0.056)	0.008	0.883
Lisbon	0.000	0.000	0.000	.
North	0.000	0.000	0.000	.
Centre	0.101 (0.102)	0.248 (0.211)	0.147	0.235
South	0.341 (0.170)	0.436 (0.202)	0.095	0.509
Portuguese Islands	0.354 (0.184)	0.261 (0.147)	-0.093	0.498
Foreign school	0.203 (0.157)	0.054 (0.053)	-0.148	0.200
N = 2 090				

Note: This table presents means, standard deviations in parentheses, and differences between the treatment and comparison groups. The sample includes exams of students from schools that had both treated and non-treated students. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively.

Table A 5 - Balance test for the mixed schools' sample and the full sample

	Mixed schools	Full sample	Difference	P-value
Age (years)	12.008 (0.330)	11.425 (0.048)	-0.584	0.077*
Gender (1 Male, 0 Female)	0.558 (0.018)	0.514 (0.001)	-0.044	0.013**
Type of School (1 Public, 0 Private)	0.760 (0.139)	0.887 (0.009)	0.127	0.355
Father College	0.325 (0.084)	0.269 (0.006)	-0.057	0.496
Father Secondary	0.306 (0.042)	0.291 (0.003)	-0.016	0.712
Mother College	0.395 (0.074)	0.378 (0.006)	-0.017	0.819
Mother Secondary	0.312 (0.033)	0.306 (0.003)	-0.006	0.856
Special needs	0.003 (0.002)	0.011 (0.0007)	0.008	0.000***
ASE level 1	0.118 (0.022)	0.127 (0.002)	0.009	0.692
ASE level 2	0.146 (0.025)	0.143 (0.002)	-0.004	0.877
ASE level 3	0.086 (0.040)	0.022 (0.003)	-0.064	0.110
Lisbon	0.000	0.205 (0.011)	0.205	0.000***
North	0.000	0.329 (0.012)	0.329	0.000***
Centre	0.157 (0.141)	0.222 (0.010)	0.065	0.644
South	0.377 (0.164)	0.212 (0.010)	-0.165	0.312
Portuguese Islands	0.319 (0.154)	0.029 (0.004)	-0.291	0.057*
Foreign school	0.147 (0.116)	0.004 (0.002)	-0.143	0.214
N = 279 480				

Note: This table presents means, standard deviations in parentheses, and differences between the treatment and comparison groups. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively.

Table A 6 - Balance test for the mixed schools' sample, including parents' education's missing values imputed as 0

	Control	Treatment	Difference	P-value
Age (years)	12.228 (0.415)	12.228 (0.449)	-0.001	0.999
Gender (1 Male, 0 Female)	0.519 (0.015)	0.521 (0.015)	0.002	0.935
Type of School (1 Public, 0 Private)	0.690 (0.130)	0.842 (0.074)	0.152	0.155
Father College	0.198 (0.060)	0.153 (0.029)	-0.046	0.378
Father Secondary	0.160 (0.028)	0.188 (0.035)	0.027	0.322
Mother College	0.241 (0.063)	0.210 (0.040)	-0.031	0.567
Mother Secondary	0.175 (0.032)	0.172 (0.033)	-0.004	0.898
Special needs	0.008 (0.004)	0.005 (0.003)	-0.003	0.499
ASE level 1	0.139 (0.024)	0.121 (0.018)	-0.018	0.331
ASE level 2	0.136 (0.024)	0.162 (0.021)	0.026	0.272
ASE level 3	0.115 (0.038)	0.082 (0.034)	-0.033	0.402
Lisbon	0	0	0	.
North	0.001 (0.008)	0.213 (0.142)	0.212	0.143
Centre	0.051 (0.051)	0.111 (0.104)	0.060	0.293
South	0.324 (0.124)	0.227 (0.105)	-0.096	0.312
Portuguese Islands	0.427 (0.130)	0.370 (0.113)	-0.057	0.621
Foreign school	0.197 (0.106)	0.078 (0.045)	-0.119	0.131
N = 5 371				

Note: This table presents means, standard deviations in parentheses, and differences between the treatment and comparison groups. The sample includes exams of students from schools that had both treated and non-treated students, and additionally students with null values for parents' education, now imputed as 0. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively.

Figure A 12 - Number of students per school category

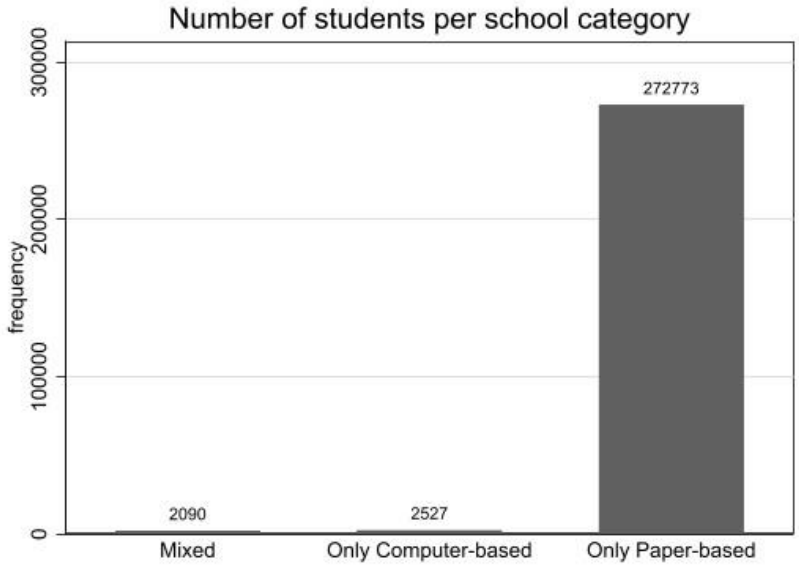


Figure A 13 - Number of schools per school category

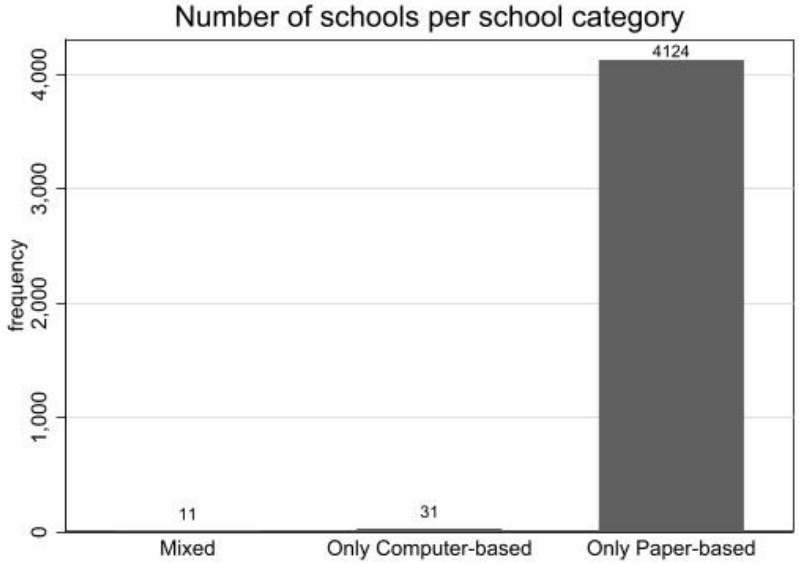


Table A 7 - Balance test for DiD with PPT students as control group

	Control	Treatment	Difference	P-value
Age (years)	14.175 (0.005)	14.159 (0.078)	-0.016	0.842
Gender (1 Male, 0 Female)	0.512 (0.002)	0.580 (0.030)	0.067	0.025**
Type of School (1 Public, 0 Private)	0.901 (0.010)	0.568 (0.344)	-0.333	0.333
Father College	0.244 (0.007)	0.523 (0.191)	0.278	0.145
Father Secondary	0.283 (0.004)	0.273 (0.089)	-0.010	0.906
Mother College	0.347 (0.008)	0.477 (0.243)	0.130	0.593
Mother Secondary	0.299 (0.004)	0.216 (0.069)	-0.083	0.230
Special needs	0.011 (0.001)	0.000	-0.011	0.000***
ASE level 1	0.116 (0.003)	0.125 (0.077)	0.009	0.908
ASE level 2	0.138 (0.003)	0.057 (0.035)	-0.081	0.022**
ASE level 3	0.019 (0.003)	0.000	-0.019	0.000***
Lisbon	0.197 (0.016)	0.000	-0.197	0.000***
North	0.341 (0.018)	0.000	-0.341	0.000***
Centre	0.228 (0.015)	0.011 (0.014)	-0.217	0.000***
South	0.210 (0.015)	0.557 (0.345)	0.347	0.314
Portuguese Islands	0.022 (0.006)	0.000	-0.022	0.000***
Foreign school	0.002 (0.001)	0.432 (0.344)	0.430	0.211
N = 114 936				

Note: This table presents means, standard deviations in parentheses, and differences between the treatment and comparison groups. The sample includes 8th grade exams of students that did the History exam on computer and the Portuguese exam on paper and students that did both exams on paper. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively.

Table A 8 - Balance test for DiD with CBT students as control group

	Control	Treatment	Difference	P-value
Age (years)	14.159 (0.081)	14.204 (0.029)	0.045	0.590
Gender (1 Male, 0 Female)	0.580 (0.031)	0.515 (0.023)	-0.065	0.082*
Type of School (1 Public, 0 Private)	0.568 (0.353)	0.960 (0.030)	0.392	0.260
Father College	0.523 (0.196)	0.275 (0.058)	-0.248	0.222
Father Secondary	0.273 (0.091)	0.276 (0.019)	0.003	0.969
Mother College	0.477 (0.250)	0.368 (0.061)	-0.109	0.662
Mother Secondary	0.216 (0.071)	0.299 (0.023)	0.083	0.265
Special needs	0.000	0.021 (0.020)	0.021	0.316
ASE level 1	0.125 (0.080)	0.097 (0.013)	-0.028	0.720
ASE level 2	0.057 (0.036)	0.156 (0.023)	0.099	0.030**
ASE level 3	0.000	0.035 (0.018)	0.035	0.062*
Lisbon	0.000	0.152 (0.127)	0.152	0.247
North	0.000	0.288 (0.143)	0.288	0.059*
Centre	0.011 (0.014)	0.325 (0.148)	0.313	0.041**
South	0.557 (0.354)	0.152 (0.102)	-0.405	0.260
Portuguese Islands	0.000	0.062 (0.040)	0.062	0.136
Foreign school	0.432 (0.353)	0.022 (0.023)	-0.410	0.240
N = 1 804				

Note: This table presents means, standard deviations in parentheses, and differences between the treatment and comparison groups. The sample includes 8th grade exams of students that did the History exam on computer and the Portuguese exam on paper and students that did both exams on computer. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively.

Table A 9 - Balance test for DiD with PPT students as control group, including parents' education's missing values imputed as 0

	Control	Treatment	Difference	P-value
Age (years)	14.222 (0.005)	14.308 (0.054)	0.086	0.112
Gender (1 Male, 0 Female)	0.510 (0.002)	0.500 (0.039)	-0.010	0.791
Type of School (1 Public, 0 Private)	0.893 (0.010)	0.804 (0.173)	-0.088	0.610
Father College	0.187 (0.006)	0.224 (0.088)	0.037	0.674
Father Secondary	0.215 (0.003)	0.125 (0.057)	-0.090	0.115
Mother College	0.263 (0.007)	0.170 (0.109)	-0.093	0.393
Mother Secondary	0.238 (0.004)	0.071 (0.040)	-0.167	0.000***
Special needs	0.012 (0.001)	0.000	-0.012	0.000***
ASE level 1	0.132 (0.003)	0.154 (0.055)	0.022	0.690
ASE level 2	0.135 (0.003)	0.083 (0.032)	-0.052	0.108
ASE level 3	0.023 (0.004)	0.022 (0.021)	-0.001	0.961
Lisbon	0.224 (0.015)	0.000	-0.224	0.000***
North	0.308 (0.016)	0.000	-0.308	0.000***
Centre	0.204 (0.013)	0.003 (0.003)	-0.200	0.000***
South	0.212 (0.014)	0.433 (0.212)	0.220	0.300
Portuguese Islands	0.045 (0.007)	0.369 (0.186)	0.323	0.082*
Foreign school	0.007 (0.002)	0.196 (0.173)	0.189	0.274
N = 174 348				

Note: This table presents means, standard deviations in parentheses, and differences between the treatment and comparison groups. The sample includes 8th grade exams of students that did the History exam on computer and the Portuguese exam on paper and students that did both exams on paper, and additionally students with null values for parents' education, now imputed as 0. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively.

Table A 10 - Balance test for DiD with CBT students as control group, including parents' education's missing values imputed as 0

	Control	Treatment	Difference	P-value
Age (years)	14.308 (0.054)	14.249 (0.031)	-0.058	0.314
Gender (1 Male, 0 Female)	0.500 (0.039)	0.494 (0.016)	-0.006	0.878
Type of School (1 Public, 0 Private)	0.804 (0.175)	0.857 (0.058)	0.053	0.753
Father College	0.224 (0.089)	0.168 (0.037)	-0.056	0.522
Father Secondary	0.125 (0.058)	0.165 (0.022)	0.040	0.486
Mother College	0.170 (0.111)	0.243 (0.047)	0.073	0.507
Mother Secondary	0.071 (0.041)	0.195 (0.030)	0.124	0.013**
Special needs	0.000	0.019 (0.012)	0.019	0.136
ASE level 1	0.154 (0.056)	0.103 (0.013)	-0.051	0.325
ASE level 2	0.083 (0.033)	0.126 (0.018)	0.043	0.226
ASE level 3	0.022 (0.022)	0.067 (0.030)	0.045	0.211
Lisbon	0.000	0.121 (0.089)	0.121	0.183
North	0.000	0.289 (0.107)	0.289	0.011**
Centre	0.003 (0.004)	0.208 (0.101)	0.204	0.045**
South	0.433 (0.215)	0.122 (0.067)	-0.310	0.130
Portuguese Islands	0.369 (0.188)	0.206 (0.068)	-0.162	0.381
Foreign school	0.196 (0.175)	0.053 (0.033)	-0.143	0.377
N = 3 360				

Note: This table presents means, standard deviations in parentheses, and differences between the treatment and comparison groups. The sample includes 8th grade exams of students that did the History exam on computer and the Portuguese exam on paper and students that did both exams on computer, including students with null values for parents' education, now imputed as 0. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively.

Table A 11 - Estimated CBT effect, full sample with robust standard errors

Treatment effect on exam's score of CBT students

	(1) Score	(2) Score	(3) Score	(4) Score	(5) Score
treatment	-0.088***	-0.087***	-0.088***	-0.092***	-0.056***
	(0.0033)	(0.0032)	(0.0032)	(0.0033)	(0.0072)
Observations	277 390	277 390	277 390	277 390	277 373
R-squared	0.321	0.404	0.406	0.414	0.469
Controls	NO	YES	YES	YES	YES
District FE	NO	NO	YES	NO	NO
Municipality FE	NO	NO	NO	YES	NO
School FE	NO	NO	NO	NO	YES

Note: This table presents estimates of the effects of CBT mode effect on exam's score. Each column corresponds to a separate regression. Robust standard errors are reported in parenthesis. In regression (5), we drop 17 singleton observations. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively. Controls include age, gender, parents' education, if students they have special education needs, social support from the government, type of school (public or private) and location.

Table A 12 - Estimated CBT effect, full sample only including mainland Portugal

Treatment effect on exam's score of CBT students

	(1) Score	(2) Score	(3) Score	(4) Score	(5) Score
treatment	-0.090***	-0.090***	-0.092***	-0.094***	-0.055
	(0.0133)	(0.0111)	(0.0115)	(0.0116)	(0.0374)
Observations	268 337	268 337	268 337	268 337	268 320
R-squared	0.321	0.403	0.406	0.413	0.468
Controls	NO	YES	YES	YES	YES
District FE	NO	NO	YES	NO	NO
Municipality FE	NO	NO	NO	YES	NO
School FE	NO	NO	NO	NO	YES

Note: This table presents estimates of the effects of CBT mode effect on exam's score. Each column corresponds to a separate regression. The sample includes all students in mainland Portugal. Standard errors, reported in parenthesis, are clustered at the school level. In regression (5), we drop 17 singleton observations. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively. Controls include age, gender, parents' education, if students they have special education needs, social support from the government, type of school (public or private) and location.

Table A 13 - Estimated CBT effect, full sample, including parents' education's missing values imputed as 0

Treatment effect on exam's score of CBT students					
	(1)	(2)	(3)	(4)	(5)
	Score	Score	Score	Score	Score
treatment	-0.077***	-0.078***	-0.078***	-0.087***	-0.075***
	(0.0095)	(0.0086)	(0.0090)	(0.0087)	(0.0178)
Observations	436 947	436 947	436 947	436 947	436 943
R-squared	0.315	0.392	0.394	0.401	0.464
Controls	NO	YES	YES	YES	YES
District FE	NO	NO	YES	NO	NO
Municipality FE	NO	NO	NO	YES	NO
School FE	NO	NO	NO	NO	YES

Note: This table presents estimates of the effects of CBT mode effect on exam's score. Each column corresponds to a separate regression. The sample includes students with null values for parents' education, now imputed as 0. Standard errors, reported in parenthesis, are clustered at the school level. In regression (5), we drop 4 singleton observations. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively. Controls include age, gender, parents' education, if students they have special education needs, social support from the government, type of school (public or private) and location.

Table A 14 - Estimated CBT effect, mixed schools' sample, including parents' education's missing values imputed as 0

Treatment effect on exam's score of CBT			
	(1)	(2)	(3)
	Score	Score	Score
treatment	-0.054***	-0.054***	-0.061***
	(0.0153)	(0.0136)	(0.0149)
Observations	5 371	5 371	5 371
R-squared	0.441	0.468	0.520
Controls	NO	YES	YES
School FE	NO	NO	YES

Note: This table presents estimates of the effects of CBT mode effect on exam's score. Each column corresponds to a separate regression. The sample includes exams of students from schools that had both treated and non-treated students, and additionally students with null values for parents' education, now imputed as 0. Standard errors, reported in parenthesis, are clustered at the school level. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively. Controls include age, gender, parents' education, if students they have special education needs, social support from the government, type of school (public or private) and location.

Table A 15 - Estimated CBT effect, DiD with PPT students as control group, including parents' education's missing values imputed as 0

Treatment effect on exam's score of CBT

	(1)	(2)
	Score	Score
1.treatment87	-0.042*	-0.021
	(0.0217)	(0.0277)
1.Exam87	-0.243***	-0.243***
	(0.0012)	(0.0012)
1.treatment87#1.Exam87	-0.187***	-0.187***
	(0.0194)	(0.0194)
Observations	174 348	174 348
R-squared	0.410	0.473
Controls	NO	YES

Note: This table presents estimates of the effects of CBT mode effect on exam's score. Each column corresponds to a separate regression. The sample includes 8th grade exams of students that did the History exam on computer and the Portuguese exam on paper and students that did both exams on paper, and additionally students with null values for parents' education, now imputed as 0. Standard errors, reported in parenthesis, are clustered at the school level. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively. Controls include age, gender, parents' education, if students they have special education needs, social support from the government, type of school (public or private) and location.

Table A 16 - Estimated CBT effect, DiD with PPT students as control group, with only mainland Portugal

Treatment effect on exam's score of CBT

	(1) Score	(2) Score
1.treatment87	-0.024***	0.014***
	(0.0057)	(0.0048)
1.Exam87	-0.241***	-0.241***
	(0.0014)	(0.0014)
1.treatment87#1.Exam87	-0.235***	-0.235***
	(0.0115)	(0.0115)
Observations	112 112	112 112
R-squared	0.409	0.479
Controls	NO	YES

Note: This table presents estimates of the effects of CBT mode effect on exam's score. Each column corresponds to a separate regression. The sample includes 8th grade exams of students in mainland Portugal that did the History exam on computer and the Portuguese exam on paper and students that did both exams on paper. Standard errors, reported in parenthesis, are clustered at the school level. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively. Controls include age, gender, parents' education, if students they have special education needs, social support from the government, type of school (public or private) and location.

Table A 17 - Estimated CBT effect, DiD with CBT students as control group, including parents' education's missing values imputed as 0

Treatment effect on exam's score of CBT

	(1)	(2)
	Score	Score
1.treatment87	-0.021	-0.015
	(0.0241)	(0.0286)
1.Exam87	-0.421***	-0.421***
	(0.0086)	(0.0086)
1.treatment87#1.Exam87	-0.009	-0.009
	(0.0214)	(0.0214)
Observations	3 360	3 360
R-squared	0.624	0.647
Controls	NO	YES

Note: This table presents estimates of the effects of CBT mode effect on exam's score. Each column corresponds to a separate regression. The sample includes 8th grade exams of students that did the History exam on computer and the Portuguese exam on paper and students that did both exams on computer, and additionally students with null values for parents' education, now imputed as 0. Standard errors, reported in parenthesis, are clustered at the school level. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively. Controls include age, gender, parents' education, if students they have special education needs, social support from the government, type of school (public or private) and location.