I Z A Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 18283

Can the Keyboard Beat the Good Old Pen? Computer-Based Testing and Students' Performance

Pedro Freitas
Luis Catela Nunes
Ana Reis
João Pereira dos Santos

NOVEMBER 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 18283

Can the Keyboard Beat the Good Old Pen? Computer-Based Testing and Students' Performance

Pedro Freitas
University of Oxford and Nova SBE Economics of Education Knowledge Center

Luis Catela Nunes
Nova School of Business and Economics

Ana Reis
Nova School of Business and Economics

João Pereira dos Santos
University of Lisbon and IZA

NOVEMBER 2025

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Can the Keyboard Beat the Good Old Pen? Computer-Based Testing and Students' Performance[*]

Computer-based testing is increasingly being adopted by educational institutions worldwide. However, whether this transition from paper-based testing leads to different outcomes in student performance remains an open question. This paper assesses the impact of computer-based testing by examining a large-scale pilot programme for low-stakes exams implemented in Portugal in 2022. We leverage rich student-level data to implement pooled OLS and difference-in-differences approaches. Our results indicate that students who used computer-based testing performed worse than their peers using paper-based testing by 5 to 14 percentage points, on average. This negative effect is concentrated in specific question formats, namely questions requiring the analysis of figures. We discuss the implications of our findings for the large-scale implementation of computer-based testing.

JEL Classification: I24, I28

Keywords: computer-based testing, paper-based testing, student performance

Corresponding author:
João Pereira dos Santos
ISEG - Univesity of Lisbon
Rua do Quelhas 6
1200-781 Lisbon
Portugal
E-mail: joao.santos@iseg.ulisboa.pt

[*] We would like to thank IAVE and DGEEC for providing the anonymised data used in this research. We thank IAVE's team: the President of the board, Luís Santos; the Teacher Training director, Rui Pires; and Paula Simões, Catarina Lains, and Marta Teixeira, for the immense support. The authors are also grateful to comments and suggestions by Pedro Pita Barros, Pedro Carneiro, Colin Green, Susana Peralta, Ana Balcão Reis, as well as by participants in the Lisbon Micro Group, the Eco research seminar at FEP (Porto), and the Lisbon Economics and Statistics of Education conference. This work was funded by Fundação para a Ciência e a Tecnologia (PTDC/EGE-ECO/4764/2021, UID/00124/2025, UID/PRR/124/2025, Nova School of Business and Economics DOI: https://doi.org/10.54499/UID/00124/2025) and LISBOA2030 (DataLab2030 - LISBOA2030-FEDER-01314200). All errors are our own.

1. Introduction

Computer-based testing (CBT) mode is becoming increasingly common across private and public education institutions worldwide (Parhizgar, 2012). In fact, many national and international institutions are planning to or are 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[5] and Sweden (Bagger, Norén, Boistrup, & Lundahl, 2019). 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 (see, *inter alia*, Agostinelli, Doepke, Sorrenti & Zilibotti, 2022; Rodríguez-Planas, 2022a, b; Werner & Woessmann, 2023).

There are several advantages of using CBT when compared with paper-and-pencil testing (PPT), namely: increase digital literacy of students and prepare them for a labour market that demands computer-related skills,[6] higher flexibility in test item design using video and animation-based questions; reduced scope for cheating (Backes & Cowan, 2019); increase in the efficiency of assessment, reducing "time lag" in reporting scores; cut paper consumption (Öz & Özturan, 2018); and lower potential biases on scoring, which can have long run implications for occupational choices and earnings at adulthood (Lavy, Sand, & Shayo, 2022).

Notwithstanding, CBT also poses some challenges such as high investment costs related with infrastructure and software quality and compatibility, the risk of testing candidates that are not used to the tool, hence testing other skills than intended or typing problems (Scheuermann & Björnsson, 2009). In the European Union (EU), recent data shows a disparity in digital education across Member States: access to broadband internet in 2019 varied from 74% of

[5] 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.

[6] Lakdawala, Nakasone, & Kho (2023) show that the impact of school-based internet access on Peruvian second graders' test scores was positive, but took some time to materialize.

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).

In this paper, we study the mode effect of CBT on low-stake national exams for Portuguese 2nd, 5th, and 8th graders, taking advantage of a pilot-project conducted by Portugal's Institute of Educational Assessment (*Instituto de Avaliação Educativa* - IAVE) 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. Our analysis includes socioeconomic background variables of students and schools, such as age, gender, parents' information, social support from school, among others. Since this pilot-programme was not implemented as a randomised control experiment, we exploit different but complementary identification strategies to assess the CBT mode effect, namely a pooled OLS with different samples, rich vectors of fixed effects and controls, and a student-level difference-in-differences approach. We find that students who used CBT performed worse than their paper counterparts by 5 pp to 14 pp, on average. These results are consistent across several econometric identification strategies, strengthening the robustness of our conclusions.

This paper contributes to the literature in three ways. First, to our knowledge, this is the first large-scale study of CBT mode effects using nationwide data collected after the Covid-19 pandemic. Previous research has mainly focused on online assessments in small-scale, often university-level, contexts. An important exception is Backes & Cowan, 2019 who study the impact of CBT in MA, U.S.A. in 2015 and 2016. Using a school-level difference-in-differences approach, this study finds negative effects associated with the transition to CBTs, effects that are attributable to the exam mode itself rather than to alternative explanations such as differential pre-treatment trends across schools or short-run transition dynamics between exam formats. The main difference is that while, in their case, they adapt the pen-and-paper testing (PPT) mode to the computer format, making it more difficult to isolate the CBT mode effect, in our case, IAVE kept the CBT and PPT exam questions as similar as possible in their pilot.

Second, we explore CBT mode effects across a broad range of subjects and school years, shedding light on how the impact of CBT may vary depending on the domain being assessed. Recognizing these heterogeneous effects is crucial for designing policies that mitigate potential negative consequences of transitioning to CBT.[7]

Third, we investigate how the transition from PPT to CBT exams may affect different types of questions in distinct ways. Specifically, we examine whether performance varies across question formats, including multiple-choice questions, questions requiring the analysis of or interaction with figures, and those requiring students to revisit earlier parts of the test to retrieve relevant information. This dimension has been largely overlooked in the existing literature comparing performance between paper-based and computer-based assessments. (Backes & Cowan, 2019; Öz & Özturan, 2018; Lavy, Sand, & Shayo, 2022).

This paper proceeds as follows. Section 2 summarizes the previous literature on the matter. Section 3 provides institutional background and describes the experiment set-up and data. Section 4 lays out the empirical strategy used. Section 5 presents the results, Section 6 explores the placebo exercises, Section 7 shows the heterogeneity effects and Section 8 suggests some potential mechanisms behind the results. Finally, Section 9 concludes.

[7] These reasons were put forward by the new Portuguese Government elected in March 2024 that decided to postpone the decision to run all 9th grade exams under CBT format.

2. Literature Review

Over the last two decades, several researchers tried to assess the impacts of different testing modes. Some studies indicate that students taking CBT outperform their peers taking PPT (Wang, Kao, & Chen 2021). Clariana & Wallace (2002) uses data from 105 first-year business undergraduates enrolled in a Computer Fundamentals course, concluding that CBT students performed better than the PPT ones. However, most of the literature supports PPT over CBT (Backes & Cowan, 2019; Wuthisatian, 2020), or finds no statistically significant differences between the two testing modes (Goodwin, Cho, Reynolds, Brady, & Salas, 2020; Öz & Özturan, 2018). For example, Beatty, Esco, Curtiss, & Ballen () found that students who preferred online CBT consistently underperformed compared to PPT students, for 305 second-year undergraduate students in Organic Chemistry.

One major 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 perform differently across testing modes due to several factors, such as item presentation. Some studies have concluded that CBT is more challenging than PPT (Laborda, 2010), but the majority have shown that these test modes are comparable (Logan, 2015; Jeong, 2014; Retnawati, 2015; Öz & Özturan, 2018). Nevertheless, the type of item and its presentation may still influence results. Wang, Kao & Chen (2021) show that CBT students performed better than PPT students in multiple-choice questions, but uncovered no statistical differences between CBT and PPT with single-item presentation, for 381 primary school fifth graders in northern Taiwan in the subject of Natural Sciences. Leeson (2006) finds that single-item presentation in CBT may lead students to answer hastily, resulting in 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 reading comprehension is stronger when reading on paper than on a screen for longer sections of text. These factors may increase the complexity of the computer interface and, if not accounted for or adjusted in the grading process, there is a risk of assessing unintended skills rather than the intended learning outcomes.

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 benefit students' performance and the development of cognitive skills (Beg, Halim, Lucas, & Saif, 2022; Machin, McNally, & Silva, 2007; Cristia, Ibarrarán, Cueto, Santiago, & Severín, 2017; Bai, Liu, & Su, 2023). These effects, however, are not always statistically significant (Hall & Lundin, 2024), and there also can be negative effects of online instruction (Alpert, Couch, & Harmon, 2016; Bettinger, Fox, Loeb, & Taylor (2017) or the presence of computers in the classroom (Carter, Greenberg, & Walker, 2017) on performance.

Heterogeneity in effects should also be considered, i.e., there may be unequal benefits from the transition to CBT mode. Hall, Lundin, & Sibbmark (2021) found that a laptop distribution programme led to increased inequality in student outcomes, whereas, for example, online tutoring programmes tend to have a positive impact (Carlana & Ferrara, 2021). Different genders may also react differently to CBT. Martin (2009) argues that male students tend to prefer dynamic stimuli and may therefore underperform dynamic media, such as videos or simulations, are lacking.

Equally relevant is the impact that the change from PPT to CBT may have on the quality of the information provided by external assessments. Figlio & Loeb (2011) explain how the existence of external assessment mechanisms can contribute to improvements in academic achievement. These results are in line with those found by McElroy (2023) for the US and Jacob & Lefgren (2004) for third and eight graders in Chicago. Using cross-country analysis, Hanushek & Woessmann (2021), Fuchs & Woessmann (2007), Bergbaeur, Hanushek, & Woessmann (2018) further reinforce the relevance of external assessment mechanisms for student learning. Additionally, students and families react to the feedback and information provided (Azmat & Iriberri, 2016; Bobba & Frisancho, 2016). If CBT influences how student learning is measured, it is likely to impact the feedback and information provided to students and their families.

3. Background and data

3.1 Institutional Setting

In Portugal, IAVE is the institution responsible for planning, designing, and validating external evaluation instruments for primary and secondary education in Portugal. The national low stake-exams for primary and low-secondary 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".[8]

[8] 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.

From 2016 until 2023 (with the exception of 2020), the students assessed were from the 2nd, 5th, and 8th grades. All exams were blindly graded by teachers from schools other than the students' own, and the results did not influence the students' final subject scores; therefore, these assessments are considered low-stakes.

3.2 IAVE's pilot

In 2022, the following exams were conducted: Portuguese, Maths, Artistic Education and Physical Education for 2nd graders; Maths & Natural Sciences and Visual & Technological Education for 5th graders; and History & Geography, Portuguese Second Language, Portuguese, and Physical Education for 8th graders. The majority of the students (98.6%) did this exam on paper, while 6 182 students (1.4%) from 75 schools used CBT[9].

The schools selected by IAVE to implement the CBT mode in mainland Portugal were chosen to ensure representation of students from different regions and socioeconomic backgrounds. These schools were required to have a reliable broadband internet connection to ensure that no technical issues disrupted the pilot. All invited schools agreed to participate. Regarding Portuguese schools abroad, participation in the pilot programme was voluntary. Taken together, these factors mean that the pilot cannot be considered a randomised control experiment. In addition, within each school, the school board selected the classes which would take the CBT version of the exam, so it cannot be guaranteed that student selection was random either.

Students wrote the exam in a room in their school, all beginning at the same time, with exactly 90 minutes to complete the tasks. CBT students used IAVE's platform, which could be accessed both online and offline in case of connectivity issues. Prior to the exam, these students

[9] In our analysis we excluded the Visual and Technological Education, Portuguese second language, and Physical Education.

were given access to a set of practice exercises prepared by IAVE familiarise them with the digital format.

Exams were designed to be as similar as possible across the two modes, with only minor differences. In the CBT version, questions were presented in single-item format, that is, one item per webpage, whereas in the PPT version, multiple questions could appear on a single page. CBT students could highlight and delete highlights, zoom in and out, and use a magnifying tool from 2x to 8x. They could skip questions, navigate forwards and backwards freely, and had access to a visible progress tracker. Unlike PPT students, CBT students could not have an overall view of the exam,and could only navigate to the immediately preceding or following question at each click, e.g., they could not jump directly from question 1 to question 5. For Math questions, students had access to a special tool to input mathematical symbols (as shown in Figure A1 in the Appendix). In some open-ended and formula-based questions, students were not allowed to use the "Enter" key to create 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). The CBT exams were validated by IAVE and the results made available to the students.

Although the structure of the CBT and PPT exams was largely equivalent, the exams differed across subjects in terms of question types. Question format has been extensively studied as a relevant driver of student performance - for example, how students with different characteristics perform under multiple-choice formats, Pekkarinen (2015), Riener & Wagner (2017), or Saygin & Atwater (2021). In Table 1, we present descriptive statistics on the distribution of questions types for each exam, namely: 1. the share of multiple-choice questions; 2. the share of questions involving the analysis or interaction with a figure; 3. the share of questions requiring students to turn the page or scroll (in the CBT case) to locate the information needed to answer.

Table 1 –Share of questions by type for each of the exams

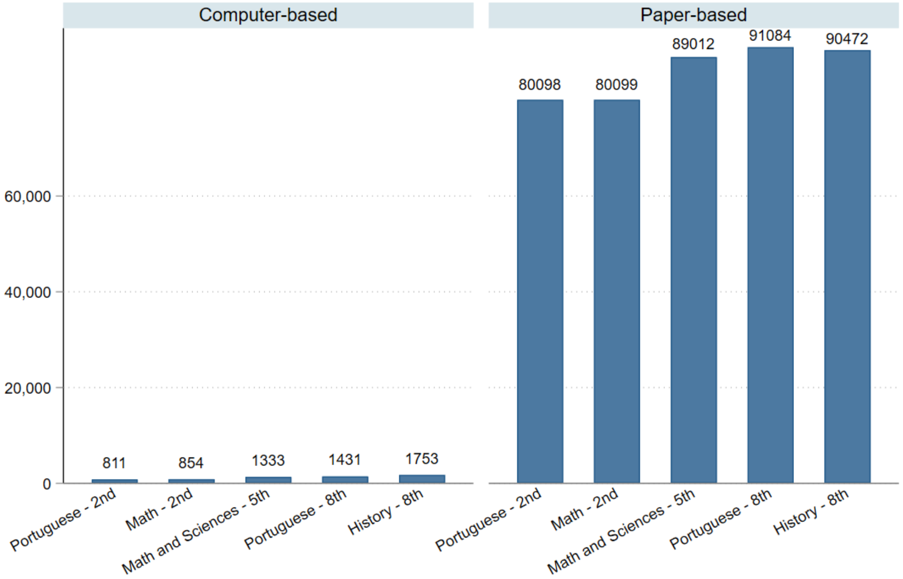
Exam	Multiple Choice	Figure	Scroll down/go back
Portuguese - 2nd grade	77%	46%	31%
Maths - 2nd grade	48%	7%	74%
Math and Science - 5th grade	58%	67%	0%
Portuguese - 8th grade	50%	54%	46%
History - 8th grade	88%	93%	18%

Note: The shares do not need to sum 100%, since each question maybe belong to one or more group type

3.3 Data

The data set is composed by 436 947 exams performed in 2022 by 270 019 students from 4 757 schools in mainland and islands of Portugal as well as Portuguese schools abroad. Figures A2 to A7 in the Appendix display the geographic distributions of both the number of schools and the number of exams in mainland Portugal, the archipelago of Azores, and the archipelago of Madeira. These students are distributed across three grades: 49 609 from 2nd grade, 60 071 from 5th grade, and 62 619 from 8th grade. The exams are divided into those completed on paper (PPT) and completed on computer (CBT), differing only in terms of test format (Figure A8 in the appendix). The number of students taking each exam is shown in Figure 1.

Figure 1 - Number of students per exam and treatment status



The dataset provided by IAVE covers the universe of students who took the exams and includes information on students' performance at the item level. For each student, we also have data on age, gender, parents' education[10], whether they have special education needs, and whether they receive social support from the government (classified into three levels of support - ASE levels 1 to 3).[11] Regarding schools, the dataset includes information on school type (public or private) and the municipality in which they are located. Parents' education is available for a smaller number of observations, around 270 thousand students.

We present the main descriptive statistics in Table 2. The gender distribution is balanced and the majority of students come from public schools in mainland Portugal. Around 30% of students receive social support (ASE levels 1 to 3)[12].

[10] 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 them in robustness checks.

[11] The first level students are the ones that receive most support from government, such as paid meals, free textbooks, and a fixed budget for school materials. As in 2025, the first/ second/ third level are measured considering family income until 3 363.01€/ 6 726.302€/ 11 434.01€. Level 3 corresponds to very limited support, not covering school meals or school materials. Remaining students do not receive other kinds of social support from the Portuguese government.

[12] In Appendix, Table A2, we present the descriptive statistics for the sub-sample which includes parents' schooling levels.

Table 2 - Descriptive Statistics

	Mean	**SD**	**Min**	**Max**
Age (years)	11.35	2.75	7	17
Gender (1 Male, 0 Female)	0.51	0.50	0	1
Type of School (1 Public, 0 Private)	0.88	0.20	0	1
Special needs	0.01	0.11	0	1
ASE level 1	0.14	0.35	0	1
ASE level 2	0.14	0.34	0	1
ASE level 3	0.02	0.15	0	1
Lisbon	0.23	0.42	0	1
North	0.30	0.46	0	1
Centre	0.20	0.40	0	1
South	0.21	0.41	0	1
Portuguese Islands	0.05	0.21	0	1
School abroad	0.01	0.10	0	1
	N= 436 947			

Note: This table presents the means, standard deviations, and minimum and maximum values of the variables used in the paper, based on the full sample of student-exam observations.

3.4 How different are students doing the computer-based exams?

The selection of schools by IAVE and the selection of students by school into a CBT constitute an important challenge when comparing the effect of this form of examination between those who did it (CBT) and those who responded to their exams with pen and paper (PPT). We take advantage of the rich administrative dataset provided by IAVE to compare the observed socioeconomic characteristics of the two groups and examine whether there are important differences between them. We present the results of this balance test in Table 3.

Table 3 - Balance test for full sample by treatment status

	PPT	CBT	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
Special needs	0.012 (0.001)	0.016 (0.007)	0.003	0.641
ASE level 1	0.141 (0.002)	0.11 (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 (0.01)	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)	0.17 (0.054)	-0.045	0.408
Portuguese Islands	0.045 (0.005)	0.259 (0.061)	0.214	0.000***
School abroad	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 control groups. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively.

We find small statistically significant differences in age, access o ASE (levels 1 and 3), and the share of schools in Portuguese Islands and Schools abroad. These last two differences are expected given that some of these schools volunteered to take part in the pilot. In Table A2 in the Appendix, we report the balance test when we restrict our sample to students in mainland Portugal. In this case, we find that selected students are slightly older and a small stronger prevalence of students with access to social support (level 1). In Tabel A3 we present a similar balance table but considering the subset of students from whom we have information on parental education. For this subset we find overall balanced samples between the students who wrote exam in paper and in a computer-based-format.

4. Empirical Strategies

In this section, we present the econometric specifications that we employ to study the effects of responding to a computer-based test when compared with those who continued to respond on a pen-and-paper format.

4.1 Pooled OLS

We start with a pooled OLS strategy considering all students who were evaluated in 2022. More specifically, for student i in school s and exam e, we estimate the following equation:

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

where y is the exam's score.[13] *CBT* is a binary variable that identifies the treatment status of the student, taking the value 1 if student i completed exam e using CBT. Therefore, δ is the parameter of interest, representing the *ceteris paribus* mean difference in exam scores between students who used CBT and students who used PPT. φ_e are exam-specific fixed effects, X are student-level control variables (including age, gender, special education needs status, receipt of social support, and school type – public or private). K is a vector of binary variables capturing

[13] These exams use a qualitative scale that we transform into a standardised score. First, we calculate the score per item. For multiple-choice questions we have considered a correct answer (=1) if the student answered the correct option. For open questions, we the value 1 if the student answered the question (partially) correctly and 0 if the student answered wrongly/ did not answer. Then, the standardised score was computed as the sum of all exam score items divided by the total number of items in the exam and is measured 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 variables was 0.98. Looking at the distribution of the scores, we observe that most students had scores above 50% and observe relevant distributional differences between the two exam modes (see Figure A9 in the Appendix).

school/region fixed effects (defined at the level of 20 districts, or 308 municipalities, or individual schools, depending on the specification). 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 disparities, we add region-specific fixed effects to the regression model, estimating models with district-level fixed effects and municipality-level fixed effects. Finally, ε denotes the error term. Across specifications, standard errors were clustered at the school level.

The results obtained were robust to many different specifications, namely: 1. all students from all grades and locations, with robust standard errors;[14] 2. all students from all grades in mainland Portugal, with clustered standard errors at the school level;[15] 3. students with information on parents' education.[16]

4.2 Within school effects in all and in mixed schools

Among the schools that participated in the pilot, some administered exams in both CBT and PPT formats simultaneously – we refer to these as mixed schools. Therefore, it is possible to estimate our regression model including school fixed effects. In that case, the parameter of interest will be identified from the exams taken in these mixed schools. The relevant variation comes from 5371 students in 32 mixed schools, with 2345 treated and 3026 non-treated exams (Figure A11 and Figure A12 in the Appendix).

Naturally, since the identification of the CBT mode effect in this case relies on a limited subset of exams and students, we cannot be confident that it is representative of the universe of all schools and students. To address this concern, we present two descriptive analyses to assess

[14] This sample includes 6 182 treated students and 430 765 control students.
[15] This sample includes 4 158 treated students and 407 733 control students.
[16] This sample includes 3 316 treated students and 274 074 control students.

the representativeness and balance of this restricted sample. To test the hypothesis that the mixed schools sample may not be representative of the average student in Portugal, we compare it with the full sample and report the results of this balance test in Table A4 in the Appendix. We observe imbalances in the variables Age, Special needs, and the regions Lisbon and North, as there were no mixed schools in these regions. Special needs' students represent 1% of the population, so this imbalance is relatively minor in terms of representativeness. We also conducted a balance test within the mixed schools' students to ensure the comparability between those students who were selected by these schools to do the exams on paper and on a computer. We display the results in Tabel A5 in Appendix. In this case, we do not find any significant imbalances.

4.3 Difference-in-differences

The last identification strategy relies on a difference-in-differences (DiD) approach, taking advantage of the fact that students in the 8th grade took two exams: some completed both on paper, some on computer only, and some used both formats depending on the exam. This setting allows us to estimate two DiDs.

The first uses the group of 8th grade students that completed the History exam on computer and the Portuguese exam on paper (88 students) together with the group that did both the History and the Portuguese exams on paper (57 380 students). In this case, we estimate the effect of changing the format of the History exam from a PPT to a CBT. The scores in the Portuguese PPT exam are used to control for differences between the two groups.

The second DiD uses the same group of 8th graders that took the History exam on computer and the Portuguese exam on paper (88 students), together with those who took both exams on computer (814 students). In this case, we estimate the effect of changing the

Portuguese exam from a CBT to a PPT. The scores in the History CBT exam are used to control for differences between the two groups.

It should be noted that, although using a DiD identification strategy may yield more robust results, since it implicitly involves a within-student analysis, the sample size for these DiDs is very small (only 88 treated students across 3 schools), which may limit the representativeness of the findings. One of the schools included in the treatment group for this set-up is a private school in Angola. As this school may differ significantly from Portuguese schools in terms of its characteristics, we conduct an additional DiD analysis that includes only schools located in mainland Portugal.[17]

For both DiD specifications, Table A6 and Table A7 in the appendix show overall balance between the treatment and control groups.

Formally, for both DiD specifications, we estimate the following equation for student *i* in school s and exam e, to estimate our parameter of interest θ:

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

where *y* is the exam score, *T* is a binary variable that identifies the group to which the student belongs, taking the value 1 if the student took the History exam on computer and the Portuguese exam on paper, and 0 otherwise. *Subject* is a binary variable that takes the value 1 for History and 0 for Portuguese in the first DiD specification, and, conversely, takes the value 0 for History and 1 for Portuguese in the second DiD. *X* are student-level control variables (age, gender, parents' education, special education needs, social support, and public/private type of school), *K* is a vector of binary variables identifying the school's location (NUTs II, district, and municipality), and ε is the error term.

[17] 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.

The variable *T* captures the mean difference in scores between the two groups, *Subject* captures the mean difference in scores between the History and Portuguese exams. The parameter of interest is the coefficient of the interaction term *T*Subject*. For the first DiD, it captures the average treatment effect of doing the CBT version of the History exam instead of the PPT version. For the second DiD, it captures the average treatment effect of doing the PPT version of the Portuguese exam instead of the CBT version.

5. Results

We now present the results for the three empirical strategies described in Section 4.

5.1 Results using the full sample

We start by presenting the results from estimating equation (1) for the full sample of students in Table 4.

Table 4 – CBT is associated with lower score

	(1) Score	(2) Score	(3) Score	(4) Score	(5) Score
CBT (δ)	-0.087***	-0.075***	-0.075***	-0.088***	-0.062***
	(0.0090)	(0.0094)	(0.0098)	(0.0091)	(0.0155)
Observations	436,947	436,947	436,947	436,947	436,943
R-squared	0.174	0.260	0.263	0.273	0.358
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. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively. Controls include age, gender, parents' education, whether students have special education needs, social support from the government, type of school (public or private), and location.

In all five specifications, the treatment effect on scores of doing an exam through CBT mode is negative. In the first four specifications, doing an exam through the computer decreases scores, on average, between 8 and 9 pp, on a scale of 0% to 100%, and the coefficients are significant at a 1% significance level.

If it is plausible that CBT students were positively selected from the population, then one could argue that the true causal effects would have been even more negative if these selection concerns were valid. However, the stability of the point estimates in columns (1) to (4) provides comforting evidence that selection on unobservable is unlikely to pose a significant threats to identification. In other words, these results appear robust to omitted variable bias, assuming that selection on a rich set of observables can be informative about selection on unobservables (Altonji, Elder, & Taber, 2005; Oster, 2019).

In column (5), we substitute regional by school fixed effects. In this setting, which corresponds to our preferred specification, the average treatment effect captures within school effects and it is approximately -6 pp. As expected, this coefficient is slightly more imprecisely estimated given that IAVE's pilot only included a small number of mixed schools.

For robustness, we re-estimated the previous regressions with robust standard errors (see Table A11 in the Appendix). As expected, standard errors are smaller. Furthermore, we also restricted the sample to students from mainland Portugal. Estimation results show similar negative effects (Table A12 in the appendix).[18]

5.2 Results for students in mixed schools

Next, we focus on students from mixed schools, i.e., schools where both testing formats were used, allowing us to compare treated and non-treated students within the same school environment. The results from estimating eq. (1) are presented in Table 5. We find that the size

[18] When we restrict the sample and include the information on parents' education, we observe that the average treatment effect is around -7/10 pp and very close to baseline results (Table A13 in the Appendix).

of the treatment is smaller than in the previous specifications. Once again, the results are stable across models with different sets of controls variables, and standard errors tend to decrease as controls and school fixed effects are added. On average, CBT students scored between 4 and 5 pp lower than PPT students.[19]

Table 5 – CBT is associated with lower scores for students from mixed schools

	(1) Score	(2) Score	(3) Score
CBT (δ)	-0.045***	-0.053***	-0.051***
	(0.0157)	(0.0144)	(0.0162)
Observations	5,371	5,371	5,371
R-squared	0.280	0.348	0.376
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 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 have special education needs, social support from the government, type of school (public or private), and location.

5.3 Difference-in differences results

The last identification strategy consists of two difference-in-differences estimations, using as treated group the 8th graders that took the History exam on computer and the Portuguese exam on paper. We compare their exam scores both with students who took both exams on paper and with those who took both exams on computer. On the one hand, this approach allows us to account for unobserved factors that are constant across the two exams (such as, for example, individual ability to use a computer). On the other hand, this strategy considers a small subset of treated students, so the results should be interpreted with this limitation in mind.

[19] When we restrict the sample and include the information on parents' education, the results are similar and the estimated average treatment effect is around (Table A14 in the Appendix).

Table 6 – Estimated CBT effect on the History exam, DiD with students doing both exams in PPT format as control group

	(1) Score	(2) Score
T (δ)	-0.031	-0.025
	(0.0308)	(0.0307)
Subject (μ)	-0.207***	-0.207***
	(0.0012)	(0.0012)
*T * Subject* (θ)	-0.145***	-0.145***
	(0.0267)	(0.0267)
Observations	174,348	174,348
R-squared	0.265	0.350
Controls	NO	YES

Note: This table presents estimates of the effects of CBT mode effect on the History 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 (*T*=1) and students that did both exams on paper (*T*=0). *Subject=1* for the History exam and *Subject=0* for the Portuguese exam. and 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 have special education needs, social support from the government, type of school (public or private) and location.

In Table 6, we present the estimated results using students who took both exams on paper as the control group. Once again, adding controls to the regression in column (2) does not significantly alter the results. We find that the average treatment effect of taking the History exam on computer, as captured by the interaction variable, is -15 pp, a stronger effect than previously observed.[20] However, these findings should be interpreted with caution given the small sample size of the treated group.

Table 7 presents the results using students who took both exams on computer as the control group. Consistent with the previous DiD specification, the estimated average treatment effect

[20] When we restrict the sample and include the information on parents' education, the treatment effect of CBT in History diminishes to -19 pp, on average. The coefficient is statistically significant at standard levels (Table A15 in the appendix).

of taking the Portuguese exam in PPT format rather than CBT is positive. However, the magnitude is substantially smaller, only 1pp.[21]

Table 7 – Estimated PPT effect on the Portuguese exam, DiD with students doing all exams in CBT format as control group

	(1) Score	(2) Score
T (δ)	-0.009	-0.021
	(0.0181)	(0.0155)
Subject (μ)	0.344***	0.344***
	(0.0105)	(0.0105)
*T * Subject* (θ)	0.009	0.009
	(0.0289)	(0.0289)
Observations	1 804	1 804
R-squared	0.497	0.565
Controls	NO	YES

Note: This table presents estimates of the effects of PPT mode effect on the Portuguese 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 (*T*=1) and students that did both exams on computer (*T*=0). *Subject=1* for the Portuguese exam and *Subject=0* for the History one. 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 have special education needs, social support from the government, type of school (public or private) and location.

6. Placebo exercises

Given some of the limitations discussed above, it is important to further strengthen the credibility and robustness of our results. To this end, we take advantage of the fact that our dataset includes information on scores from previous years to conduct two placebo exercises. The first exercise is based on students' prior scores, while the second relies on schools' past performance for the same grades. This approach allows us to assess two potential confounding

[21] When we restrict the sample and include the information on parents' education, the treatment effect of PPT format in Portuguese is around 5 pp, on average. The coefficient is statistically significant at standard levels (Table A16 in the appendix).

explanations for our findings.

6.1 Was CBT students' performance already worse in 2019?

One potential explanation for our results is that CBT students may have had lower scores than their peers even before the intervention. We can test this hypothesis, as the 5th grade students who performed the Mathematics exam in 2022 also sat in a Mathematics exam three years earlier, in the 2nd grade. Likewise, we can include in this analysis the students who took the 8th grade History exam, since these students also took a History exam three years earlier, in the 5th grade. For these students, we can estimate a placebo regression analogous to eq. (1).

In Table 8, we show that CBT students did not perform differently than their peers two years earlier, in 2019, as the point estimates are statistically indistinguishable from zero and precisely estimated. This suggests that these students were not selected based on their past performance in these subjects.

Table 8 – CBT students did not perform worse in 2019

	(1)	(2)	(3)	(4)	(5)
	Score	Score	Score	Score	Score
CBT (δ)	-0.002	-0.003	-0.002	-0.006	0.009
	(0.0085)	(0.0080)	(0.0079)	(0.0085)	(0.0090)
Observations	152,037	152,113	152,113	152,113	152,037
R-squared	0.040	0.112	0.115	0.133	0.296
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 in 2019. Each column corresponds to a separate regression. 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 have special education needs, social support from the government, type of school (public or private), and location.

6.2 Was CBT schools' performance worse in 2019?

Another potential explanation for our estimated effects is that CBT schools may have had, even before the intervention, lower average performance in the particular grades and courses tested. To examine this possibility, we use data from the previous edition of these exams, conducted in 2019. Naturally, the students that sat on these exams were not the same as those observed in 2022.

In Table 9, we show that CBT schools did not perform worse than PPT schools in 2019 for the same grades and subjects that were tested in 2022. We interpret these findings as additional reassuring evidence that the selection of CBT schools was not driven by prior performance in these exams.

Table 9 – CBT schools' performance was not worse than that of remaining schools

	(1) Score	(2) Score	(3) Score	(4) Score
CBT (δ)	-0.000 (0.0085)	-0.005 (0.0071)	-0.007 (0.0072)	-0.011 (0.0078)
Observations	331,451	331,451	331,451	331,451
R-squared	0.316	0.343	0.345	0.355
Controls	NO	YES	YES	YES
District FE	NO	NO	YES	NO
Municipality FE	NO	NO	NO	YES
School FE	NO	NO	NO	NO

Note: This table presents estimates of the effects of CBT mode effect on exam's score in 2019. Each column corresponds to a separate regression. 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 have special education needs, social support from the government, type of school (public or private), and location.

7. Heterogeneity effects

In addition to assessing the overall impact of CBT on performance, it is important to understand which types of students are most affected by the exam format. To investigate this, we estimate additional regressions by including in equation (1) an interaction term between the treatment variable and variables capturing potential heterogeneity in the treatment effect:

$$(3) \quad y_{eis} = \delta CBT_{ei} + \varphi_e + \beta X_{is}\varphi_e + \alpha CBT_{ei} * X_{is} + \gamma K_s + \varepsilon_{eis}$$

For this analysis, we focus on our preferred specification, which includes school fixed effects. We begin by estimating separate regressions that interact the treatment variable, CBT_{ei}, with student characteristics such as gender, whether the mother has college education, social support (ASE), and special education needs status. We find only small heterogeneous effects across these characteristics. Specifically, we observe that the effect of CBT is 1.2 pp more negative for male students. These results suggest that introducing computer-based assessments could potentially widen inequalities related to gender and student ability. However, the magnitude of these heterogeneous effects is relatively small, aligning with the findings by Backes & Cowan (2019).

To further explore whether the results differ across subjects, we re-estimate equation (3), this time interacting the treatment variable, CBT_{ei}, with indicator variables for different subjects. In Table 11 the coefficient on the treatment variable corresponds to the average effect of computer-based testing on the Portuguese test for 2nd graders. We observe that the interaction terms between the treatment variable and the subject indicators generally yield negative and statistically significant effects. However, the overall negative effect of CBT is particularly pronounced for the History exam in the 8th grade. In the following section, we investigate the mechanisms that might explain these substantial differences across subjects.[22]

[22] We re-estimate the baseline estimation presented in Table 3 without including the 8th grade History exam. The results are presented in Table A17 in the Appendix. They remain negative and significant, but with lower magnitude, as expected.

Table 10 – CBT, Heterogeneity effects

	(1) Score	(2) Score	(3) Score
CBT (δ)	-0.055***	-0.062***	-0.062***
	(0.0060)	(0.0156)	(0.0157)
Male (γ)	-0.007***		
	(0.0005)		
CBT#Male (α)	-0.012***		
	(0.0043)		
ASE (γ)		-0.050***	
		(0.0009)	
CBT#ASE (α)		0.001	
		(0.0081)	
Special Needs (γ)			-0.049***
			(0.0044)
CBT#Special Needs (α)			0.033
			(0.0305)
p-value(δ +α=0)	0.00	0.00	0.08
Observations	436,943	436,943	436,943
R-squared	0.357	0.357	0.357
Controls	NO	YES	YES
District FE	NO	NO	YES
Municipality FE	NO	NO	NO
School FE	YES	YES	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. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively. Controls include age, gender, parents' education, if students have special education needs, social support from the government, type of school (public or private), and location.

Table 11 – CBT by subject

	(1) Score
CBT (δ)	-0.006
	(0.0094)
Maths -2th grade (γ_1)	0.048***
	(0.0008)
Maths and Sciences – 5th grade (γ_2)	-0.060***
	(0.0058)
Portuguese – 8th grade(γ_3)	0.260***
	(0.0070)
History – 8th grade (γ_4)	0.054***
	(0.0070)
CBT#Maths -2th grade (α_1)	-0.024***
	(0.0083)
CBT#Maths and Sciences – 5th grade (α_2)	-0.024**
	(0.0100)
CBT#Portuguese – 8th grade (α_3)	-0.005
	(0.0103)
CBT#History -8th grade (α_4)	-0.137***
	(0.0099)
p-value($\delta + \alpha_1 = 0$)	0.001
p-value($\delta + \alpha_2 = 0$)	0.000
p-value($\delta + \alpha_3 = 0$)	0.121
p-value($\delta + \alpha_4 = 0$)	0.000
Observations	436,943
R-squared	0.357
Controls	YES
District FE	NO
Municipality FE	NO
School FE	YES

Note: This table presents estimates of the effects of CBT mode effect on exam's score. 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 have special education needs, social support from the government, type of school (public or private), and location.

8. Mechanisms

We further investigate the mechanisms that may drive the negative effects associated with CBT exams. We first explore how question format interacts with exam mode. Although this interaction is relatively underexplored in the literature, it represents an important dimension of exam design, particularly in identifying which types of questions are more susceptible to the transition from paper-based to computer-based testing.

Using the information presented in Table 1, we extend our dataset from the student-exam level to the student-exam-question level and estimate the following specification for exam question q in exam e, for student i in school s:

$$(4) \quad y_{qeis} = \delta CBT_{ei} + \gamma F_{qe} + \alpha F_{qe} CBT_{ei} + \varphi_e + \beta X_{is} + \gamma K_s + \varepsilon_{eis}$$

Here, y_{qeis} equals one if student i answered question q in exam e correctly. F_{qe} captures the question format: multiple-choice, interaction with a figure, or the need to scroll or search through the material to locate relevant information. The baseline category consists of open-ended questions that require neither figure interaction nor scrolling/searching. The remaining control variables are as previously defined in specification (1).

Results presented in Table 12 show that the negative impact of the CBT format is particularly pronounced for questions involving figures. On average, questions requiring interpretation or interaction with figures are 6pp lower less likely to be answered correctly in CBT exams compared to PPT exams. Conversely, questions that require students to recall or locate previously displayed information show a 6pp higher likelihood of being answered correctly under the CBT format. Importantly, these findings highlight that the transition from paper to computer-based exams may not affect all types of questions uniformly. Table 12 – Estimated CBT impact on students' performance by question type

Table 13 –Estimated CBT by type of question

	(1) Score
CBT (δ)	-0.018
	(0.0191)
Multiple Choice (γ_1)	0.151***
	(0.0013)
Figure (γ_2)	-0.031***
	(0.0014)
Go back (γ_3)	-0.072***
	(0.0012)
CBT#Multiple Choice (α_1)	-0.022**
	(0.0091)
CBT#Figure (α_2)	-0.060***
	(0.0121)
CBT#Go back (α_3)	0.058***
	(0.0091)
p-value($\delta + \alpha_1 = 0$)	0.00
p-value($\delta + \alpha_2 = 0$)	0.00
p-value($\delta + \alpha_3 = 0$)	0.01
Observations	12,353,845
R-squared	0.082
Controls	YES
District FE	NO
Municipality FE	NO
School FE	YES

Note: This table presents estimates of the effects of CBT mode effect on exam's score. 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 have special education needs, social support from the government, type of school (public or private), and location.

The variation in student performance across question types raises the question of whether lower performance is driven by different patterns of missing answers between the two exam modes. Previous literature has examined missing-answer behaviour, particularly in multiple-choice questions, focusing on how boys and girls respond differently to the presence of penalties for incorrect answers (Griselda, 2024; Saygin & Atwater, 2021; Riener & Wagner, 2017; Pekkarinen, 2015).

To explore this, we estimated specification (4) using as the outcome variable, y_{qeis}, a binary indicator equal to 1 if a student did not answer a question, and thus the response was missing. The results, presented in Table 13, indicate that exam mode—paper-based or computer-based—affects students' choices of leaving questions unanswered, which may be an important driver of performance differences between the two exam types[23]. Specifically, for multiple-choice questions, we estimate that the probability of leaving a question blank is 12pp higher in the computer-based exam. We also find a slightly higher probability of omitting questions involving figures, around 3 pp, and a smaller probability of leaving unanswered questions that require students to refer back to information, around 5 pp.

The results on the impact of exam mode on the probability of answering an item incorrectly or leaving it blank help to explain the pattern observed in Table 11, which shows a particularly large negative effect for the 8th-grade History exam— the subject with the highest concentration of multiple-choice and figure-based questions, as shown in Table 1. These findings indicate that exam mode may influence performance differently across question types. However, we are not yet able to identify the specific cognitive or behavioural mechanisms that cause students to perform differently across the two formats. Understanding these mechanisms are key to better understand which format types.

[23] To note that wrong answers did not have an additional penalty, so unanswered questions or questions wrongly answer received both zero points.

Table 13 –Estimated CBT impact on the probability of a blank answer question by type of question

	(1) Score
CBT (δ)	-0.030*
	(0.0176)
Multiple Choice (γ_1)	-0.205***
	(0.0017)
Figure (γ_2)	-0.012***
	(0.0008)
Go back (γ_3)	0.018***
	(0.0006)
CBT#Multiple Choice (α_1)	0.119***
	(0.0150)
CBT#Figure (α_2)	0.030***
	(0.0073)
CBT#Go back (α_3)	-0.052***
	(0.0078)
p-value($\delta + \alpha_1 = 0$)	0.00
p-value($\delta + \alpha_2 = 0$)	0.99
p-value($\delta + \alpha_3 = 0$)	0.00
Observations	12,353,845
R-squared	0.149
Controls	YES
District FE	NO
Municipality FE	NO
School FE	YES

Note: This table presents estimates of the effects of CBT mode effect on the probability. 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 have special education needs, social support from the government, type of school (public or private), and location.

We explore a third mechanism related with the quality of internet access in schools. Lower performance could be driven by inadequate internet connections, which may have compromised the adequate implementation of the computer-based assessments. Using data collected by the Portuguese Ministry of Education, we classify schools according to internet speed: Speed 1 – 32 Mbps; Speed 2 – 64 Mbps; Sped 3 – 100 Mbps; Speed 4 – 200 Mbps; Speed 5 – 1024 Mbps. This analysis is more limited, as we have information for only 29 of the 75 schools that implemented computer-based testing. Based on results in Table 14, we do not observe sizeable that schools with slower internet connection drove the observed result.

Table 14 – Estimated CBT impact by internet quality in school

	(1) Score
CBT (δ)	-0.039**
	(0.0179)
Speed 2 (γ_1)	-0.003
	(0.0048)
Speed 3 (γ_2)	-0.002
	(0.0050)
Speed 4 (γ_3)	0.009
	(0.0056)
Speed 5 (γ_3)	0.013*
	(0.0072)
CBT#Speed 2 (α_1)	-0.044**
	(0.0191)
CBT#Speed 3 (α_2)	-0.032
	(0.0263)
CBT#Speed 4 (α_3)	-0.037*
	(0.0192)
CBT#Speed 5 (α_4)	-0.074***
	(0.0245)
p-value($\delta + \alpha_1 = 0$)	0.00
p-value($\delta + \alpha_2 = 0$)	0.01
p-value($\delta + \alpha_3 = 0$)	0.00
p-value($\delta + \alpha_4 = 0$)	0.00
Observations	353,524
R-squared	0.250
Controls	YES
District FE	NO
Municipality FE	YES
School FE	NO

Note: This table presents estimates of the effects of CBT mode effect on exam's score. Standard errors, reported in parenthesis, are clustered at the school level. Internet speed variables at school level given by: Speed 2 – 64 Mbps; Speed 3 – 100 Mbps; Speed 4 – 200 Mbps; Speed 5 – 1024 Mbps. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively. Controls include age, gender, parents' education, if students have special education needs, social support from the government, type of school (public or private), and location.

9. Conclusion and policy implications

This study investigates the impact of computer-based testing by leveraging a large-scale pilot-programme for Portuguese exams in the 2nd, 5th, and 8th grades, implemented in the aftermath of the pandemic. The analysis draws on rich data covering 270 019 students and 436 947 exams, as well as detailed socioeconomic information at both the student and school levels. To assess the effect of test mode, we use several econometric strategies. We find that CBT students consistently underperform their PPT peers, with differences ranging from -5 to -14 pp, even when controlling for student-specific characteristics.

Our findings have significant policy implications that should be considered by Portuguese and international institutions before transitioning to computer-based testing. One possible explanation for the lower performance of CBT students is a lack of familiarity with computers. Not all Portuguese schools have the necessary infrastructure, internet capacity, and human resources to support this transition. For example, in April 2021, it was estimated that one third of Portuguese teachers lacked adequate technological training for classroom computer use,[24] and at the start of the 2022 school year, there was a shortage of ICT teachers.[25] Notwithstanding, the Portuguese government is investing highly in ICT in classrooms, as part of the digital transition goal of the Recovery and Resilience Plan. By 2026, it is expected that students will progressively adopt digital textbooks, schools will have reliable broadband internet, and new ICT classrooms will be adapted to the reality of students having their own portable computer.[26]

[24] "Ensino à distância expôs a falta de formação tecnológica dos professores", *SIC Notícias*, last accessed December 14, 2022, https://sicnoticias.pt/pais/2022-02-21-ensino-a-distancia-expos-a-falta-de-formacao-tecnologica-dos-professores.

[25] "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.

[26] 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.

Another important policy implication concerns exam design. The literature suggests that CBT students perform better when multiple items are presented on the screen, 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). Our analysis goes further by examining the impact of different question types, finding that some of the estimated negative effects are concentrated in questions requiring interaction with or analysis of figures, particularly in comparison to other types, such as open-ended questions or those that ask students to locate information from earlier sections of the exam. However, further research on the details of the interaction between the student and the exam under both formats is needed to fully understand how different question types are differently affected by the change of exam mode.

Regarding heterogeneity, policymakers should consider how the implementation of CBT may affect different segments of the student population, potentially amplifying pre-existing inequalities. Our results indicate some negative effects are concentrated on male students and students with special education needs.

There are, however, some limitations to this study. The selection of schools and students to participate in the pilot project was not random. To circumvent this limitation, we employed several estimation strategies to address potential selection bias. Nevertheless, the sample size for the pooled OLS with "mixed schools" and difference-in-differences approach was substantially reduced, which may limit the external validity of the results. Future research could address these limitations by conducting randomized control trials, or by increasing the sample size of treated students to produce more accurate and generalizable findings.

Finally, we highlight that this paper does not seek to recommend for or against transitioning to CBT. Rather, it describes the short-run impacts of moving to CBT and the need for policies that guarantee an adequate transition between test modes. A full cost-benefit analysis is beyond the scope of this paper, as we do not consider the likely benefits associated with online testing, such as lower administrative costs.

References

Alpert, W. T., Couch, K. A., & Harmon, O. R. (2016). "A randomized assessment of online learning. *American Economic Review*." 106(5), 378-382.

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools." *Journal of Political Economy*, 113(1), 151-184.

Agostinelli, F., Doepke, M., Sorrenti, G., & Zilibotti, F. (2022). When the great equalizer shuts down: Schools, peers, and parents in pandemic times. *Journal of Public Economics*, 206, 104574.

Azmat, G., & Iriberri, N. (2010). “The importance of relative performance feedback information: Evidence from a natural experiment using high school students”. *Journal of Public Economics*, 94(7–8), 435–452

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. (n.d.). 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.

Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics, 86*(1), 226-244.

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.

McElroy, K. (2023). Does test-based accountability improve more than just test scores? *Economics of Education Review, 94*, 102381.

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.

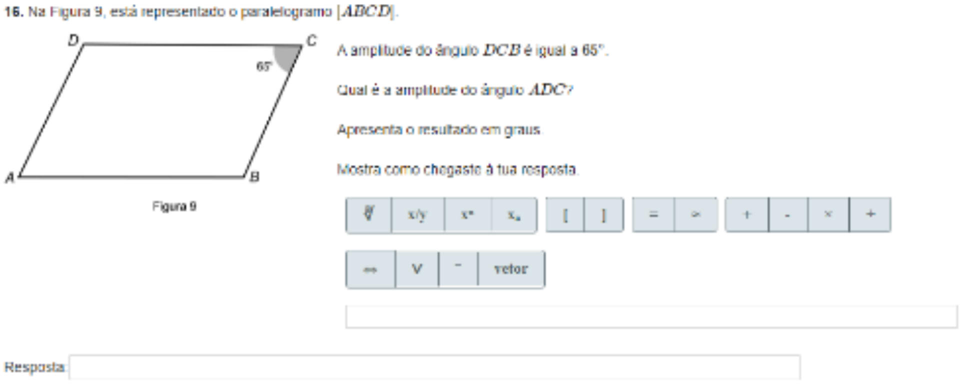
Wang, T.-H., Kao, C.-H., & Chen., H.-C. (2021). Factors Associated with the Equivalence of the Scores of Computer-Based Test and Paper-and-Pencil Test: Presentation Type, Item Difficulty and Administration Order. *Sustainability, 13*(17), 9548.

Wang, Y. Y., Jiang, Y., Li, Q., & Lie, Y. (2022). Innovative online learning strategies for the successful construction of student self-awareness during the COVID-19 pandemic: Merging TAM with TPB. *Journal of Innovation & Knowledge, 7*(4), 100252.

Wuthisatian, R. (2020). Student exam performance in different proctored environments: Evidence from an online economics course. *International Review of Economics Education, 35*, 100196.

Appendix – Figures and tables

Figure A1 - Example of a mathematical question on CBT mode



16. Na Figura 9, está representado o paralelogramo $[ABCD]$.

A amplitude do ângulo DCB é igual a 65°.

Qual é a amplitude do ângulo ADC?

Apresenta o resultado em graus.

Mostra como chegaste à tua resposta.

Resposta:

Figure A2 - Distribution of CBT exams in mainland Portugal

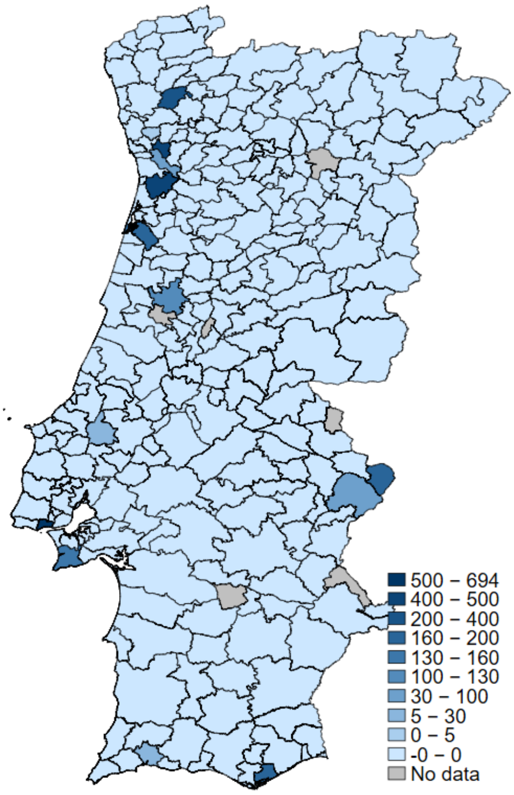


Figure A3 - Distribution of schools with CBT exams in mainland Portugal

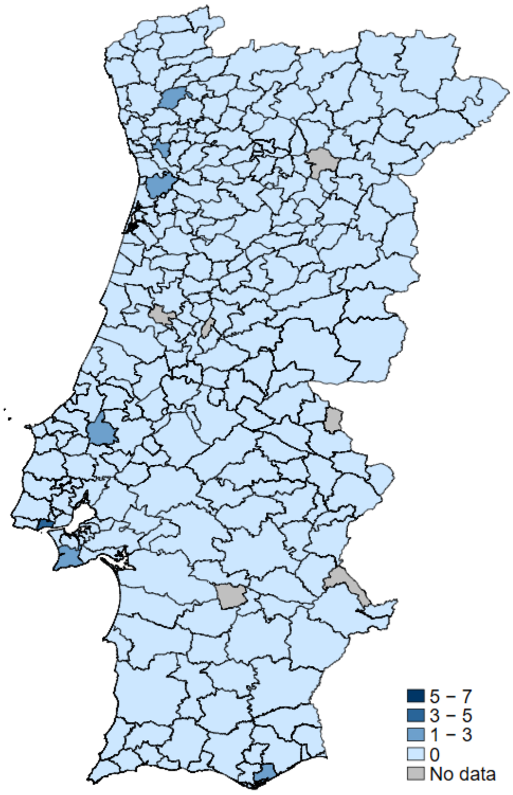


Figure A 4 - Distribution of CBT exams in Azores (Island of Portugal)

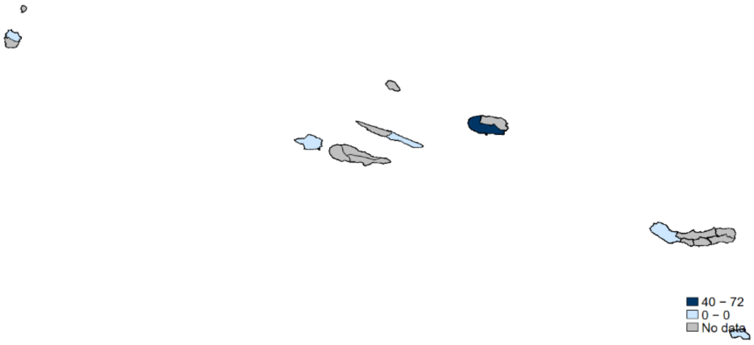


Figure A5 - Distribution of schools with CBT exams in Azores (Island of Portugal)

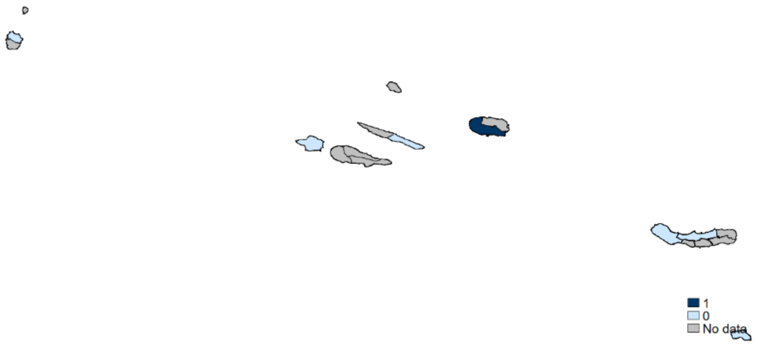


Figure A6 - Distribution of schools with CBT exams in Madeira (Island of Portugal)

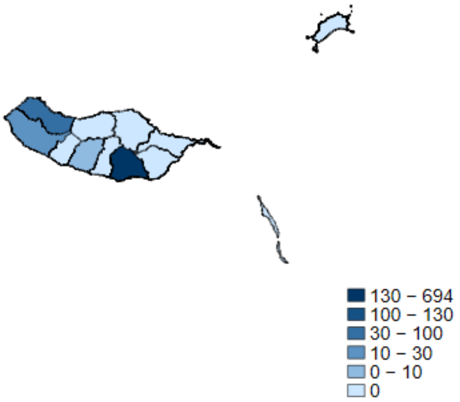


Figure A6 - Distribution of schools with CBT exams in Madeira (Island of Portugal)

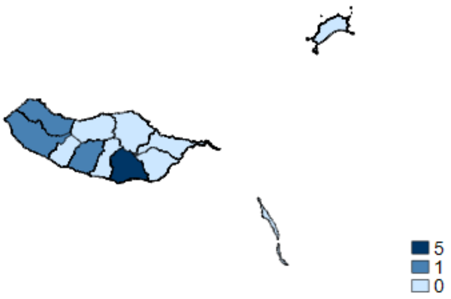
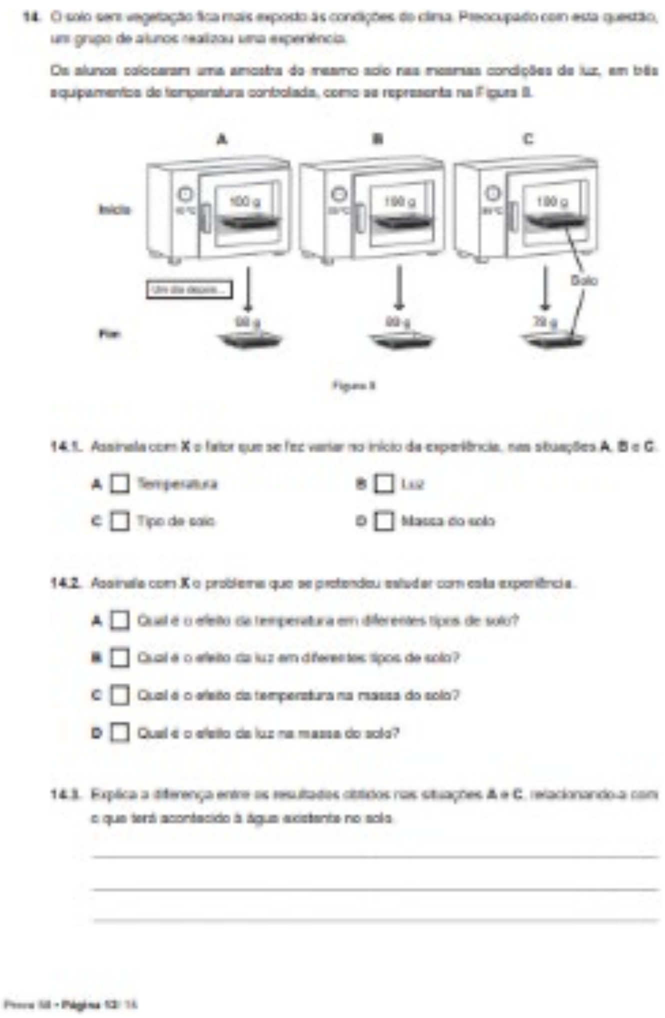


Figure A7 - Example of PPT item (left hand-side) and CBT item (right hand-side)



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.

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?

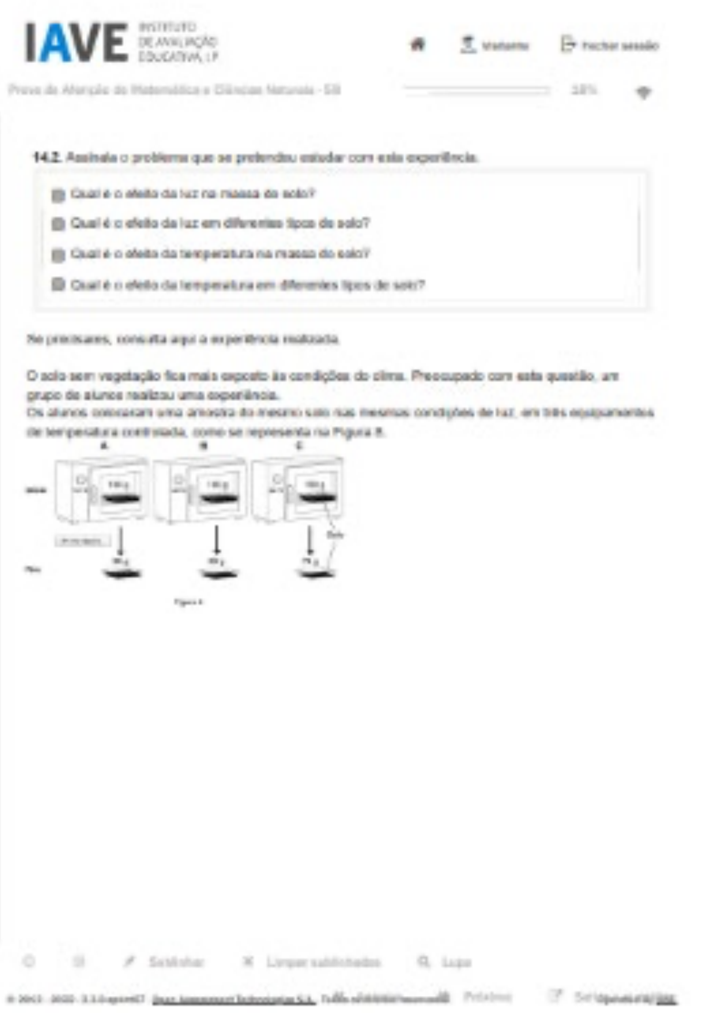
14.3. Explica a diferença entre os resultados obtidos nas situações **A** e **C**, relacionando-a com o que terá acontecido à água existente no solo.

____

____

____

Prova 58 • **Página 12**/ 16



IAVE INSTITUTO DE AVALIAÇÃO EDUCATIVA, I.P.
Visitante
Fechar sessão
Prova de Aferição de Matemática e Ciências Naturais - 58
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.

© 2013 - 2022 - 3.3.0-[illegible] [illegible] Assessment Technologies S.A. Todos os direitos reservados

43

Table A1 - Descriptive Statistics

	Mean	**SD**	**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
School abroad	0.004	0.06	0	1
	N= 436,947			

Note: This table presents means, standard deviations, minimum and maximum values of the variables used in the paper, for the full sample student-exam observations

Figure A8 - Distribution of scores

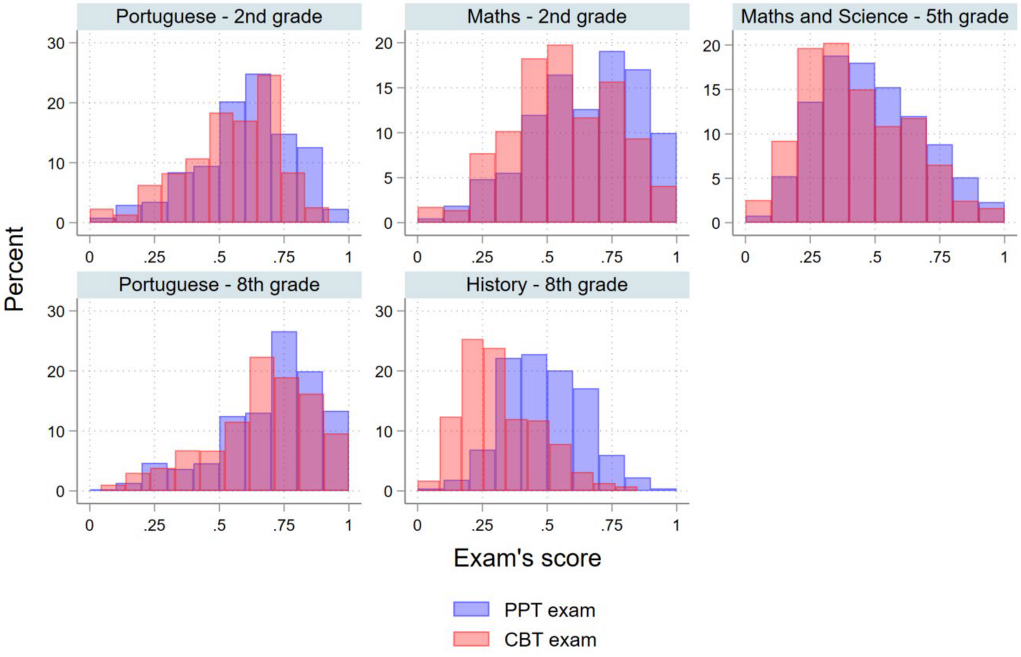


Figure A 9 - Number of students per treatment status

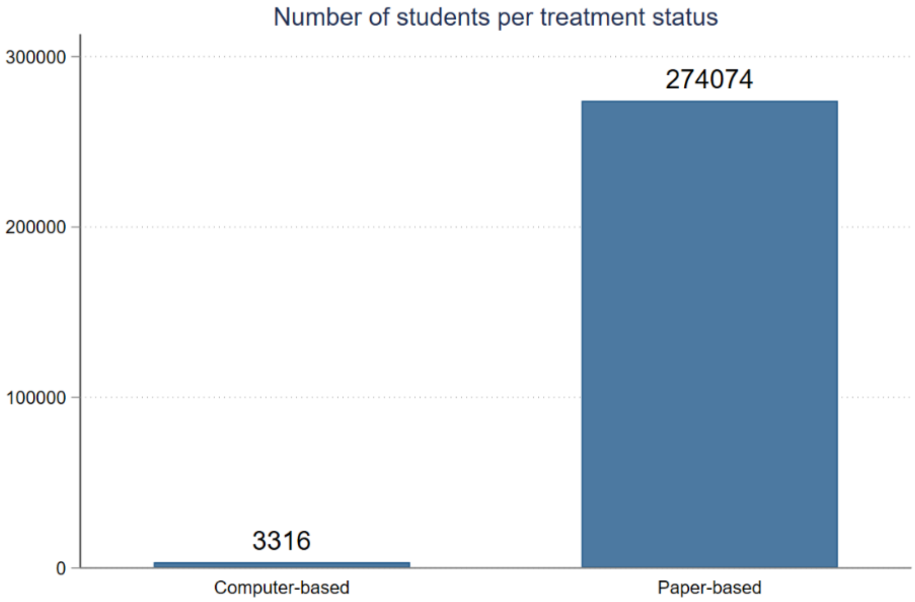


Figure A10 - Number of students per school category

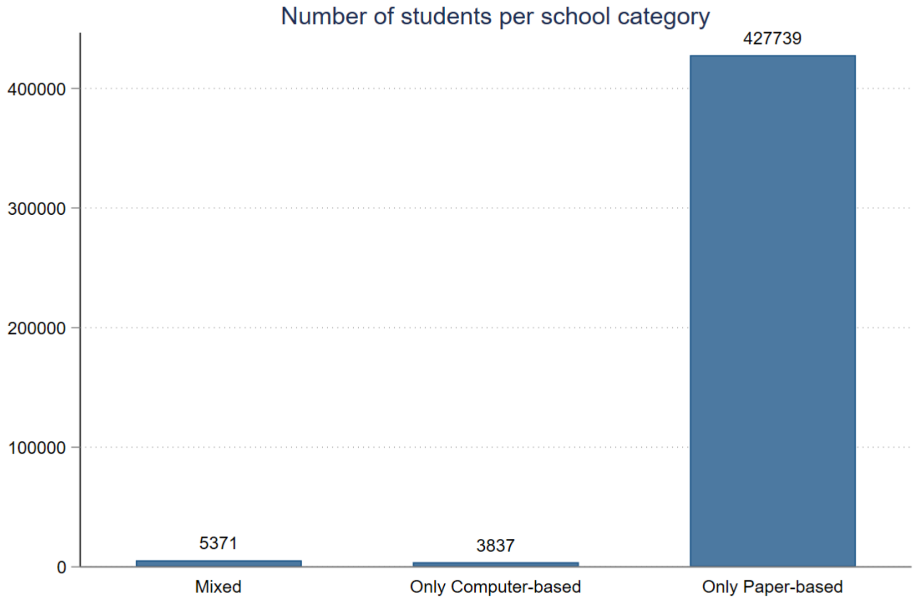


Figure A11 - Number of schools per school category

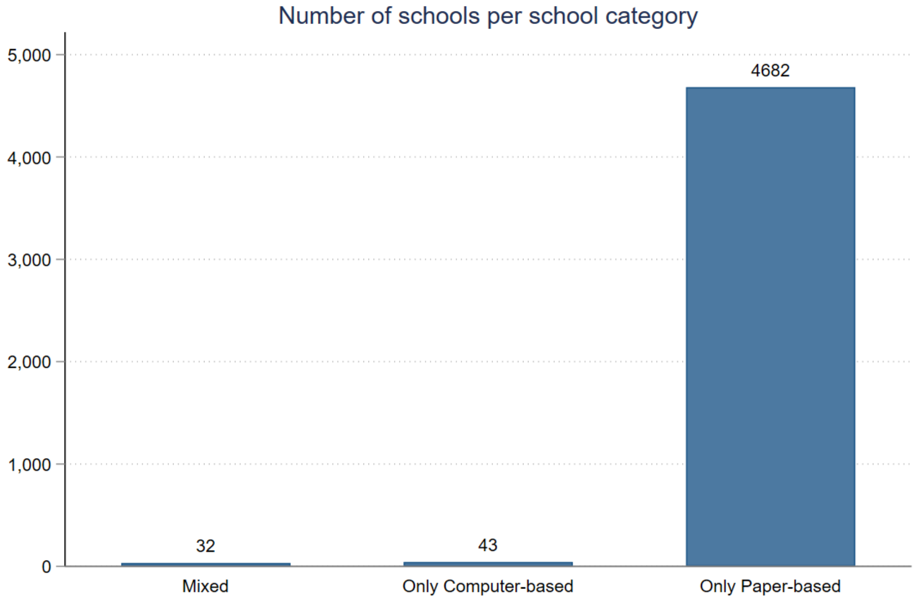


Table A1 - Balance test, full sample only considering mainland Portugal

	Control	Treatment	Difference	P-value
Age (years)	11.341 (0.047)	12.183 (0.39)	0.841	0.033**
Gender (1 Male, 0 Female)	0.514 (0.001)	0.503 (0.012)	-0.011	0.358
Type of School (1 Public, 0 Private)	0.886 (0.008)	0.883 (0.056)	-0.002	0.967
Special needs	0.012 (0.001)	0.015 (0.01)	0.003	0.756
ASE level 1	0.144 (0.002)	0.123 (0.012)	-0.021	0.096*
ASE level 2	0.139 (0.002)	0.14 (0.017)	0.001	0.956
ASE level 3	0.015 (0.002)	0.036 (0.021)	0.022	0.299
Lisbon	0.248 (0.011)	0.222 (0.092)	-0.026	0.784
North	0.313 (0.011)	0.333 (0.113)	0.020	0.859
Centre	0.212 (0.009)	0.192 (0.095)	-0.020	0.835
South	0.228 (0.01)	0.253 (0.083)	0.025	0.759
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 A2 - Balance test for full sample, including parents' education

	Control	Treatment	Difference	P-value
Age (years)	11.417 (0.049)	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)	-0.002	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 (0.001)	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	0.33 (0.012)	0.238 (0.11)	-0.092	0.406
Centre	0.222 (0.01)	0.215 (0.105)	-0.006	0.953
South	0.213 (0.01)	0.179 (0.077)	-0.033	0.670
Portuguese Islands	0.028 (0.004)	0.103 (0.042)	0.075	0.069*
School abroad	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. 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 A4 - Balance test for the mixed schools' sample and the full sample

	Mixed schools	Full sample	Difference	P-value
Age (years)	11.339 (0.045)	12.228 (0.213)	0.890	0.000***
Gender (1 Male, 0 Female)	0.513 (0.001)	0.52 (0.012)	0.006	0.576
Type of School (1 Public, 0 Private)	0.879 (0.008)	0.757 (0.098)	-0.122	0.217
Special needs	0.012 (0.001)	0.007 (0.003)	-0.005	0.094*
ASE level 1	0.141 (0.002)	0.131 (0.019)	-0.010	0.601
ASE level 2	0.138 (0.002)	0.148 (0.02)	0.010	0.610
ASE level 3	0.023 (0.002)	0.1 (0.03)	0.077	0.011**
Lisbon	0.236 (0.01)	0.000	-0.236	0.000***
North	0.298 (0.011)	0.093 (0.07)	-0.204	0.004***
Centre	0.201 (0.009)	0.077 (0.073)	-0.124	0.094*
South	0.214 (0.009)	0.282 (0.105)	0.068	0.523
Portuguese Islands	0.043 (0.005)	0.402 (0.108)	0.359	0.001***
School abroad	0.008 (0.003)	0.145 (0.077)	0.138	0.073*
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 A5 - Balance test for mixed schools

	Control	Treatment	Difference	P-value
Age (years)	12.228	12.228		0.999
	(0.415)	(0.449)	-0.001	
Gender (1 Male, 0 Female)	0.519	0.521		0.935
	(0.015)	(0.015)	0.002	
Type of School (1 Public, 0 Private)	0.69	0.842		0.155
	(0.13)	(0.074)	0.152	
Special needs	0.008	0.005		0.499
	(0.004)	(0.003)	-0.003	
ASE level 1	0.139	0.121		0.331
	(0.024)	(0.018)	-0.018	
ASE level 2	0.136	0.162		0.272
	(0.024)	(0.021)	0.026	
ASE level 3	0.115	0.082		0.402
	(0.037)	(0.034)	-0.033	
Lisbon	-	_	-	-
North	0.001	0.213		0.143
	(0.001)	(0.142)	0.212	
Centre	0.051	0.111		0.293
	(0.051)	(0.104)	0.060	
South	0.324	0.227		0.312
	(0.124)	(0.105)	-0.096	
Portuguese Islands	0.427	0.37		0.621
	(0.13)	(0.113)	-0.057	
School abroad	0.197	0.078		0.131
	(0.106)	(0.045)	-0.119	
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. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively.

Table A6 - Balance tests for DiD with PPT students as control group

	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 (0.01)	0.804 (0.173)	-0.088	0.610
Special needs	-	-	-	-
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	-	-	-	
North	-	-	-	
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*
School abroad	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. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively.

Table A7 - Balance test for DiD with CBT students as control group

	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
Special needs	-	-	-	
ASE level 1	0.103 (0.0134)	0.154 (0.056)	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 (0.03)	0.045	0.211
Lisbon	-	-	-	
North	-	-	-	
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
School abroad	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. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively.

Table A11 - Estimated CBT effect, full sample with robust standard errors

	(1)	(2)	(3)	(4)	(5)
	Score	Score	Score	Score	Score
CBT (δ)	-0.087***	-0.075***	-0.075***	-0.088***	-0.062***
	(0.0025)	(0.0025)	(0.0025)	(0.0027)	(0.0055)
Observations	436,947	436,947	436,947	436,947	436,943
R-squared	0.174	0.260	0.263	0.273	0.358
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 have special education needs, social support from the government, type of school (public or private) and location.

Table A12 - Estimated CBT effect, full sample only including mainland Portugal

	(1)	(2)	(3)	(4)	(5)
	Score	Score	Score	Score	Score
CBT (δ)	-0.078***	-0.080***	-0.081***	-0.092***	-0.065***
	(0.0114)	(0.0119)	(0.0121)	(0.0104)	(0.0174)
Observations	411,891	411,891	411,891	411,891	411,887
R-squared	0.173	0.259	0.261	0.271	0.355
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, whether students have special education needs, social support from the government, type of school (public or private) and location.

Table A13 - Estimated CBT effect, full sample, including parents' education

	(1)	(2)	(3)	(4)	(5)
	Score	Score	Score	Score	Score
CBT (δ)	-0.094***	-0.088***	-0.090***	-0.095***	-0.051**
	(0.0127)	(0.0097)	(0.0101)	(0.0102)	(0.0237)
Observations	277,390	277,390	277,390	277,390	277,373
R-squared	0.184	0.305	0.308	0.317	0.389
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. 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 have special education needs, social support from the government, type of school (public or private) and location.

Table A14 - Estimated CBT effect, mixed schools' sample, including parents' education

	(1)	(2)	(3)
	Score	Score	Score
CBT (δ)	-0.042	-0.048**	-0.044**
	(0.0249)	(0.0198)	(0.0196)
Observations	2,090	2,090	2,090
R-squared	0.266	0.386	0.395
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. . 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 have special education needs, social support from the government, type of school (public or private) and location.

Table A15- Estimated CBT effect on the History exam, DiD with PPT students as control group, including parents' education

	(1) Score	(2) Score
T (δ)	-0.009	-0.007
	(0.0117)	(0.0087)
Subject (μ)	-0.208***	-0.208***
	(0.0014)	(0.0014)
*T * Subject* (θ)	-0.195***	-0.195***
	(0.0146)	(0.0146)
Observations	114,936	114,936
R-squared	0.268	0.382
Controls	NO	YES

Note: This table presents estimates of the effects of CBT mode effect on the History 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 (*T*=1) and students that did both exams on paper (*T*=0). *Subject=1* for the Portuguese exam and *Subject=0* for the History one. and 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 have special education needs, social support from the government, type of school (public or private) and location.

Table A16 - Estimated PPT effect on the Portuguese exam, DiD with CBT students as control group, including parents' education

	(1) Score	(2) Score
T (δ)	-0.042***	-0.003
	(0.0132)	(0.0132)
Subject (μ)	0.345***	0.345***
	(0.0128)	(0.0128)
*T * Subject* (θ)	0.058**	0.058**
	(0.0203)	(0.0204)
Observations	1,804	1,804
R-squared	0.497	0.572
Controls	NO	YES

Note: This table presents estimates of the effects of PPT mode effect on the Portuguese 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 (*T*=1) and students that did both exams on paper (*T*=0). *Subject=1* for the Portuguese exam and *Subject=0* for the History one. The samples only include students in mainland Portugal. 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 have special education needs, social support from the government, type of school (public or private) and location.

Table A17 - Estimated CBT effect, without including the History exam in 8th grade

	(1) Score	(2) Score	(3) Score	(4) Score	(5) Score
CBT (δ)	-0.053***	-0.039***	-0.039***	-0.053***	-0.027***
	(0.0030)	(0.0030)	(0.0030)	(0.0032)	(0.0068)
Observations	344,722	344,722	344,722	344,722	344,718
R-squared	0.144	0.227	0.238	0.250	0.347
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. Significance at the 1, 5 and 10 percent levels is indicated by ***, ** and *, respectively. Controls include age, gender, parents' education, if students have special education needs, social support from the government, type of school (public or private) and location.