dECON
Facultad de Ciencias Sociales
UNIVERSIDAD DE LA REPÚBLICA

Documentos de Trabajo

Virtual Instruction effects within University Courses

A Boon for Those Who Need it, a Bane for Others.

Flores, Manuel - Gerstenblüth, Mariana - Suarez, Lucía - Cantera, Luciana.

Documento No. 01/24
Febrero 2024

ISSN 0797-7484

Noname manuscript No.
(will be inserted by the editor)

Virtual Instruction effects within University Courses

A Boon for Those Who Need it, a Bane for Others

Flores, Manuel* · **Gerstenblüth, Mariana*** · **Suarez, Lucía*** · **Cantera, Luciana***

February, 2024

Abstract We examine the effects of virtual instruction on academic achievement at the *Universidad de la República*, Uruguay, in 2022. We analyze student performance by considering the sequential nature of the evaluation process within the courses. Our results reveal that students in virtual courses are less likely to be active or achieve course approval. When possible, we use alternative identification strategies that show the stability of the estimated effects. We also find that the gap in the results is explained by the sequence of intermediate tests, which combine different performances in terms of retention and test scores. We highlight the importance of effective targeting as the negative effects disappear for students facing constraints on attendance.

Keywords Virtual education · Student performance · Sequential Treatment Effects · Heterogeneous Treatment Effects · Program Targeting

JEL Classification C21 · H75 · I20 · I21 · I23

* Departamento de Economía, Facultad de Ciencias Sociales, Universidad de la República, Constituyente 1502, piso 6, 11200, Montevideo, Uruguay. Corresponding author's email: manuel.flores@cienciassociales.edu.uy.

1 Introduction

This paper evaluates the impact of virtual instruction on students' performance in university courses based on the Online Learning Pilot Experience (OLPE) conducted in 2022 in the Social Sciences Faculty (FCS) at *Universidad de la República* (Udelar) in Uruguay. The COVID-19 pandemic led to a massive shift to distance learning methods. Currently, the question of how to combine traditional instruction with virtual teaching appears as an important educational policy issue that must be addressed.

Thus far, the evaluation of educational delivery methods has yielded mixed results. A series of experimental designs at the tertiary level in economics have found that virtual instruction leads to worse academic outcomes than face-to-face instruction (Alpert et al., 2016; Escueta et al., 2017; Figlio et al., 2013; Jaggars and Xu, 2016; Xu and Jaggars, 2014). However, blended courses have not yet been found to significantly underperform purely face-to-face (Alpert et al., 2016; Bettinger et al., 2017; Bowen et al., 2014; Escueta et al., 2017; Joyce et al., 2015). In addition, Alpert et al. (2016) found that the increased dropout of virtual courses amplifies the potential negative impact of virtual coursework compared with face-to-face groups.

More recently, some studies have analyzed the effects of online teaching on educational outcomes during the pandemic. In line with previous findings, virtual students in higher education perform worse than their face-to-face counterparts (Altindag et al., 2021; Foo et al., 2021) and have lower completion rates (Bird et al., 2022; Bulman and Fairlie, 2022). Similar findings are attained for primary and secondary schools, where Jack et al. (2023) find that declines in student pass rates are larger in districts with less in-person schooling. Cacault et al. (2021) note that the effects of virtual instruction are heterogeneous, having a negative impact on low-ability students but improving results on high-ability fellows.

However, the previous studies mostly focused on developed countries, and research for tertiary education in Uruguay is limited. This paper contributes to the existing literature on assessing student outcomes in various teaching modalities, tracking students throughout the course, rather than simply looking at the final outcomes. Our approach provides a more nuanced understanding of the impact of different teaching methods on student performance.

Our study focuses on a post-pandemic program in which all students, regardless of the course modality, underwent the same assessments throughout the course. Furthermore, analyzed courses included a diverse range of disciplines and belong to a public university. These factors make our evaluation an original contribution to the ongoing discussion of the effectiveness of virtual teaching methods.

Regarding final outcomes, the results showed that students performed worse in virtual environments than in face-to-face. However, differences in performance are not always significant when looking at intermediate results. In addition, the Heterogeneous Treatment Effects (HTE) results show that the negative effects are concentrated in students who were not targeted by the program. The negative effects lose significance in the sub-sample of targeted students, and in some cases they even become positive.

The remainder of this paper is organized as follows. Section 2 presents the OLPE program and the context in which it was implemented. Section 3 describes data. The evaluation strategy is described in Section 4. The main empirical results are presented in Section 5. Alternative identification strategies are shown in Section 6 and heterogeneity of treatment effects analyzed in Section 7. Section 8 discusses our main findings, and Section 9 concludes.

2 Program and Context

The OLPE program took place at FCS-Udelar, the main public university in Uruguay. Udelar offers almost 100 careers and has approximately 145,000 enrolled students, accounting for 85% of all undergraduates in the country. FCS is one of its 15 faculties, which has approximately 7,000 students and offers four careers: Social Work, Political Science, Sociology, and Development.

The program aimed to accommodate students who faced geographic, employment, or care-giving challenges. However, it was open to all students, with no priority given to the target group. Around 40% of first-year students are full- or part-time workers and the same proportion have care-giving responsibilities, leading to a high first-year dropout rate (around 30%), which is comparable to other Latin American universities.

Uruguay has good infrastructure and Internet access, with over 90% of the population having cell phones (D'Almeida and Margot, 2018). Additionally, 88% of households have Internet access and 71% have broadband connectivity. The high penetration of Internet connectivity across all income quintiles, from 84% in the lowest to 95% in the highest, highlights the favorable conditions for online learning in the country (AGESIC, 2020).

Since 2008, the university has been using a Moodle platform named EVA to distribute study materials and facilitate communication between students and instructors. From the pandemic on, EVA was also used for evaluations. The university provides free Zoom licenses to all faculty members and students.

We evaluate the four required courses of the first semester: Problems of Development (PD), Principles of Economics (PE), Mathematics (MAT) and The Social Question in History (SQH).[1]

The evaluated courses have some shared characteristics, including concurrent start dates and duration (15 weeks), a shared pool of students, and 12-14 groups each. All courses in FCS use a common scoring system, where students must earn a minimum of 81 out of 100 points in intermediate tests to receive Full Approval (FA). A score between 50 and 80 points indicates Partial Approval (PA), which requires students to take a supplementary exam. No minimum score is required in any intermediate test. Each course designed its pedagogical proposal for the virtual groups, leading to different combinations of synchronous and asynchronous instances. Assessment strategies were also course-specific, but tests were always the same for virtual and face-to-face students, both in terms of test papers and modalities (details on test characteristics are provided in Appendix A).

Students could apply to a face-to-face or virtual group. Table 1 shows the number of students and groups in each course and modality. Three of the courses offered three or four virtual

[1] The four courses have different curricular objectives. PD aims to enhance students' comprehension of the intricacies of development using a combination of theoretical and practical approaches. PE covers the concepts and analytic tools necessary to understand the macro-and microeconomic aspects of social reality. The objective of MAT is to equip students with a robust mathematical background. Finally, SQH brings students closer to the "social question" as a field of study within the social sciences, encompassing various theoretical and ideo-political perspectives.

groups (PE, PD and MAT). In the case of SQH there was only one virtual group but four blended groups were offered.[2]

Table 1: Number of groups and students by course and modality

	PD			PE			MAT			SQH			
	Tot	F2F	Virt	Tot	F2F	Virt	Tot	F2F	Virt	Tot	F2F	Blen	Virt
# Groups	14	11	3	13	9	4	12	9	3	13	8	4	1
# Students	819	570	249	974	635	339	846	578	268	806	575	152	79

Notes: Own elaboration using DECID 2022. "Tot": total; "F2F": face-to-face; "Virt": virtual; "Blen": blended.

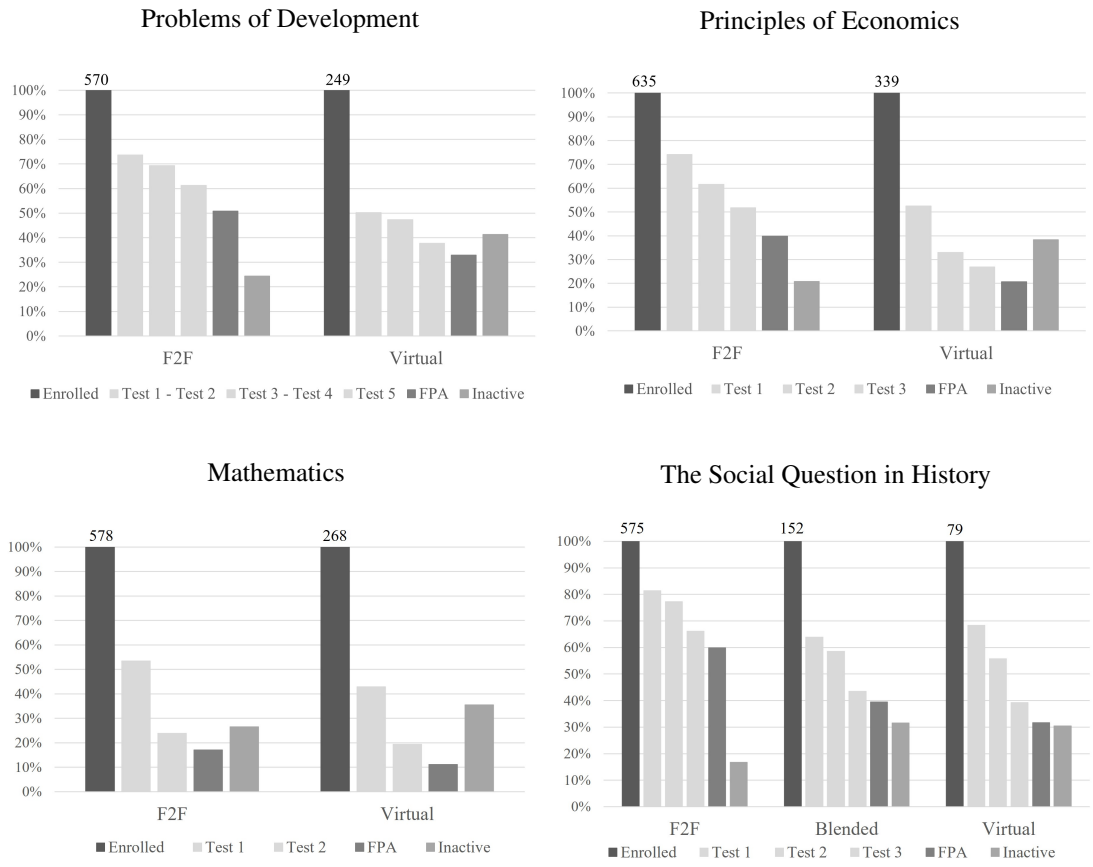
Figure 1 shows the retention rates for each test by modality and course. This reveals that virtual learning environments have a higher proportion of students with no activity and a declining participation rate in evaluation instances. Furthermore, the percentage of students who successfully completed the course is lower in the virtual groups.

Table A.1 presents the average grades for each evaluation instance for each course and modality. In general, there are no significant differences in average grades between face-to-face and virtual learning modalities. Descriptive statistics give no evidence that one modality consistently outperformed the others.

However, these differences in retention or average grades could be due to differences in the characteristics of the students who self-select into virtual groups. In particular, these students have distinct characteristics such as a higher average age, a larger proportion from public secondary education, full-time employment, care-giving responsibilities, and admission to faculty prior to 2022 (see Appendix B for more details).

[2] The difference between the blended and virtual groups is that the blended group had one face-to-face class per week.

Fig. 1: Number of students and performance by course and treatment group



Notes: Own elaboration using DECID 2022.

Table 2: Maximum and average scores by course, modality and test

	PD			PE			MAT			SQH			
Test	Max	F2F	Virt	Max	F2F	Virt	Max	F2F	Virt	Max	F2F	Blen	Virt
1	5	3.4	3.1	15	10.7	10.3	35	12.6	11.8	20	16.7	16.8	16.8
2	20	15.8	15.9	50	24.6	22.2	50	28.2	23.1	35	27.5	26.4	27.4
3	5	3.5	3.4	25	17.3	17.4				45	29.6	31.6	29.9
4	20	16.0	16.1										
5	50	26.7	29.1										

Notes: Own elaboration using DECID 2022. “Max”: maximum; “F2F”: face-to-face; “Virt”: virtual; “Blen”: blended.

3 Data sources

Combining academic results and socioeconomic data, we created the Database on First-Year Students and Their Performance (DECID, for its name in Spanish). It includes information for 1651 students who were enrolled in at least one of the evaluated courses.

Information was collected from several sources. In particular, the official records of the University's IT Central Service (SECIU, for its name in Spanish) provided data on student enrollment and performance. Mid-term results and initial diagnostic tests were obtained from course coordinators. By combining these sources, a detailed record of each student's performance and final course results was obtained.

Two sources were combined to obtain socioeconomic information. On the one hand, the university requests all students to fill a compulsory statistical questionnaire (available for years 2018 to 2021). Missing observations were completed using an additional statistical questionnaire requested by the FCS to each new generation (available for generations 2022 and 2021). The relevant questions were compatible between these two sources. The resulting database includes socioeconomic information on age, gender, residence, working condition (full- or part-time), and care-giving tasks. It also includes educational information such as the year and department in which secondary school was finished, the year in which the degree was started, the number of courses to which the student was enrolled, and a dummy variable for repeat students.

4 Methods

Our main strategy for the identification of treatment effects was based on controlling selection on observables through matching estimators, which allowed us to measure the effects of the OLPE program in all four courses.

As shown in the previous section, there were significant number of non-attendees, and the dropout rates were abnormally high compared to universities in developed countries. These factors made it difficult to accurately identify effective participants in the program. Nevertheless, the list of selected students was clearly defined, which provided an indicator variable of intention to treat (ITT).[3]

Treatment effects were evaluated separately for each course because, as illustrated in the previous section, virtual delivery techniques varied among courses. In every case we estimated average treatment effects (ATE) for retention rates, test scores and final results.

[3] While access to virtual classes in most courses was limited to those students assigned to virtual groups, this was not the case for MAT, where virtual classes were available to students in any group, and movements across groups were permitted.

4.1 Effects on Final Results

Self-selection is the main identification challenge because the treatment variable is no longer independent of potential outcomes. The information only allows the observation of the average outcome under treatment for the treated sample, $E[Y_i^T|T_i = 1]$, and the average outcome with no treatment for the control sample, $E[Y_i^C|T_i = 0]$. However, the evaluation of the treatment effects requires two unobserved counterfactuals: the average outcome under treatment for the untreated sample, $E[Y_i^T|T_i = 0]$, and the average outcome with no treatment for those that were treated $E[Y_i^C|T_i = 1]$.

As shown by Barnow et al. (1980) and Rosenbaum and Rubin (1983, 1985), if selection can be assumed to rely exclusively on observable variables X_i (and provided that a common support assumption is also met), it is possible to build a control group that is similar to the treated group in terms of X. Matrix X must gather all confounding variables (i.e., those that affect the probability of being treated and the outcomes). In our model the X matrix includes variables that control for factors that explain a greater need to use the virtual modality, such as a dummy variable that indicates whether the person resides in Montevideo, where the FCS is located, dummies indicating if the student works part- or full-time, and a dummy that indicates whether the person has care-giving responsibilities. Other controls that may be associated with preferences for virtuality are also included, such as a dummy variable for sex, dummy variables for five age groups, dummies indicating if secondary education was completed in Montevideo and if it was in a private institution, four dummies indicating generations 2019 to 2022, a continuous variable counting the number of other virtual courses in which she is enrolled, and a dummy variable indicating whether the student is repeating the course.

Matching estimators rely on two critical assumptions. The Conditional Independence Assumption (CIA) requires that the vector of potential outcomes is independent of the treatment once conditioned to the observed values of the variables in X, that is, $\left(Y_i^T, Y_i^C\right) \perp T_i|X_i$. The Common Support Assumption requires that in each combination of values in X both treated and control observations can be found, that is, $0 < p_i < 1$ with $p_i = p(X_i) = Pr(T_i = 1|X_i)$. In our particular setting, the CIA requires that, after conditioning on observables, the assignment to a virtual group is independent of potential outcomes.

Another important assumption is the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1974, 1980). In our study, SUTVA requires that no mechanism exists through which students in face-to-face groups are affected by the fact that other students are receiving virtual lessons. However, this assumption is potentially problematic since online content might be accessed by all enrolled students, although this has not been a common practice. Therefore, we define our control group as those students who were not enrolled in virtual or blended groups but who had potentially accessed online materials intended for use in virtual groups.

Our estimation method relies on Propensity Scores (PS), i.e. the probability of being treated according to observable characteristics. This means that each student in a virtual group is compared with one or more face-to-face students selected according to their proximity in terms of PS, p_i. This allows estimation of the ATE conditional on the PS for student i as proposed in equation 1 (subscripts are omitted for simplicity).

$$E\left[Y|T=1,p(X)\right]-E\left[Y|T=0,p(X)\right]=E\left[Y^T-Y^C|p(X)\right] \tag{1}$$

Integrating the last expression over $p(X)$ gives the $\Delta^{ATE}=E\left[Y^T-Y^C\right]$.

Matching estimators also need to fulfill a Common Support condition, meaning that the distributions of the PS for the treatment and control groups overlap. The results shown in the following section maintain observations pertaining to the region of common support, dropping observations that fell outside the overlap region.[4]

We used the full sample of students to evaluate the final results, comparing the treatment and control groups in terms of three outcomes. The first is the probability of being active, a category that signals students taking any of the tests during the course. Thus, the differences in the probability of being active can hardly be interpreted as a consequence of the treatment itself. Instead, it describes the potential differences in the engagement of enrolled students.[5] The other two outcomes studied were the probabilities of "Full or Partial Approval" (FPA) and of "Full Approval" (FA).

[4] The balance property is fulfilled in every observable characteristic of the vector X except age. Treated individuals tend to be older than those in the control group.

[5] Inactive students are those who enroll but never start a course, or those who assist to the first lectures and promptly abandon, either in virtual or face-to-face groups.

4.2 Sequential Treatment Effects

A distinctive feature of our setting is the possibility of measuring differences between the two groups in each intermediate test. An additional challenge emerges, because dropouts occur at every intermediate stage, meaning that the treated and control samples change during the semester. In addition, an outcome in one stage may influence another outcome in the following stages. More specifically, a low/high score in one stage can reduce/increase the probability of being retained in the following stages.

To deal with this complex relationship between test scores and retention we identify the stages s in each course's progression that coincide with the moment of the tests, where $s \in [1, 2, 3]$.[6] In every stage we analyze these two outcomes, which are now denoted as $Y_{is} = \{R_{is}, S_{is}\}$ where R_{is} is a dummy variable signaling retention of student i in stage s and S_{is} a continuous variable gathering the score obtained by student i in stage s.

In contrast to an important strand in the literature on Dynamic Treatment Effects (Okamura and Islam, 2021; Ding and Lehrer, 2010), students' affiliation to the treatment or control group is a fixed attribute (T_i) that does not change during the different stages of the course. We also assume that the confounding factors (X_i) do not vary in time.[7]

Estimation of treatment effects in this context requires taking into account the fact that dropouts have an impact on average exam scores. As the course progresses, students who are still enrolled are selected from among those who have a higher chance of approval and higher expected scores. To address this issue, in each stage, we re-matched treatment and control observations among the students who were still present in that stage.

On the other hand, the score that the student obtained in the previous stages may be relevant in explaining retention in stages two and three. Thus, intermediate scores are confounders that need to be controlled when analyzing the effects of treatment on the probability of retention.

[6] In PD tests 1 and 2 are subsumed in stage $s = 1$, tests 3 and 4 correspond to $s = 2$, because the two tests fall within a week of each other.

[7] Some variables can vary in time but we have only information at the initial stage, these are residence department, or the condition of work or care-giving. This can be seen as an important limitation, because any effect of the treatment that is channeled through changes in residence, work or caring conditions will be ignored.

As discussed in Lechner (2004) this particular issue modifies the assumptions required for identification, and the Dynamic Conditional Independence Assumptions should be met. This means that the vector of potential outcomes in period $s > 1$ must be orthogonal to the treatment conditional on the variables in X and to the score(s) in the previous period(s). In our approach, sequential matching consists of including past scores in the conditioning set when evaluating the effects on retention.

4.3 Estimation Methods

A main characteristic of our setting is that the treatment status is defined at the onset and is stable thereafter, while the outcomes appear as a sequence of retention and scores in stages two to three. Two aspects of this process must be considered in the estimation strategy. First, the retention outcome in advanced stages depends on the history of scores in previous stages (because students with low cumulative scores are expected to withdraw, as they predict that they will fail).[8] Second, dropouts can be seen as selective attrition that increases expected scores as stages go by.[9]

To address the first aspect, we need to separately estimate an equation for the retention outcome in each stage, where the probability of being retained depends on the scores accumulated up to each stage. The second aspect is faced through a re-estimation of the propensity scores using a sample of students retained up to the corresponding stage. Inverse Probability Weighted Regression Adjustment (IPWRA) estimators allow separately estimating an equation for the outcome and an equation for the probability of being treated in each stage (Robins et al., 2000). An advantage of IPWRA estimators is that they are "doubly-robust", since as was shown by Wooldridge (2007) it only requires correct specification of either the treatment or the outcome equation for the estimators to be consistent.

[8] Scores in one stage should not explain scores in subsequent stages, as contents evaluated are different across tests.

[9] An alternative approach to face sample selection issues could be to estimate upper and lower bounds for the true effects, based on Lee (2009). However, Lee bounds require a monotonicity assumption, meaning that assignment to the treatment group only is allowed to affect sample selection into treatment in one direction, whether increasing or decreasing the probability of remaining in the treatment group. In our case the treatment could increase or decrease the probability of dropping out, and thus the monotonicity assumption would be violated.

In each stage s the sample is composed of those retained in the previous stage ($R_{ij} = 1 \; \forall \; j < s$). We obtain a stage-specific propensity score as the predicted value $\widehat{p}_{is}$ from a logistic treatment regression (equation 2).

$$T_i = \Lambda(X_i\alpha) + e_{is}, \tag{2}$$

where $\Lambda(\cdot)$ is the cumulative logistic function, α is the vector of the corresponding coefficients, and e_{is} is an idiosyncratic error term. Then, a Regression Adjustment (RA) strategy is performed using $1/\widehat{p}_{is}$ as observation weights, and equations 3 and 4 are separately estimated using linear models.

$$S_{is} = \tau T_i + Z_i\beta + u_{is} \tag{3}$$

$$R_{is} = \theta T_i + W_i\gamma + \delta \sum_{j=0}^{s-1} S_{ij} + v_{is} \tag{4}$$

where W and Z are matrices gathering control variables for scores and retention, β and γ are coefficient vectors, δ is a scalar coefficient, u_{is} and v_{is} are idiosyncratic error terms, and τ and θ are the treatment effect estimators for scores and retention, respectively.[10]

In all cases, standard errors of the estimators were obtained from bootstrapping using 400 repetitions.

5 Results

The final outcomes were evaluated by considering differences in activity status, FPA, and FA, between virtual or blended groups and the control group of students who attended the course in person. Table 3 presents the IPWRA estimation results.[11] When significant, all coefficients are negative, showing that in terms of activity and approval expected probabilities, virtual students perform worse.

[10] Matrices Z and W contain the same variables in X, except for the dummy signaling students that live in Montevideo, as the place of residence is not expected to affect academic performance.

[11] Weights are propensity scores that are obtained from a probit estimation presented in Appendix C.

Table 3: Treatment effects on final outcomes: IPWRA

	Active	FPA	FA
Problems of Development			
ATE	-0.09*	-0.14***	0.01
	(0.05)	(0.05)	(0.03)
N	819	819	819
Principles of Economics			
ATE	-0.09**	-0.16***	-0.05**
	(0.04)	(0.04)	(0.02)
N	974	974	974
Mathematics			
ATE	-0.01	-0.07***	-0.04***
	(0.04)	(0.02)	(0.01)
N	846	846	846
Social Question in History Virtual			
ATE	-0.48***	-0.27***	-0.04
	(0.10)	(0.09)	(0.08)
N	654	654	654
Social Question in History Blended			
ATE	-0.12***	-0.16***	-0.04
	(0.04)	(0.05)	(0.04)
N	727	727	727

Notes: Own elaboration using DECID 2022. The results correspond to estimations of ATE from IPWRA models, using observations in the common support. The *teffects ipwra* command from Stata was used (Cattaneo, 2010). Bootstrap standard errors with 400 repetitions in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Regarding intermediate outcomes, Table 4 shows that there is no common pattern of effects across courses. The negative final results are driven by different performances in specific instances, both in terms of the scores and probabilities of retention.

From Tables 3 and 4, a specific pattern of effects emerges in each course. PD students who took virtual courses were less likely to be active by 9 percentage points (pp) and less likely to obtain a FPA (14 pp). This is a result of the lower likelihood of retention in the first stage for virtual students (9 pp). The greatest difference in performance was observed in the first test, where virtual students obtained, on average, a score 0.34 points lower out of a total of 5. However, no significant differences were observed in the remaining intermediate outcomes.

The final results for PE virtual students mirror those for PD, with a slightly higher effect on FPA (16 pp) and adding a significant negative effect on FA (5 pp). The sequence of intermediate results clearly differs from PD, since PE virtual students had significant lower scores in stages 2 and 3. The lower probability of retention in stage 1 (15 pp) is now partially offset by the higher retention rate in stage 3 (8 pp).

Table 4: Treatment effects on intermediate outcomes: IPWRA

Problems of Development								
	Ret1	Score1	Score2	Ret2	Score3	Score4	Ret3	Score5
ATE	-0.09***	-0.34***	0.42	-0.03	0.03	0.67*	-0.03	1.18
	(0.03)	(0.13)	(0.37)	(0.03)	(0.15)	(0.38)	(0.04)	(2.12)
N	577	547	545	516	474	511	449	318
Max	1	5	20	1	5	20	1	50
Principles of Economics								
	Ret1	Score1	Ret2	Score2	Ret3	Score3		
ATE	-0.15***	-0.37	0.00	-6.71***	0.08*	-2.18***		
	(0.05)	(0.30)	(0.04)	(2.00)	(0.04)	(0.82)		
N	712	649	649	496	496	410		
Max	1	15	1	50	1	25		
Mathematics								
	Ret1	Score1	Ret2	Score2				
ATE	0.03	-2.75***	-0.04	-6.00				
	(0.04)	(0.82)	(0.05)	(4.04)				
N	598	424	424	188				
Max	1	35	1	50				
Social Question in History Virtual								
	Ret1	Score1	Ret2	Score2	Ret3	Score3		
ATE	0.08	0.77	0.00	3.05	0.15	-8.45*		
	(0.05)	(0.85)	(0.12)	(3.16)	(0.17)	(4.61)		
N	534	522	522	484	484	403		
Max	1	20	1	35	1	45		
Social Question in History Blended								
	Ret1	Score1	Ret2	Score2	Ret3	Score3		
ATE	-0.05*	0.47	-0.05	-0.08	-0.08*			
	(0.03)	(0.30)	(0.04)	(0.53)	(0.05)			
N	583	565	565	529	529			
Max	1	20	1	35	1	45		

Notes: Own elaboration using DECID 2022. The results correspond to estimations of ATE from IPWRA models, using observations in the common support. The *teffects ipwra* command from Stata was used (Cattaneo, 2010). In the case of the CSH blended group, the number of observations did not allow obtaining an estimate of the treatment effects in Score 3. Bootstrap standard errors with 400 repetitions in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Differences between virtual and in-person students in the MAT course arise exclusively from lower scores in the first (2.75/35) test, leading to lower probabilities of FPA (7 pp) and FA (4 pp).

Turning to SQH, the estimated coefficients indicate a lower probability of activity and FPA for students in the virtual and blended groups. Intermediate results show no significant differences between the groups, except for blended course students exhibiting lower probabilities of taking the first and last tests (5 and 8 pp, respectively).

The robustness of the results is further supported by estimates obtained using alternative strategies, namely Regression Adjustment (RA) and PSM techniques, which results are presented in Appendix D.[12]

[12] In PSM results, matches are obtained using Epanechnikov Kernel functions and computing the bandwidth in each estimation through Silverman's method.

Tables D.2 and D.3 exhibit the resilience of the coefficient estimates in the final results, thus highlighting their robustness. The analysis consistently indicates that both the magnitude and statistical significance of the coefficients are preserved. While a slight reduction in significance is observed in certain cases, the coefficient values remain stable in the PSM approach.

The intermediate results obtained using PSM and RA are presented in Tables D.4 and D.5. These are similar to those obtained by the IPWRA, except for a small number of coefficients that are slightly smaller in magnitude.

6 Identification

The presented estimations of the treatment effects require observables to fully capture the differences in enrollment probabilities between students. Although we include a comprehensive set of controls, this is a bold assumption. In this section we perform two exercises aimed at assessing the stability of the results of alternative identification strategies.[13] Information restrictions prevent the extension of these two exercises to the four courses and the entire sample.

6.1 Randomized Controlled Trial

Although the OLPE program was open to all applicants for the online and blended groups, a maximum of 100 students were allowed for each group. If the quota exceeded, a lottery was held, and those who were not selected were assigned to a face-to-face group of their choice. However, most groups did not reach the cap; therefore, all applicants were accepted.[14] Furthermore, a noticeable proportion of those selected did not attend, further reducing the size of both the treatment and control groups. Therefore, the RCT strategy was employed in only one course.

In the case of PD, we were able to exploit an RCT design by bringing the two virtual groups together. The estimation of the effects in this case is a valuable exercise to test our identification strategy, because it does not require the assumption that selection is based only

[13] Following Oster (2019) we were able to estimate the relative degree of selection on unobservable controls. Table D.1 shows that it is less than 5% of the selection on observables in most cases.

[14] Draws were conducted in two groups for PD (119 and 163 applicants), two groups for MAT (113 and 157 applicants), and one online group for SQH (145 applicants).

on observables. In this case, the control group was no longer composed of all face-to-face students, but only of those who applied and were not selected. Table 5 presents the results, validating the corresponding estimations in Table 3 in terms of sign and significance, although the magnitudes were larger.[15,16]

Table 5: Treatment effects on final PD outcomes: RCT

	Active	FPA	FA
ATE	-0.15	-0.29***	-0.10
	(0.09)	(0.10)	(0.09)
N	193	193	193

Notes: Own elaboration using DECID 2022. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

As the number of control group observations is not sufficiently large to rely on asymptotic properties, alternative p-values are obtained through Randomization Inference (Heß, 2017). Table 6 shows that the p-values obtained through the permutations of the control group are similar to the classic p-values.

Table 6: Treatment effects on final PD outcomes: RCT using Randomization Inference

	Active	FPA	FA
ATE	-0.15	-0.29	-0.10
p-value	0.15	0.01	0.26
p-value RI	0.13	0.04	0.21

Notes: Own elaboration using DECID 2022. Randomization inference p-values are obtained using the *ristat* command in Stata (Heß, 2017) using 100 permutations.

6.2 Controlling for ability

We acknowledge that the lack of controls for students' ability in the previous estimations might be a source of bias. Our database does not include any variables that allow to control for students ability before starting courses. However, the FCS performs a diagnostic math test for freshmen

[15] Unclustered robust standard errors are used for inference, following Abadie et al. (2023) who show that clustering is not appropriate in RCT estimations even if there is correlation within clusters.

[16] The parameters being identified in the RCT estimations do not coincide exactly with our previous IPWRA results if selection into the program is endogenous. We estimated a Heckman selection model where the variable used for the exclusion restriction is the distance in kilometers between the capital city of the student's residence province and Montevideo. The coefficients estimated in this case were exactly the same as those presented in Table 5.

students, in which the knowledge of secondary education math courses is evaluated. This test, conducted in 2020 and 2022, is non-compulsory and has no credits assigned. Approximately half of each generation takes it; therefore, using this information leads to an important reduction in our sample.

Considering the characteristics of the courses included in our study, basic math knowledge is a reasonable proxy for the ability required in MAT and PE, while it does not seem to be relevant for SQH and PD courses. In Table 7 we include the diagnostic test scores in the outcome equation of the baseline IPWRA estimations, restricting the sample to the generation of 2020 and 2022 students who took the test.

Table 7: Treatment effects on final outcomes: IPWRA including math diagnostic scores

	Active	FPA	FA
Principles of Economics			
ATE	-0.06	-0.25***	-0.06*
	(0.05)	(0.06)	(0.03)
N	385	385	385
Mathematics			
ATE	-0.22**	-0.20***	-0.12***
	(0.09)	(0.05)	(0.02)
N	249	249	249

Notes: Own elaboration using DECID 2022. The results correspond to estimations of ATE from IPWRA models, using observations in the common support for students from generations 2020 and 2022 who took the math diagnostic test. The *teffects ipwra* command from Stata was used (Cattaneo, 2010). Bootstrap standard errors with 400 repetitions in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

These results differ only slightly from our baseline estimations for PE in Table 3. In the case of MAT, all coefficients became significant, and their magnitudes increased. This is an expected outcome, as we are now controlling for the specific abilities required in this course.

The two exercises performed in this section seem to support our main identification strategy, confirming the sign and significance of most of the estimated coefficients. Regarding magnitude, alternative strategies estimate stronger treatment effects.

7 Heterogeneity of Treatment Effects

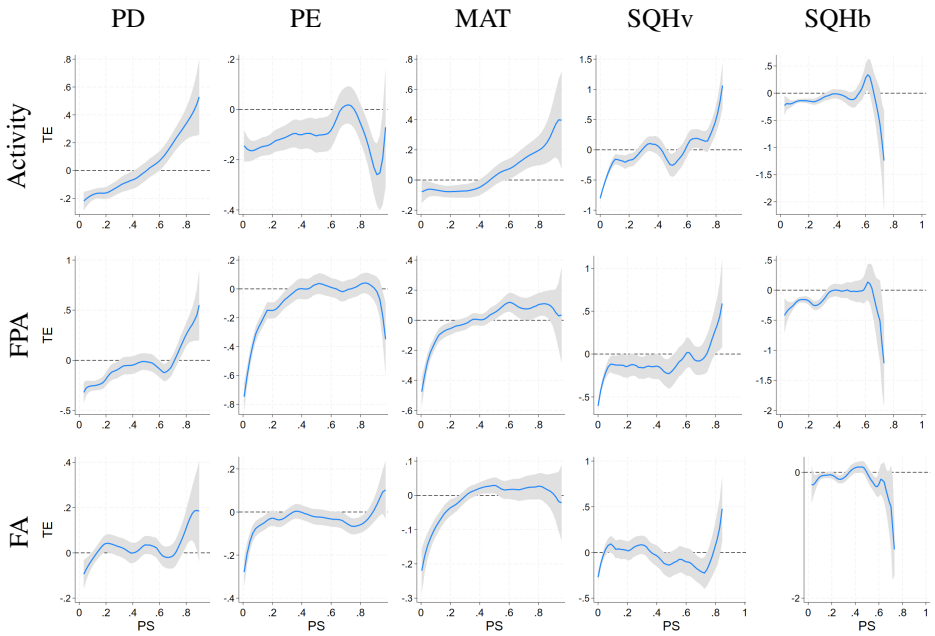
The effect of the program can differ according to student characteristics, and as long as observable variables are available, this source of bias can be controlled using matching methods. The treatment effects were estimated under the assumption that the program had the same effect on all the participants. However, students could be heterogeneous in how the program affected them, and the effect could even be positive for some groups and negative for others. This aggregation bias is ignored in standard matching techniques; however, specific approaches allow for computing the differences between the treatment and control groups when the effects are heterogeneous across the sample. As Heckman et al. (2006) demonstrate, the only interaction that results in a selection bias for causal inference under the premise of ignorability is between the treatment of interest and the propensity for selection into treatment.

Following Xie et al. (2012) we use a Matching-Smoothing Method to graphically assess the existence of heterogeneity in the program's effects. This technique is based on nonparametric local polynomial regression (Fan and Gijbels, 1996) of the difference in a pair of treated and untreated units as a function of the propensity score. We employ this method consistently throughout the evaluation sequence to assess the heterogeneous sequential effects of OLPE, as well as the heterogeneous effects on the final results.

In Figure 2 we analyze the HTEs by course and final outcomes (Activity, FPA and FA). The difference in outcomes between each observation and its counterfactual is now allowed to vary according to the value of the PS. Thus, high values of PS identify individuals who were more likely to receive treatment, such as students who were employed, had care-giving responsibilities, or resided at a considerable distance from the FCS. Estimations of PS in Appendix C show that, in addition to the aforementioned characteristics, students with high values of PS are, on average, older in age and pertain to older generations.

In most instances, negative effects were observed among students with a low probability of receiving treatment, indicating that virtual courses led to poorer outcomes for individuals who were not the target group. This negative effect was the only heterogeneous effect detected in some cases, as the effects for higher values of PS were not statistically significant (probabil-

Fig. 2: HTEs on final outcomes, by course



Notes: Own elaboration using DECID 2022. Estimations performed in Stata using *hte ms* (Jann et al., 2014). 95% confidence intervals in gray.

ity of being active in PE and SQHb, probability of FPA in PE, MAT, SQHv, and SQHb, and probability of FA in PE and MAT).

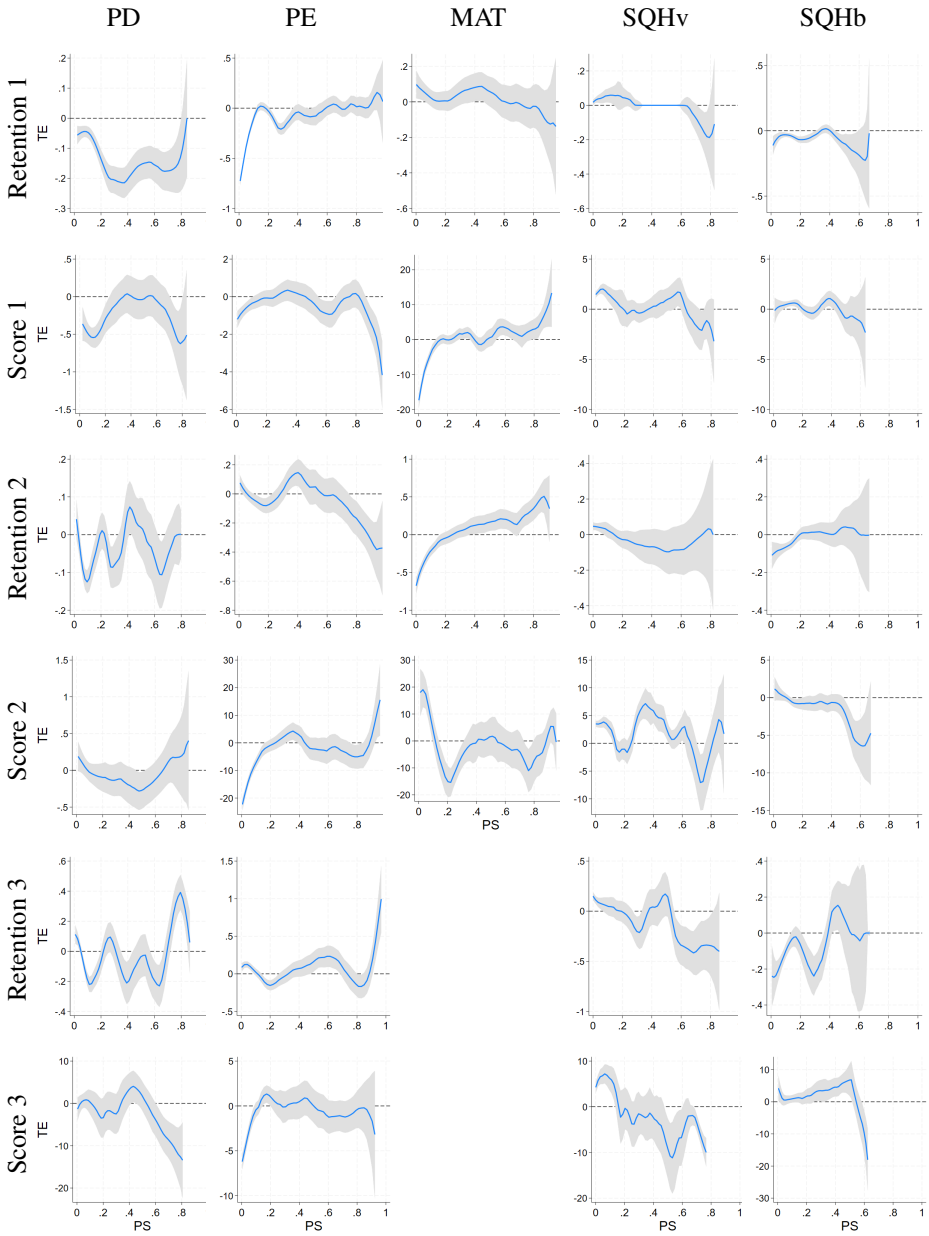
However, in other cases, negative treatment effects for low PS values were accompanied by significant positive effects for higher values of PS. These cases include the probability of being active in PD and SQHv as well as the probability of FPA in PD.

Finally, the cases that were not significant in terms of the overall effect in Table 3 remained generally insignificant across the entire range of PS, except for the probability of activity in MAT. In this case, the negative effect for individuals with low PS values appears to be offset by the positive effect for those with high PS values.

Figure 3 shows the HTEs for intermediate outcomes. In most cases where the overall result was significant and negative in Table 4, the same effect was observed for low PS values in the heterogeneity analysis (Retention 1 in PE, Test 1 in PD and MAT, Tests 2 and 3 in PE, and Retention 3 in SQHb). In one case, the negative overall effect responds to negative and significant effects in the entire range of PS (Retention 1 in PD). A particular situation emerged for Test 2 in MAT, with positive effects for the lowest values of PS and negative effects for intermedi-

ate propensity scores. Finally, Retention 1 in SQHb shows a very small negative coefficient in Table 4, which is in line with the slightly negative heterogeneous effects.

Fig. 3: HTEs on sequential outcomes, by course



Notes: Own elaboration using DECID 2022. Estimations performed in Stata using *hte ms* (Jann et al., 2014). PD Scores of group works are not presented. 95% confidence intervals in gray.

The analysis reveals that in cases in which the overall effect was not significant, HTEs were also insignificant along the PS, with some exceptions that did not describe any specific pattern.

8 Discussion

The previous sections show the negative estimated effects of virtual instruction on students' performance, in line with previous literature findings. However, further analysis makes this result more nuanced, revealing that the negative effects are not generalizable to all courses, instances, or students. The latter is particularly relevant, as online education can positively fulfill the needs of certain students.

A significant limitation of our approach is the inability to explore the mechanisms that may account for these findings. Existing literature suggests various channels through which virtuality could have a detrimental effect on student performance. One group is related to students' behavior, whereas the other focuses on interactions within the learning environment.

Procrastination, overcommitment, and time management issues significantly impact virtual learning performance, especially for those prone to task delays (De Paola et al., 2023; Doherty, 2006). Doo et al. (2023) highlight how students' self-regulation, in the line of Zimmerman (2000), is crucial for successful learning in virtual environments where there is a diminished sense of instructor control (Castro and Tumibay, 2021).

Regarding interaction factors, Doherty (2006) underscores that limited communication with instructors and insufficient contact time play a crucial role in explaining lower student retention. Failache et al. (2022), based on students' perceptions during the pandemic, report that the lack of interaction with teachers and fellow students correlates negatively with the number of approved courses and the average grade at Universidad de la República.

By exploring HTEs, we show that the estimated negative treatment effects are mostly explained by the poorer outcomes of students who are not part of the program's intended population. This indicates that students who had the opportunity to attend the course in person but chose not to were adversely affected by virtuality. Conversely, students who were the intended audience for the program, such as those with work or care-giving responsibilities or those residing far away, did not appear to be negatively affected by the virtual courses. In this case, the negative effects might be offset by the positive effect arising from the fact that virtuality allows overcoming their relative disadvantages.

Failache et al. (2022) discover a positive correlation between academic performance during the pandemic and students who reported benefiting from reduced travel times. Although OLPE was intended to target students residing far from the university, no specific strategy was implemented in that direction. Our findings underscore the importance of adequately targeting and communicating the aims and characteristics of the program, suggesting a need to discuss measures to limit access to targeted students.

These findings highlight the necessity of a well-targeted program that effectively reaches and benefits the intended audience. Analyzing the HTEs provided valuable insights into the effectiveness of OLPE for different student subgroups.

9 Conclusion

This study focused on the evaluation of the OLPE program at FCS-Udelar, which aimed to accommodate students facing challenges such as geographic constraints, employment, or caregiving responsibilities.

We analyzed data from the mandatory first-semester courses, including information on socioeconomic characteristics and the intermediate and final outcomes of all students. We estimated the coefficients using matching techniques to identify any differences in the results between students attending virtual and face-to-face courses.

The evaluation of the program highlights several important insights. The comparison between virtual and in-person students, both in terms of final outcomes and intermediate stages, revealed negative coefficients for virtual instruction, indicating poorer performance among virtual students in terms of activity and approval probability. Sequential analysis further revealed that the discrepancies between virtual and face-to-face students were a result of different performance and retention rates.

The HTE analysis revealed that the negative treatment effects primarily affected students who were not part of the program's intended population. This emphasizes the importance of adequately targeting and communicating the program's aims and characteristics. Furthermore, the study highlights that students who had the option to attend the course in person but chose virtual

learning were more adversely affected, while students with work or care-giving responsibilities or those residing far away did not show significant negative impacts from virtual courses.

Overall, the findings emphasize the necessity for a well-targeted program that effectively reaches and benefits the intended audience. Analyzing HTEs provides valuable insights into the effectiveness of the program for different student subgroups, further contributing to the understanding of virtual learning in the context of higher education in a developing country.

While online instruction has increased access to tertiary education, more work is required to ensure that all students derive maximum benefits from it. The findings of our evaluation provide valuable insights that can guide future interventions and policies aimed at enhancing access to FCS-Udelar, similar institutions in Latin America, and other developing countries.

References

Abadie A, Athey S, Imbens GW, Wooldridge JM (2023) When Should You Adjust Standard Errors for Clustering? The Quarterly Journal of Economics 138(1):1–35, DOI 10.1093/qje/qjac038

AGESIC (2020) Estudio sobre conocimientos, actitudes y prácticas de ciudadanía digital. Tech. rep., AGESIC, Montevideo, Uruguay. url: https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/datos-y-estadisticas/estadisticas

Alpert WT, Couch KA, Harmon OR (2016) A Randomized Assessment of Online Learning. American Economic Review 106(5):378–82, DOI 10.1257/AER.P20161057

Altindag DT, Filiz ES, Tekin E (2021) Is Online Education Working? Tech. rep., National Bureau of Economic Research, Cambridge, MA, DOI 10.3386/W29113

Barnow BS, Cain GG, Goldberger AS (1980) Issues in the Analysis of Selectivity Bias. In: Stromsdorfer E, Farkas G (eds) Evaluation Studies, vol 5, Sage Publications, Sanf Francisco

Bettinger EP, Fox L, Loeb S, Taylor ES (2017) Virtual Classrooms: How Online College Courses Affect Student Success. American Economic Review 107(9):2855–75, DOI 10.1257/AER.20151193

Bird KA, Castleman BL, Lohner G (2022) Negative Impacts From the Shift to Online Learning During the COVID-19 Crisis: Evidence From a Statewide Community College System. AERA Open 8(1):1–16, DOI 10.1177/23328584221081220

Bowen WG, Chingos MM, Lack KA, Nygren TI (2014) Interactive Learning Online at Public Universities: Evidence from a Six-Campus Randomized Trial. Journal of Policy Analysis and Management 33(1):94–111, DOI 10.1002/PAM.21728

Bulman G, Fairlie RW (2022) The Impact of COVID-19 on Community College Enrollment and Student Success: Evidence from California Administrative Data. DOI 10.2139/ssrn.4156927

Cacault MP, Hildebrand C, Laurent-Lucchetti J, Pellizzari M (2021) Distance Learning in Higher Education: Evidence from a Randomized Experiment. Journal of the European Economic Association 19(4):2322–2372, DOI 10.1093/JEEA/JVAA060

Castro MDB, Tumibay GM (2021) A literature review: Efficacy of online learning courses for higher education institution using meta-analysis. Education and Information Technologies 26(2):1367–1385, DOI 10.1007/s10639-019-10027-z

Cattaneo MD (2010) Efficient semiparametric estimation of multi-valued treatment effects under ignorability. Journal of Econometrics 155(2):138–154, DOI 10.1016/j.jeconom.2009.09.023

D'Almeida F, Margot D (2018) La Evolución de las Telecomunicaciones Móviles en América Latina y el Caribe. BID Invest 4(54)(54)

De Paola M, Gioia F, Scoppa V (2023) Online teaching, procrastination and student achievement. Economics of Education Review 94:102378, DOI 10.1016/j.econedurev.2023.102378

Ding W, Lehrer SF (2010) Estimating Treatment Effects from Contaminated Multiperiod Education Experiments: The Dynamic Impacts of Class Size Reductions. The Review of Economics and Statistics 92(1):31–42, 25651388

Doherty W (2006) An analysis of multiple factors affecting retention in Web-based community college courses. The Internet and Higher Education 9(4):245–255, DOI 10.1016/j.iheduc.2006.08.004

Doo MY, Bonk CJ, Heo H (2023) Examinations of the relationships between self-efficacy, self-regulation, teaching, cognitive presences, and learning engagement during COVID-19. Educational Technology Research and Development 71(2):481–504, DOI 10.1007/s11423-023-10187-3

Escueta M, Quan V, Nickow AJ, Oreopoulos P (2017) Education Technology: An Evidence-Based Review. DOI 10.3386/w23744, 23744

Failache E, Fiori N, Katzkowicz N, Machado A, Méndez L (2022) Impact of COVID-19 on Higher Education: Evidence from Uruguay. Tech. rep., Instituto de Economía, Facultad de Ciencias Económicas, Montevideo

Fan J, Gijbels I (1996) Local Polynomial Modelling and Its Applications. Monographs on Statistics and Applied Probability 66, Routledge, New York, DOI 10.1201/9780203748725

Figlio D, Rush M, Yin L (2013) Is It Live or Is It Internet? Experimental Estimates of the Effects of Online Instruction on Student Learning. Journal of Labor Economics 31(4):763–784, DOI 10.1086/669930, 10.1086/669930

Foo Cc, Cheung B, Chu Km (2021) A comparative study regarding distance learning and the conventional face-to-face approach conducted problem-based learning tutorial during the COVID-19 pandemic. BMC Medical Education 21(1):141, DOI 10.1186/s12909-021-02575-1

Heckman JJ, Urzua S, Vytlacil E (2006) Understanding Instrumental Variables in Models with Essential Heterogeneity. The Review of Economics and Statistics 88(3):389–432, 40043006

Heß S (2017) Randomization Inference with Stata: A Guide and Software. The Stata Journal 17(3):630–651, DOI 10.1177/1536867X1701700306

Jack R, Halloran C, Okun J, Oster E (2023) Pandemic Schooling Mode and Student Test Scores: Evidence from US School Districts. American Economic Review: Insights 5(2):173–190, DOI 10.1257/aeri.20210748

Jaggars SS, Xu D (2016) How do online course design features influence student performance? Computers \& Education 95:270–284, DOI 10.1016/J.COMPEDU.2016.01.014

Jann B, Brand JE, Xie Y (2014) HTE: Stata module to perform heterogeneous treatment effect analysis. Statistical Software Components

Joyce T, Crockett S, Jaeger DA, Altindag O, O'Connell SD (2015) Does classroom time matter? Economics of Education Review 46:64–77, DOI 10.1016/J.ECONEDUREV.2015.02.007

Lechner M (2004) Sequential Matching Estimation of Dynamic Causal Models. IZA Discussion Papers (1042)

Lee DS (2009) Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. The Review of Economic Studies 76(3):1071–1102, 40247633

Leuven E, Sianesi B (2018) PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Boston College Department of Economics

Okamura K, Islam N (2021) Effects of the timing of childbirth on female labor supply: An analysis using the sequential matching approach. Applied Economics 53(28):3253–3266, DOI 10.1080/00036846.2020.1855320

Oster E (2016) PSACALC: Stata module to calculate treatment effects and relative degree of selection under proportional selection of observables and unobservables. Statistical Software Components

Oster E (2019) Unobservable Selection and Coefficient Stability: Theory and Evidence. Journal of Business & Economic Statistics 37(2):187–204

Robins JM, Hernán MA, Brumback B (2000) Marginal structural models and causal inference in epidemiology. Epidemiology (Cambridge, Mass) 11(5):550–560, DOI 10.1097/00001648-200009000-00011

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55, DOI 10.1093/BIOMET/70.1.41

Rosenbaum PR, Rubin DB (1985) Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. The American Statistician 39(1):33, DOI 10.2307/2683903, 2683903

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66(5):688–701, DOI 10.1037/H0037350

Rubin DB (1980) Discussion of 'Randomization Analysis of Experimental Data in the Fisher Randomisation Test' by Basu. Journal of the American Statistical Association 75(591):593, DOI 10.2307/2287653, 2287653

Wooldridge JM (2007) Inverse Probability Weighted Estimation for General Missing Data Problems. Journal of Econometrics 141(2):1281–1301, DOI 10.1016/j.jeconom.2007.02.002

Xie Y, Brand JE, Jann B (2012) Estimating Heterogeneous Treatment Effects with Observational Data. Sociological Methodology 42(1):314–347, DOI 10.1177/0081175012452652

Xu D, Jaggars SS (2014) Performance gaps between online and face-to-face courses: Differences across types of students and academic subject areas. Journal of Higher Education 85(5):633–659, DOI 10.1353/JHE.2014.0028

Zimmerman BJ (2000) Chapter 2 - Attaining Self-Regulation: A Social Cognitive Perspective. In: Boekaerts M, Pintrich PR, Zeidner M (eds) Handbook of Self-Regulation, Academic Press, San Diego, pp 13–39, DOI 10.1016/B978-012109890-2/50031-7

Appendix A Courses' Assessments

Table A.1: Assessments' timing, types and modalities by course

	PD			PE			MAT			SQH		
Test	Week	Type	Mod	Week	Type	Mod	Week	Type	Mod	Week	Type	Mod
1	5	MC	Virt	4	MC	Virt	8	F2F	E	5	MC	Virt
2	5	E	GTH	11	E	F2F	15	F2F	E	10	E	GTH
3	12	MC	Virt	15	MC	Virt	Cont	Virt	MC	15	E	F2F
4	12	E	GTH	Cont	MC	Virt						
5	15	E	F2F									

Notes: Own elaboration. Tests weeks (column Week) are at most 15, and continuous assessments are marked as 'Cont'. Test types (column Type) are multiple choice (MC) or essays (E). Test modalities (column Mod) are virtual (Virt), face-to-face (F2F), or group take-home examinations (GTH).

Appendix B Descriptive Statistics

Table B.1: Students characteristics by course and modality

	PD			PE			MAT			SQH		
	F2F	Virt	Diff	F2F	Virt	Diff	F2F	Virt	Diff	F2F	Blend	Virt
Woman	0.75	0.74	-0.01	0.74	0.78	0.05*	0.74	0.82	0.08***	0.76	0.74	0.73
Age	23.58	29.24	5.66***	23.65	29.72	6.07***	23.78	29.63	5.86***	24.17	25.13	29.75
Res Mvd	0.57	0.50	-0.07*	0.57	0.52	-0.05	0.60	0.50	-0.09**	0.58	0.53	0.34
HS Mvd	0.44	0.43	-0.02	0.44	0.39	-0.05	0.48	0.39	-0.09***	0.46	0.37	0.30
HS Priv	0.11	0.07	-0.05**	0.10	0.06	-0.05***	0.12	0.06	-0.06***	0.10	0.08	0.09
Work Full	0.26	0.49	0.23***	0.26	0.54	0.29***	0.28	0.54	0.26***	0.32	0.30	0.58
Work Part	0.09	0.08	-0.01	0.09	0.10	0.01	0.11	0.11	0.00	0.09	0.11	0.01
Care	0.11	0.21	0.10***	0.09	0.25	0.16***	0.08	0.24	0.16***	0.10	0.18	0.24
Gen 22	0.62	0.45	-0.17***	0.61	0.40	-0.21***	0.41	0.23	-0.18***	0.63	0.53	0.47
Other Virt	0.18	0.64	0.46***	0.20	0.63	0.43***	0.19	0.47	0.28***	0.18	0.43	0.86
Repeat	0.05	0.06	0.02	0.03	0.03	0.01	0.06	0.08	0.02			
Observations	570	249	819	635	339	974	578	268	846	575	152	79

Notes: Own elaboration using DECID 2022. Tests on the equality of means were performed using *ttest* command in Stata allowing unpaired data to have unequal variances. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Appendix C Propensity Score results

Table C.1: Estimation of the Propensity Scores

	PD	PE	MAT	CSHv	CSHb
Woman	0.06	0.23**	0.29**	0.07	-0.01
	(0.12)	(0.12)	(0.13)	(0.17)	(0.12)
Age 17-19	0.00	0.00	0.00	0.00	0.00
	(.)	(.)	(.)	(.)	(.)
Age 20-24	0.21	0.53***	0.41**	0.19	0.24
	(0.15)	(0.14)	(0.17)	(0.26)	(0.15)
Age 25-34	0.53***	0.93***	0.97***	0.79***	0.36**
	(0.18)	(0.17)	(0.19)	(0.29)	(0.18)
Age 35-44	0.46**	0.89***	0.85***	0.62*	0.14
	(0.22)	(0.20)	(0.25)	(0.34)	(0.23)
Age 45+	1.11***	0.81***	1.33***	1.00***	-0.27
	(0.23)	(0.22)	(0.26)	(0.38)	(0.32)
HS Mvd	-0.14	-0.34***	-0.50***	-0.62***	-0.11
	(0.11)	(0.11)	(0.11)	(0.17)	(0.12)
HS Priv	-0.09	-0.04	0.07	0.44	-0.00
	(0.21)	(0.19)	(0.19)	(0.28)	(0.19)
Work Full	0.30**	0.50***	0.35***	0.10	-0.33**
	(0.14)	(0.13)	(0.13)	(0.19)	(0.15)
Work Part	0.17	0.42**	0.27	-1.21***	0.06
	(0.20)	(0.17)	(0.18)	(0.47)	(0.19)
Care	0.09	0.41***	0.58***	0.35	0.34**
	(0.15)	(0.14)	(0.15)	(0.22)	(0.16)
Gen 19	-0.50*	-0.01	-0.01		
	(0.30)	(0.24)	(0.21)		
Gen 20	-0.03	-0.42	-0.02	0.67*	-0.13
	(0.35)	(0.27)	(0.23)	(0.40)	(0.36)
Gen 21	-0.28	-0.09	0.18	-0.90**	0.37**
	(0.19)	(0.17)	(0.15)	(0.37)	(0.19)
Gen 22	-0.36**	-0.49***	-0.49***	-0.34*	-0.05
	(0.14)	(0.13)	(0.15)	(0.18)	(0.14)
Other virt PD	1.17***				
	(0.11)				
Repeat PD	0.31				
	(0.23)				
Other virt PE		1.21***			
		(0.10)			
Repeat PE		0.14			
		(0.26)			
Other virt MAT			0.97***		
			(0.12)		
Repeat MAT			0.20		
			(0.21)		
Other virt CSHv				1.53***	
				(0.15)	
Other virt CSHs					0.46***
					(0.12)
Constant	-1.16***	-1.52***	-1.56***	-2.30***	-1.10***
	(0.19)	(0.18)	(0.21)	(0.26)	(0.18)

Notes: Own elaboration using DECID 2022. Regressions were performed using the *probit* command in Stata. Age 17-19 is the reference category in age ranges. Generations previous to 2019 are the reference category. $^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$

Appendix D Robustness to Different Specifications

D.1 Final Results

Table D.1: OLS Estimations

	Active	FPA	FA
Problems of Development			
ATE	-0.05	-0.09**	0.02
	(0.04)	(0.04)	(0.03)
delta	0.035	0.049	-0.060
N	819	819	819
Principles of Economic			
ATE	-0.08**	-0.05	-0.03
	(0.04)	(0.04)	(0.02)
delta	0.043	0.029	0.026
N	974	974	974
Mathematics			
ATE	0.00	0.02	0.00
	(0.04)	(0.03)	(0.01)
delta	-0.003	-0.027	-0.007
N	846	846	846
Social Question in History Virtual			
ATE	-0.02	-0.10	-0.00
	(0.07)	(0.07)	(0.06)
delta	0.014	0.044	0.000
N	654	654	654
Social Question in History Blended			
ATE	-0.11***	-0.14***	-0.04
	(0.04)	(0.04)	(0.04)
delta	0.230	0.226	0.052
N	727	727	727

Notes: Own elaboration using DECID 2022. Linear Probability Model estimation using command *reg* in Stata. Delta proportions computed using *psacalc* command in Stata (Oster, 2016). Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.2: Treatment effects on final outcomes: PSM

	Active	FPA	FA
Problems of Development			
ATE	-0.08*	-0.15***	0.00
	(0.05)	(0.05)	(0.03)
N	819	819	819
Principles of Economics			
ATE	-0.13***	-0.14***	-0.04
	(0.05)	(0.04)	(0.03)
N	974	974	974
Mathematics			
ATE	-0.03	-0.03	-0.02*
	(0.04)	(0.03)	(0.01)
N	846	846	846
Social Question in History Virtual			
ATE	-0.17	-0.12	0.05
	(0.14)	(0.11)	(0.08)
N	654	654	654
Social Question in History Blended			
ATE	-0.12***	-0.16***	-0.05
	(0.04)	(0.05)	(0.04)
N	727	727	727

Notes: Own elaboration using DECID 2022. Matching performed using Kernel algorithm with Epanechnikov function and a bandwith computed using Silverman's method for each subsample. Bootstrap standard errors with 400 repetitions in parentheses. Results obtained using *psmatch2* command in Stata (Leuven & Sianesi, 2018). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.3: Treatment effects on final outcomes: RA

	Active	FPA	FA
Problems of Development			
ATE	-0.10**	-0.14***	0.01
	(0.04)	(0.04)	(0.03)
N	819	819	819
Principles of Economics			
ATE	-0.11***	-0.11***	-0.04
	(0.04)	(0.04)	(0.03)
N	974	974	974
Mathematics			
ATE	-0.06	-0.05*	-0.04***
	(0.05)	(0.03)	(0.01)
N	846	846	846
Social Question in History Virtual			
ATE	-0.34**	-0.24	-0.00
	(0.15)	(0.15)	(0.14)
N	654	654	654
Social Question in History Blended			
ATE	-0.12***	-0.16***	-0.05
	(0.04)	(0.05)	(0.04)
N	727	727	727

Notes: Own elaboration using DECID 2022. The results correspond to estimations of ATE from RA models, using observations in the common support. Bootstrap standard errors with 400 repetitions in parentheses. The *teffects ra* command from Stata was used (Cattaneo, 2010). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

D.2 Sequential Effects: Intermediate Evaluations

Table D.4: Treatment effects on intermediate outcomes: PSM

Problems of Development								
	Ret1	Score1	Score2	Ret2	Score3	Score4	Ret3	Score5
ATE	-0.10**	-0.40**	0.48	-0.07*	-0.04	0.73**	-0.06	0.57
	(0.04)	(0.16)	(0.40)	(0.04)	(0.17)	(0.35)	(0.06)	(3.05)
N	577	547	545	516	474	511	449	318
Max	1	5	20	1	5	20	1	50
Principles of Economics								
	Ret1	Score1	Ret2	Score2	Ret3	Score3		
ATE	-0.11**	-0.30	-0.06	-5.13*	0.03	-1.19		
	(0.05)	(0.31)	(0.05)	(2.76)	(0.05)	(1.07)		
N	712	649	649	484	484	398		
Max	1	15	1	50	1	25		
Mathematics								
	Ret1	Score1	Ret2	Score2				
ATE	0.02	-0.65	-0.02	-6.04				
	(0.05)	(1.21)	(0.08)	(4.03)				
N	598	424	424	188				
Max	1	35	1	50				
Social Question in History Virtual								
	Ret1	Score1	Ret2	Score2	Ret3	Score3		
ATE	0.02	0.70	-0.04	2.20	-0.05	1.02		
	(0.02)	(0.64)	(0.07)	(1.65)	(0.12)	(4.20)		
N	490	478	478	440	440	370		
Max	1	20	1	35	1	45		
Social Question in History Blended								
	Ret1	Score1	Ret2	Score2	Ret3	Score3		
ATE	-0.03	0.43	-0.04	-0.48	-0.08	1.56		
	(0.03)	(0.32)	(0.04)	(0.68)	(0.06)	(1.70)		
N	583	565	565	529	529	420		
Max	1	20	1	35	1	45		

Notes: Own elaboration using DECID 2022. The matching was carried out using a Kernel algorithm with Epanechnikov function and bandwidth determined in each case by the Silverman rule (1986). Results obtained using *psmatch2* command in Stata (Leuven & Sianesi, 2018). Bootstrap standard errors with 400 repetitions in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.5: Treatment effects on intermediate outcomes: RA

Problems of Development								
	Ret1	Score1	Score2	Ret2	Score3	Score4	Ret3	Score5
ATE	-0.10***	-0.31*	0.68*	-0.03	-0.07	0.55	-0.10*	1.33
	(0.04)	(0.16)	(0.39)	(0.03)	(0.16)	(0.38)	(0.05)	(2.48)
N	577	547	545	516	474	511	449	318
Max	1	5	20	1	5	20	1	50
Principles of Economics								
	Ret1	Score1	Ret2	Score2	Ret3	Score3		
ATE	-0.08*	-0.17	0.01	-2.25	0.00	0.06		
	(0.04)	(0.35)	(0.05)	(2.25)	(0.06)	(1.13)		
N	712	649	649	496	496	410		
Max	1	15	1	50	1	25		
Mathematics								
	Ret1	Score1	Ret2	Score2				
ATE	0.04	-0.73	0.04	-7.27*				
	(0.05)	(1.12)	(0.06)	(4.10)				
N	598	424	424	188				
Max	1	35	1	50				
Social Question in History Virtual								
	Ret1	Score1	Ret2	Score2	Ret3	Score3		
ATE	0.10	0.59	0.04	6.16*	0.02	-1.56		
	(0.08)	(1.14)	(0.13)	(3.41)	(0.19)	(5.26)		
N	534	522	522	484	484	403		
Max	1	20	1	35	1	45		
Social Question in History Blended								
	Ret1	Score1	Ret2	Score2	Ret3	Score3		
ATE	-0.04	0.45	-0.05	-0.26	-0.10**			
	(0.03)	(0.31)	(0.04)	(0.60)	(0.05)			
N	583	565	565	529	529			
Max	1	20	1	35	1	45		

Notes: Own elaboration using DECID 2022. The results correspond to estimations of ATE from RA models, using observations in the common support. Bootstrap standard errors with 400 repetitions in parentheses. The *teffects ra* command from Stata was used (Cattaneo, 2010). In the case of the SQH blended group, the number of observations did not allow obtaining an estimate of the treatment effects in Score 3. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.6: Treatment effects on intermediate PD outcomes: RCT

	Ret1	Score1	Score2	Ret2	Score3	Score4	Ret3	Score5
ATE	-0.20***	-0.75**	0.40	0.02	-0.51*	0.62	-0.05	-4.01
	(0.05)	(0.33)	(0.82)	(0.06)	(0.29)	(0.54)	(0.10)	(3.76)
N	125	118	109	102	93	104	84	51
Max	1	5	20	1	5	20	1	50

Notes: Own elaboration using DECID 2022. Robust standard errors of the estimators in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$