I Z A Institute of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17526

The Effect of Extended Education on Educational Quality and Inequality: An Event Study Approach

Ron Diris
Kim Fairley

DECEMBER 2024

DISCUSSION PAPER SERIES

IZA DP No. 17526

The Effect of Extended Education on Educational Quality and Inequality: An Event Study Approach

Ron Diris
Leiden University and IZA

Kim Fairley
Radboud University

DECEMBER 2024

ISSN: 2365-9793

IZA DP No. 17526 DECEMBER 2024

ABSTRACT

The Effect of Extended Education on Educational Quality and Inequality: An Event Study Approach*

This study examines the effect of extended education on school achievement and inequality in Dutch primary schools. We apply a panel event study design using rich longitudinal data on the use of extended education and school achievement in grades 1 through 6, to estimate the causal effect of extended education. The analysis reveals (precisely estimated) zero or low effects from the use of extended education. Interestingly, we identify a modest Ashenfelter dip right before the start of extended education, suggesting a reaction to an incidental poor school result. We explain the overall low effectiveness by the typical low-intensity use of extended education among Dutch primary school students, while we also identify high effectiveness for (a very small subset) of more intensive forms. We conclude that extended education has no meaningful implications for educational achievement or inequality in Dutch primary education.

JEL Classification: I21, I24

Keywords: education economics, extended education, tutoring, event study design

Corresponding author:
Ron Diris
Department of Economics
Leiden University
Steenschuur 25
2311 ES Leiden
The Netherlands

E-mail: r.e.m.diris@law.leidenuniv.nl

* We would like to thank Ernst-Jan de Bruijn, Marion Collewet, Max van Lent, Suzanne Pietersma, Toep van Dijk, Gijs van der Vlugt, attendants of the Leiden Economics seminar on November 14 2024, the 2024 LEER conference in Leuven, Belgium, the 2024 International Workshop on Applied Economics of Education in Catanzaro, Italy, the 2024 KVS New Paper Sessions in The Hague, the Netherlands, the seminar on March 7 2023 at the Ministry of Education in The Hague, the Netherlands, and the Educational Inspectorate conference on September 6 2023 in Noordwijkerhout, the Netherlands, for their helpful comments. This research has received funding from the Dutch Ministry of Finance.

1 Introduction

The use of shadow education and other forms of supplemental or extended education is expanding worldwide. While traditionally mainly prevalent in South-East Asia (Bray, 1999), it has spread rapidly in Europe (Bray, 2021) and the United States (Fortune Business Insights, 2021) in the last two decades. The private tutoring market in the United States has been estimated at 24.9 billion U.S. dollars in 2020 (Fortune Business Insights, 2021). Figures for the Netherlands, the context of this study, show that tutoring expenses have increased from 26 million in 1995 to 284 million in 2018 (Statistics Netherlands, 2021). This indicates that students are spending a substantial and increasing part of their learning time outside regular school hours. There is still little empirical evidence on how the substantial investments in such practices impact student learning. Additionally, extended education is usually offered to selected subsets of students, in sharp contrast to in-school services which are typically available to all children. There is a continuous debate on whether this selected use might progress educational equity (by being used to address achievement gaps for disadvantaged students) or hinder it (by being used primarily by higher-achieving children with wealthier parents), see, e.g., OECD (2016); De Donder and Martinez-Mora (2017); Byun et al. (2018).

The aim of this study is to identify the impact of extended education (EE) on both the effectiveness and equality of student learning. We use an event study design, which compares performance before and after the start of EE. We exploit rich longitudinal test data, which are matched to survey data that registers the current and past use of EE. Dutch primary school students take standardized tests, in multiple subjects, twice per year in each grade of primary school. The 1,531 matched students from the survey data made a total of 28,159 tests. The analysis reveals (precisely estimated) zero or low effect sizes from the use of EE on achievement. Our preferred specification provides a point estimate of 0.015 of a standard deviation, with a 95% confidence interval that excludes effects above 0.045 of a standard deviation. We do find a statistically significant positive effect for reading (0.040) but mainly in the short run.

We further identify a small Ashenfelter's dip (Ashenfelter, 1978), in which achievement is at the exact same level at T $\leq$ -2 and T $\geq$ 0, but dips right before the start of EE (T=-1). Hence, EE is initiated just after an uncommonly poor performance of the student. We similarly identify no statistically or economically significant effect from EE when we split by type and level of expense of EE, or targeted population group. The only exception applies to a (very small subset of) highly intensive types of tutoring, which produce strongly positive and statistically significant findings.

EE is a broad concept and can be studied in a variety of settings. Our choice of target group and scope is driven by the potential relevance of EE towards educational quality and equity. First, we focus on primary school students. Current evidence on EE is mainly focused on secondary school.[1] Educational achievement gaps typically arise and grow in primary school and stabilize largely thereafter (see, e.g., Carneiro and Heckman (2002)). Discussions on why such early achievement gaps arise are generally centered around quality differences in preschool and primary school education and in parental environment, but rarely consider extended education as a potential mechanism.

In addition, we consider various types of educational instruction that takes place outside of regular school hours. While extended (or shadow) education is often strictly defined as paid use of additional schooling after regular school hours, our definition contains both the paid use through private suppliers (both online and offline) and the subsidized use offered through schools. We choose this broad setup for various reasons. Only focusing on paid use of EE offered through private parties would not capture schools' involvement in providing EE as well. Students with high socioeconomic status (hereafter referred to as SES) are expected to rely more on private educational assistance, but this may be countered if low SES students receive similar instruction that is made available for free through schools. Naturally, quality differences between these two types of EE are highly relevant for their implications for educational inequality. The broad setting of this study allows us to estimate the effectiveness of each type separately.

[1]E.g., in a recent review by Luo and Chan (2022), 73 percent of included studies focus on secondary school students.

The analysis of longitudinal information on the usage of EE and on standardized testing data at the individual level provides a key contribution to the literature. Reliable causal empirical evidence on this topic is rare because of lack of data availability and selection bias. Since, by definition, EE takes place after school hours, registration of the use of EE is typically poor and limited to (current) use at one particular time.[2] From a methodological perspective, effectiveness is difficult to assess given the endogenous selection into EE (Zumbuehl et al., 2024). Moreover, most studies on EE are conducted in Asia (68,9 percent according to a recent review by Luo and Chan (2022)), hence evidence on the effectiveness of non-Asian extended educational practices is sparse, let alone causal estimates.

While causal evidence on the effectiveness of commonly used forms of EE is rare, there is an extensive literature that uses randomized controlled trials to evaluate the use of high-impact tutoring (HIT) programs, typically set within the United States; see, e.g., Dietrichson et al. (2017); Fryer Jr (2017); Guryan et al. (2023); Nickow et al. (2024). These studies demonstrate positive estimates of the effectiveness of these (typically intensive and costly) interventions. Interestingly, when we narrow down our treatment to mimic the conditions of HIT tutoring (small groups, multiple hours per week), we do find strong positive and statistically significant effects. As this form of tutoring is rare in our sample, the overall contribution of EE to educational quality and equity remains low. By estimating the effectiveness of "typical forms" of EE, the current study analyzes its contribution to educational achievement and inequality "as it is", rather than "as it could be" (if high intensity and high quality programs would be provided at a large scale). This can inform the current discussion on whether more public resources should be made available for EE, and thus reduce the presumed socioeconomic gaps in (quality of) access.

The remainder of the paper is organized as follows. Section 2 discusses relevant institutional details of Dutch primary education. In Section 3, we explain data collection and present descriptive statistics. The empirical strategy is laid out in Section 4. Section 5 provides and discusses results

[2]Exceptions are provided by, e.g., Ku et al. (2022); Ryu and Kang (2013), both focusing on private tutoring in secondary school

of the empirical analysis. Section 6 concludes.

2 Institutional context

Dutch primary schools consist of two years of kindergarten (ages 4-6) and grades 1 to 6 (ages 6-12). Education is compulsory from the age of 5 onward. With very few exceptions, Dutch primary schools are publicly funded. Most students transition to secondary education at the age of 12. There are three different tracks of secondary education: vocational, general and pre-academic.

The allocation to tracks in secondary education is based on the primary school teacher's recommendation and the results of a nationwide standardized high-stakes test at the end of grade 6. Teacher recommendations are provided before the test, but can be adjusted upward in case of a good testing result. Many Dutch primary schools additionally administer standardized tests from grades 1 to 6 to track student's performance over time. The Netherlands Cohort Study on Education (in Dutch abbreviated as NCO, for *Nationaal Cohortonderzoek Onderwijs*) collects these testing data. This longitudinal data makes it possible to analyse the development of student performance across grades 1 to 6 (Haelermans et al., 2020). As these tests serve as an important input for the pre-advice given at the start of grade 6, they can be considered high-stakes.[3] Test results are communicated to students and parents. We will refer to these data as the Dutch student tracking system.

School participation in the Dutch student tracking system is voluntarily. The system is relatively new, and participation has been steadily increasing. Currently, around 35% of all Dutch primary schools register their testing data with Statistics Netherlands, meaning that they are participating in the student tracking system and providing permission for these data to be registered externally. The Dutch educational context with standardized testing in primary schools make this an ideal setting to study the effectiveness of extended education.

[3] They are also an input for decisions on whether to retain or progress students in earlier grades.

3 Data

For the purpose of this research, a survey on EE was sent out by Kantar International to a representative sample of the Dutch population. 7,000 households with school-aged children were invited to answer the survey and 3,429 parents participated, who answered on behalf of 4,695 children in total. We merge the survey data with standardized test results from the Dutch student tracking system, registered at Statistics Netherlands. In line with the school participation rate in the Dutch student tracking system, we are able to match 1,531 students to the testing data, for a total of 28,159 tests. We find no evidence of differences between the matched sample and the full questionnaire sample on observable characteristics.[4] Table 1 shows summary statistics on demographic characteristics of the sample, also split by use of extended education.

The main survey question asked parents whether they used EE for their child this school year. In asking parents whether they make use of EE and in which form, we provide the following options:

Figure 1: Main survey question on extended education

Does your child use extended education this schoolyear?

a) Yes, general tutoring
b) Yes, homework support
c) Yes, test training
d) Yes, tutoring for specific learning needs (e.g. dyslexia, giftedness)
e) Yes, supplemental online software

Our main goal is to analyze the implications of EE for Dutch primary education as broadly as possible (and to analyze more specific forms in subsequent analyses), which is reflected in the setup of the question. Additionally, the administration of a preliminary survey to a focus group of parents induced multiple question on whether tutoring for "specific needs" such as dyslexia and giftedness should be considered part of EE. Similar questions were asked with respect to the increased use of commercial educational software for after-school study. We therefore specifically included these

[4] A joint significance test on parental education, income, marital status, working hours and homework help fails to be rejected. It should be noted that the overall use of EE is slightly lower in the matched sample (32% versus 34%).

forms in the main survey question as well, thereby also allowing to consider the impact of each separate form as well. Because very few parents selected options b) and c), and because they are subforms of tutoring, we label options a) through c) as “general tutoring”. We label option d) as “targeted tutoring” and option e) as “online extended education” for the remainder of this paper.

If answering yes, parents could indicate the specific format of EE, its intensity, and whether they paid for it privately. Parents were also asked in which month EE started in order to precisely pinpoint the start of the event. If parents did not use EE for their child in the current school year, they were asked whether they used it in the past by repeating the question for each previous school year. Hence, the survey is administered at a single point in time but relies on parents answers to a retrospective question to deduce the starting point of EE. This allows us to categorize our treatment variable longitudinally. Parents were clearly instructed to assess the use of EE during normal in-school education periods, and to disregard schooling during Covid-19 lockdowns. The survey was administered at the end of March 2022, when no lockdown was effective in the Netherlands.

For the sample that was matched to the test score data, 32.1% had ever used EE during their primary school career (see Table 2). When considering the use of EE by parental education, we find a 28.9% share for those without highly educated parents and a 33.8% share for those with at least one highly educated parent.[5] This difference is marginally statistically significant (p=0.051).[6] Interestingly, the shares are virtually equal in the full sample (34.6% vs. 34.2%; p=0.759). In any case, there are at most small differences in the use of EE by parental education. This also holds when we look at other parental background variables; see Table 1. The main differences between users and non-users are that the former tend to be older, and that their parents are more involved in helping with school homework.

Table 2 further details the characteristics of EE, and how these differ by parental education. Of

[5] Parental education is defined according to the official definition of low, middle and high education by Statistics Netherlands. The categories low and middle translate to a non-college degree, and the high category to a college degree, according to international standards. We classify students as either without highly educated parents (non-HE) or with at least one higher educated parent (HE).

[6] As shown in Table 1, the difference by dropout status is statistically significant at the 5% level, although this is a relatively small group.

those that engaged in EE, 10% of students engaged in general forms of tutoring, 41% in tutoring that was targeted towards specific needs, and 49% uses online educational software. The latter is used more often by children of highly educated parents, while general tutoring is more common for students without highly educated parents. The difference by parental education is strongest when we distinguish paid versus unpaid forms. Additionally, there are marginally significant differences in supplier (more often through school for children without highly educated parents) and intensity (more often for more than 2 hours for children without high educated parents). Although not statistically significant, it is worth noting that the use of 1-on-1 forms is more prevalent among children with at least one higher-educated parent (HE) (60% versus 52%). Hence, low-SES students primarily used subsidized forms of extended education that were offered through their primary school, while high SES students more often used paid forms, typically offered outside the school setting. This result, for one, shows that the presumed inequality in access to EE largely disappears if we consider unpaid forms as well. Additionally, this implies that the impact of EE on educational inequality will largely depend on the relative effectiveness of these different forms, and mainly those of paid versus unpaid EE.[7]

4 Methodology

Examining the effectiveness of extended education is complicated by selection effects. Selection can be both positive - parents striving for excellence in their child(ren) leading to EE mainly being adopted by high-performing students - as well as negative - parents trying to remediate achievement gaps leading to EE mainly being adopted by low-performing students.

In identifying the causal link between extended education and school results, a panel event study design (ESD) is employed. This approach allows us to condition on individual fixed effects and compare within-student test performance before and after the initiation of EE. An additional ad-

[7] It should be noted that the power of some of these comparisons is relatively low, because they operate on the subset of 492 treated students, and in some cases also exclude the online form. The group size variable excludes online forms because there is no group size to be determined in this case. We also exclude the online form when we compare suppliers, since "online" is a supplier type by itself and we would be repeating this dimension.

Table 1: Summary statistics

	Mean	Std. dev.	EE=0	EE=1	P-value
Age (in 2022)	9.39	1.554	9.31	9.55	0.005
Female	0.500	0.500	0.495	0.512	0.523
Parents dropout	0.053	0.224	0.062	0.035	0.027
Parents vocational education	0.295	0.456	0.302	0.278	0.341
Parents high education	0.653	0.476	0.636	0.687	0.051
Parents low income	0.224	0.417	0.226	0.218	0.752
Parents high income	0.256	0.437	0.249	0.272	0.393
Parents married	0.662	0.473	0.660	0.667	0.804
Grade level (in 2022)	3.94	1.540	3.88	4.07	0.022
Hours worked	27.66	12.26	27.84	27.29	0.405
Hours worked partner	32.19	12.31	31.80	33.05	0.079
Hours homework help	1.13	1.543	0.975	1.46	0.000
Hours homework help lagged	1.14	1.825	0.959	1.52	0.000

Notes: 'EE=0' and 'EE=1' report means for non-users and users of extended education, separately. The final column shows the p-value of a Wald test on the statistical significance of the difference. Statistics are reported for all survey participants who could be matched to testing data (N=1,531).

vantage of the panel event study approach is that it allows for a visual inspection of the underlying assumption of parallel trends before the event. Furthermore, this approach allows for the inspection of any dynamics in the effects (Clarke and Tapia-Schythe, 2021). The econometric specification of the panel event study design can be presented as:

$$y_{st} = \sum_{j \neq -1} \beta_j b_{stj} + \mu_s + \lambda_t + \omega + \varepsilon_{st} \tag{1}$$

The model analyzes test outcomes for student s at time t by a set of events j indicated as lags ($j \leq$ -2) and leads ($j \geq 0$) (compared to the baseline event -1 which indicates the last period before initiation of extended education), student fixed effects μ_s, time fixed effects λ_t, and a collection of test fixed effects labelled as ω, which include test month, test year, and test version.[8] Standard errors are clustered at the student level. The model is estimated with the 'eventdd' command in

[8]Versions indicate whether the test was pen and paper or digital, and for which grade and testing period it is designed. In the analyses that pool all tests, it also incorporates the school subject.

Table 2: Characteristics of extended education

	Non-HE	HE	Total	P-value
All EE (N=1,531)	28.9%	33.8%	32.1%	0.051
Type (N=492)				
General tutoring	15.6%	7.4%	10.0%	0.034
Tutoring for spec. need	40.3%	41.1%	40.9%	0.326
Online supplemental education	44.2%	51.5%	49.2%	0.023
Intensity (N=492)				
1 hour	39.6%	47.0%	44.7%	0.125
2 hours	20.8%	22.2%	21.8%	0.726
>2 hours	39.6%	30.8%	33.5%	0.054
Cost burden (N=492)				
Paid by parent	31.2%	43.8%	39.8%	0.008
Supplier (N=250)				
School	53.7%	43.7%	46.8%	0.093
Private	40.7%	46.7%	44.8%	0.318
Acquintance	5.6%	9.6%	8.4%	0.214
Group size (N=250)				
1-on-1	51.7%	59.6%	57.0%	0.228
2 to 10	39.1%	33.2%	35.1%	0.343
>10	9.2%	7.3%	7.9%	0.594
Alternative definitions (N=1,531)				
Tutoring (Nickow et al., 2024)	11.7%	12.2%	12.0%	0.749
HIT (Robinson and Loeb, 2021)	1.7%	2.9%	2.5%	0.147

Notes: Non-HE are those without highly educated parents. HE are those with at least one highly educated parent. Type, intensity and costs are reported for the subgroup that has used EE (N=492). Supplier and group size are reported for users excluding online forms (N=250). The final column shows the p-value of a Wald test on the statistical significance of the difference between HE and non-HE students.

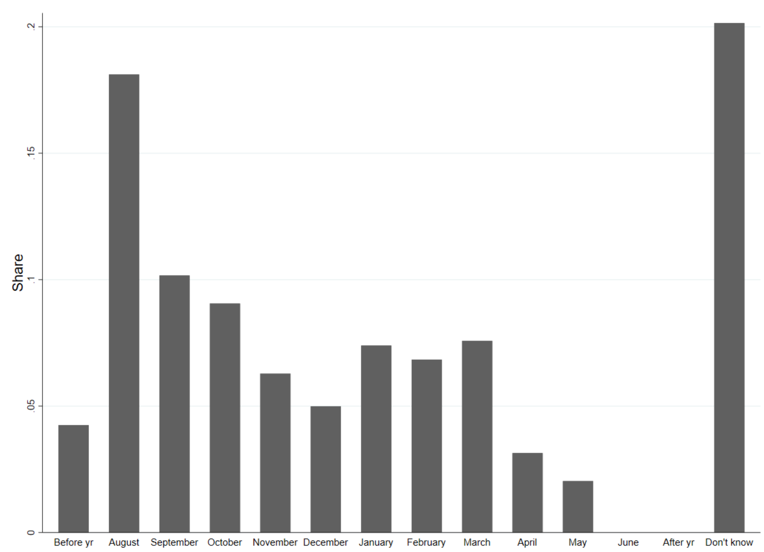
Stata (Clarke and Tapia-Schythe, 2021).

We also estimate a complementary two-way fixed effect model, for which the main treatment variable E_{st} measures whether the student is engaged in EE at the moment of the test:

$$y_{st} = \theta E_{st} + \mu_s + \lambda_t + \omega + \varepsilon_{st} \tag{2}$$

Whereas the benefit of the ESD model is that it provides a complete overview of pre-treatment and post-treatment trends, the benefit of the fixed effect model is that it captures the treatment effect in a single coefficient θ. Furthermore, this model is better equipped to deal with the fact that some students stop using EE between initial treatment and the last observed test period.

Figure 2: Distribution of starting months of extended education



Notes: The figure shows the distribution of calendar months in which extended education started (parent-reported)

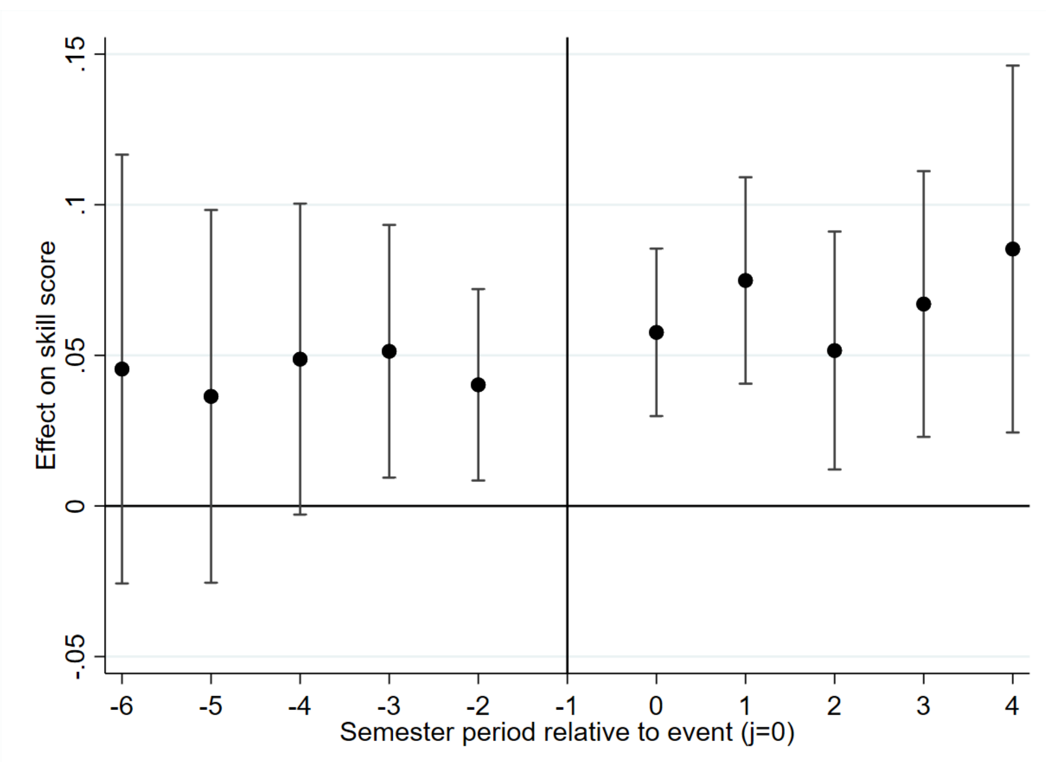
The design requires identifying the timing of both the standardized achievement tests and EE. The former is based on the administrative test data, which records the exact date of the test. For the latter, we rely on the survey data which asks parents about the school year and month in which EE started. The distribution of starting months is shown in Figure 2. Around 20% of parents indicate that they do not remember the starting month. We recode missing answers to the mode starting month of August. Note that tests are typically administered in January and June, hence setting this date to September or December does not impact the timing (the whole period before January covers around two thirds of all non-missing answers). Robustness analysis will analyse whether results are sensitive to how missing starting months are treated.

5 Results

5.1 Main Results

We first present results from a naive OLS regression model that does not take selective use of EE into account. This analysis identifies a negative effect from the use of EE on achievement, see Table 3. This either implies that the effectiveness of EE is negative, or that EE is mainly used for remedial purposes by students of lower (conditional) ability.

Figure 3: Event study analysis: main approach



Notes: The figure shows the results of estimating Model 1 for the full sample of students and using all tests.

Table 3: The effect of extended education on school achievement: OLS and fixed effects estimates

	OLS	FE all	FE no T=–1	N
All EE				
All tests	-0.119***	0.025*	0.015	28,164
	(0.022)	(0.013)	(0.015)	
Reading	-0.047**	0.046***	0.040**	7,751
	(0.021)	(0.015)	(0.017)	
Spelling	-0.238***	0.018	0.004	10,058
	(0.035)	(0.023)	(0.026)	
Arithmetic	-0.063***	0.014	0.005	10,353
	(0.022)	(0.015)	(0.017)	
Paid EE				
All tests	-0.025	0.039*	0.026	28,164
	(0.031)	(0.021)	(0.024)	
Reading	-0.003	0.046**	0.043*	7,751
	(0.031)	(0.023)	(0.026)	
Spelling	-0.095*	0.038	0.019	10,058
	(0.051)	(0.037)	(0.042)	
Arithmetic	0.023	0.035	0.025	10,353
	(0.031)	(0.025)	(0.027)	
Unpaid EE				
All tests	-0.166***	0.015	0.006	28,164
	(0.029)	(0.017)	(0.020)	
Reading	-0.070**	0.044*	0.036	7,751
	(0.027)	(0.018)	(0.022)	
Spelling	-0.299***	0.003	-0.007	10,058
	(0.045)	(0.029)	(0.033)	
Arithmetic	-0.113***	-0.001	-0.009	10,353
	(0.029)	(0.019)	(0.022)	

Notes: The table shows the estimated effect of extended education across different specifications (in columns) and test subjects (rows). The second and third panel only consider respectively paid (either partly or fully) EE or unpaid EE, in separate regressions. The third column excludes the last pre-treatment period, in light of the identified Ashenfelter dip. Standard errors are clustered at the student level.
$*p < 0.10, **p < 0.05, ***p < 0.01$

Figure 3 captures the results of our event study design (underlying coefficients are provided in Appendix Table A1). Estimates are expressed relative to the baseline period of -1, which is the last test period before EE started. Although it seems that EE has a positive effect on school achievement if we move from period -1 to 0, upon closer inspection one can see that event -1 is the single deviant period. This pattern can be classified as an Ashenfelter's dip (Ashenfelter, 1978), and suggests that parents overreact to a spurious negative test result of their child.[9]

Disregarding this dip, Figure 3 shows that student achievement is unresponsive to the use of EE.[10] The two-way fixed effects estimates can be observed in the first column of Table 3. We make a distinction between estimates for the full sample, and for the sample excluding $T = -1$. While the Ashenfelter's dip biases the estimates upward, the comparison shows that the impact of this bias is small. Estimates across all tests are marginally significant with an effect of 0.025 of a standard deviation when including all time periods, and 0.015 (statistically insignificant) without $T = -1$.

Figure 4 shows a largely similar pattern, for arithmetic, reading and spelling separately. Reading shows a comparatively stronger increase in achievement just after treatment, although this appears short-lived. This is reflected in Table 3 by a statistically significant impact for reading, also when we exclude the Ashenfelter's dip. The magnitude of the effect (0.046 or 0.040) is small, however.[11]

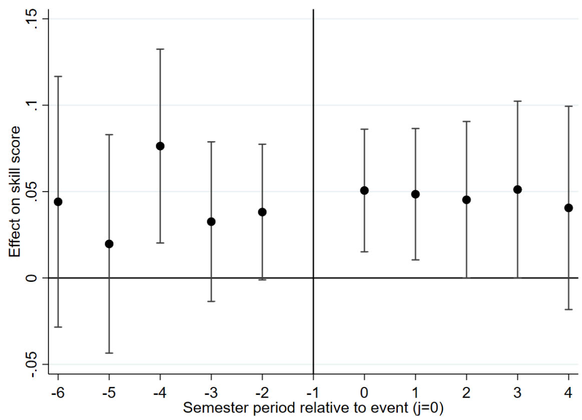
Aside from direct impacts on school achievement, EE may also impact other skill investments. The survey data asks parents about the time they spent helping their children with their homework in the current and previous year. This allows us to analyse whether EE complements or substitutes parental inputs. Note that this analysis is limited to cases in which EE started within one year before the survey. Appendix Table A5 shows that there is a positive relation between starting EE and parental homework support. This is most apparent for online extended education, although

[9] Appendix Table A6 analyzes heterogeneity in the degree of the Ashenfelter's dip. It shows that the dip is of rather similar size across population groups, but does differ by grade in which EE is initiated. This may relate to the higher stakes of test results in later grades, given their weight towards the teacher track recommendation.

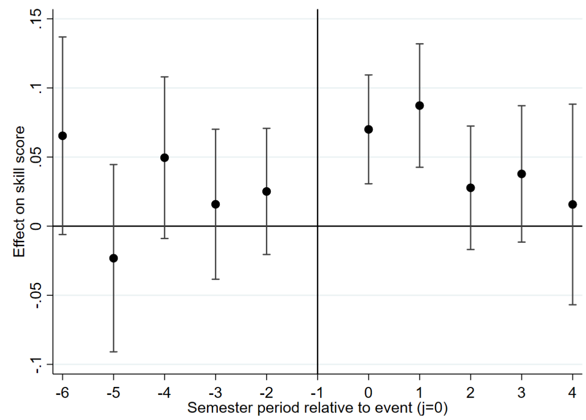
[10] Appendix Figure A7 shows the ESA figure when excluding those that stopped EE somewhere during the post-treatment period. Results are similar.

[11] Ultimately, it is not possible to formally distinguish between an Ashenfelter's dip and a true (albeit small) bounce-back effect. We note, however, that it seems implausible that EE would be responsible for the immediate recovery from the dip, but has no positive effects for later periods in which it is still ongoing.

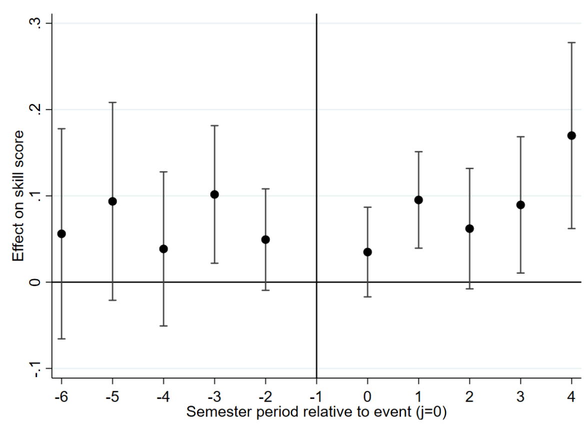
Figure 4: Event study analysis: effects by subject



(a) Arithmetic



(b) Reading



(c) Spelling

Notes: The figure shows the results of estimating Model 1 for the full sample of students, separately for each test subject

point estimates are positive for other forms as well. This suggests that EE and parental support act as complements, and that the previous estimates are an upper bound of the impact of EE alone.

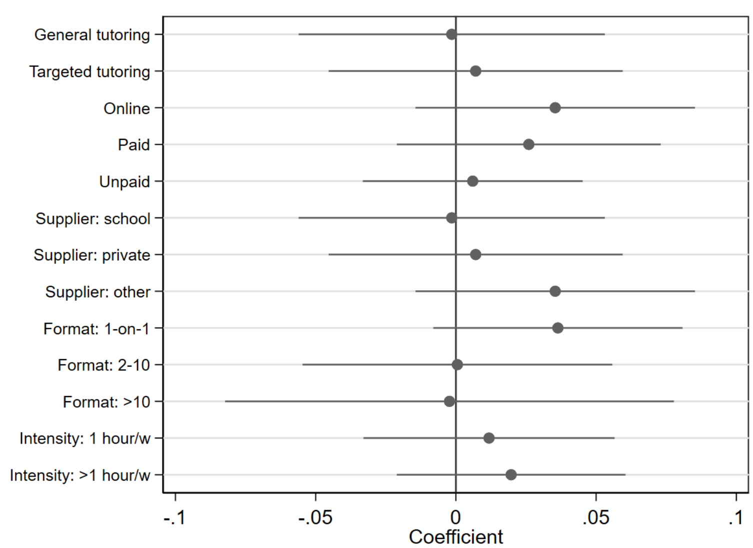
5.2 Effect heterogeneity

The previous results pertain to the effect of the average form of EE. As summary statistics show, the forms of EE differ strongly. Moreover, the type of EE correlates substantially more with SES than use of EE in itself. Thus, the implications for educational inequality mainly depend on whether there are differences in effectiveness by format.

We first examine the difference between privately paid and unpaid (subsidized) forms of EE, as this correlates most strongly with SES. Table 3 shows that effectiveness is low for both forms. While point estimates are consistently higher for the paid form, the magnitude of these differences is low. If we do a back-of-the-envelope calculation using the estimates of paid and unpaid EE from Table 3 and the figures on use by SES from Table 2, we find that EE raises achievement of high SES students by 0.015 and of low SES students by 0.012. The total achievement gap between these groups typically exceeds half a standard deviation. Even if we assume estimates for paid EE at the top of its confidence interval, the contribution of EE to this difference is less than 0.01 of a standard deviation.

Additionally, Figure 5 and Appendix Figures A11 and A12 show that effect heterogeneity is low and point estimates are consistently small across other dimensions of the type of EE, across individual background characteristics, or across timing (of the test and of the use of EE). There is some evidence for slightly higher effectiveness for girls versus boys (Figure A11).

Figure 5: Effects of extended education: effect heterogeneity by format



Notes: The figure shows estimates from the two-way fixed effects model (Model 2) for different forms of extended education, depending on type, supplier, format, intensity, and whether it is paid by parent or not.

5.3 Effect of "high impact tutoring"

The absence of a statistically or economically significant effect of EE is consistent with (the largely correlational findings of) the literature if we consider the typical characteristics of how EE is targeted to primary school students in the Netherlands. Descriptive statistics have shown that the typical form is for around one hour per week, and in small groups. Meta-analyses indicate that only intensive and high-quality programs are effective. Robinson and Loeb (2021) reveal a certain set of criteria for effective programs, which they label as "high-impact" tutoring. These are, among others, maximally three students per tutor and a minimum of three sessions a week. Since we observe group size and instruction time per week in our survey, we can define a more stringent form of "high-impact" extended education and re-estimate effects.[12]

If we do so, the estimates are positive and statistically significant, with a point estimate of around 0.144 of a standard deviation (see Table A2). Moreover, the pattern from the ESA model shows that these effects gradually increase as EE goes on (see Figure 6). The effect is especially large for spelling (0.253). This result is consistent with the notion that our previous null findings are driven by a low average intensity, which is how we can describe the typical EE in the Dutch context.

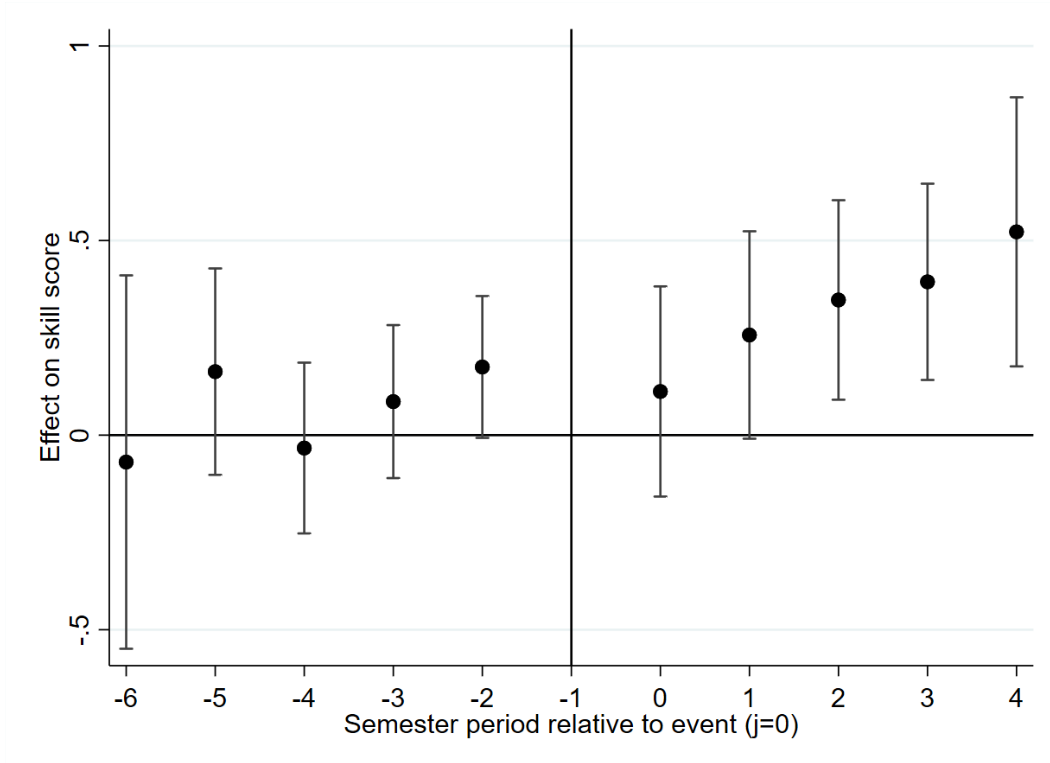
Table A2 also shows results for an alternative definition of tutoring. As stated before, we purposely take a broad definition of EE. Nickow et al. (2024)'s recent overview defines tutoring as one-on-one or small group instruction by teachers, paraprofessionals, volunteers, or parents. In order to provide a comparison to this standard in the literature, we exclude EE when it takes place online or in large groups (>10 students). Estimates for this definition are low and statistically insignificant (point estimate of -0.001).

5.4 Robustness

We finally analyze the sensitivity of our results to potential identification threats and forms of measurement error.

[12] Group size in the survey distinguishes between 1, 2-10 and >10. We take the more restrictive approach by applying the condition of one-on-one tutoring.

Figure 6: Event study analysis: high impact tutoring



Notes: The figure shows the results of estimating Model 1, when redefining the treatment as "High Impact Tutoring" (offline, one-on-one, at least three hours per week).

A first potential issue concerns the choice regarding the timing of the initiation of EE. There are cases with missing/unknown data about the exact start (see Figure 2). This induces measurement error, which may be classical or not. We assess sensitivity by considering alternatives. Note that these alternatives are fairly limited in number, because of the concentrated testing periods (see Figure A3). We impute the missing cases in the main analysis to August, but there is virtually no difference if we take October or December instead, because very few tests are made before January. Alternatively, we can assign April to the missing values, which lies between test period 1 and 2.[13] Another alternative is to exclude all cases with missing timing data. The analyses provide highly consistent estimates across both alternatives (see Figures A4 and A5). Hence, how we treat the missing cases has no effect on the estimated effects.

[13] Another option would be to assign missing values to July, indicating a start of EE after test period 2 (which effectively means changing by one full year). This seems not a realistic scenario given that virtually no non-missing answers are recorded there.

A related issue is that we do not consider EE to be in effect if it starts in the exact month of the administered test. Note that there are relatively few incidences where these coincide given that most EE starts early in the year and most tests are taken later in the year. When we also consider these coinciding timings as events where EE started before the test, results are again virtually identical.

A final related issue with respect to timing is that there may be a general recall error. Again, this is partially mitigated in this setting because most EE starts in September and a recall error of one to three months has no effect on the position with respect to the winter and spring testing periods. Still, some EE starts later in the year and recall error may be more substantial when going back more years. To assess sensitivity, we analyze whether results change when we gradually exclude cases in which EE started before 2017, 2018, etc. Naturally, this gradually reduces the precision, but the point estimates are consistent (see Table A5).

A separate issue is that the observation period for this study overlaps with the Covid-19 pandemic. This may have two relevant implications. For one, the pandemic and the consequent school closures may impact student achievement. Second, parents may have initiated EE in response to Covid-induced achievement gaps. This may imply that both the achievement gaps that EE needs to address and the typical form of EE are different from a more general setting. A benefit of the data setup of this study is that it also contains a large enough pre-Covid period. We can still identify precise estimates when we only rely on observations from 2019 or before. This provides a point estimate of -0.020 with a standard error of 0.025 (see Table A4), which is again qualitativey similar to the main results.

Finally, Appendix Figures A7, A8, A9 and A10 show that results for the event study analysis are consistent when we exclude those who stopped EE at some point after initiation, when we exclude those who have gaps in their testing data (which potentially indicates selective test taking) and when we estimate on a balanced panel.

6 Conclusion

This study has analyzed the efficiency and equality effects of extended education among primary school students in the Netherlands. Around 34% of Dutch students opt for EE during their primary school career. Hence, the potential impacts for educational efficiency are large. Using an event study design, we identify (precisely estimated) low impacts of EE, however. With respect to equity, differences in overall use of EE is small, but SES gaps in the use of paid versus unpaid forms are substantial. Hence, the potential implications of EE for equity depend on the relative effectiveness of these different forms. We find that effectiveness is low for both forms, and therefore conclude that the implications of EE in Dutch primary education are negligible with respect to both efficiency and equality, given how it is currently being used by schools and parents. Moreover, our finding of a modest Ashenfelter dip just before the initiation of EE highlights the potential of overreaction to spuriously poor test results, which may hinder the effective targeting of educational resources. Additional analysis further suggests that concerns that EE substitutes for parental support are unwarranted.

Our results contrast with the high effectiveness of high-impact tutoring programs in the United States, yet can simultaneously be explained by the same literature. Reviews on high-impact tutoring emphasize that effectiveness is only found for intensive programs that fulfill a certain set of criteria: maximally three students per tutor, a minimum of three sessions a week, having one consistent tutor, adjustment with the school curriculum, and data-supported (Robinson and Loeb, 2021). The reality is that the typical EE for Dutch primary school students is not organized with this suggested intensity in mind and so our null-results are consistent with previous more pessimistic results regarding low intensity EE (Heinrich et al., 2010).

We confirm the positive effects of high-impact tutoring in our sample, although these findings are based on a very small subgroup that engages in such a form of tutoring. More research on high-impact tutoring in this context is needed to see whether these results hold in larger samples, and whether effects are large enough to offset their higher costs as well. Nonetheless, it has potential to enhance educational efficiency among Dutch primary students. On the other hand, it has been recognized that effective educational interventions such as intensive tutoring are typically difficult to scale and therefore may not translate to meaningful higher educational achievement or equality (Cullen et al., 2013; Bhatt et al., 2024). Additionally, such a potential move toward HIT has the potential to induce stronger efficiency/equality trade-offs. Schools may be more constrained to (consistently) provide such a service compared to high-SES parents. Moreover, scaling HIT services such that they can reach a larger part of the student population can be expected to further boost the private tutoring market and increase prices for tutoring. Policy-makers need to consider these different trade-offs when deciding on whether to expand access to (intensive) forms of tutoring. A potential way to scale HIT would be via online delivery, and its positive effects have been alluded to by Gortazar et al. (2024).

Back-of-the-envelope calculations show that the contribution of EE to educational inequality in Dutch primary education is negligible, even when we assume that the impact of paid forms is at the top of its confidence interval. Given its low effectiveness, this would still apply if subsidies for unpaid forms would substantially decrease (e.g. through budgetary pressure). This conclusion would only change if high SES parents would strongly expand their use of more intensive forms of tutoring. An interesting question for future research is whether the low use of HIT at the moment is due to low willingness to pay, or lack of supply. More generally, monitoring the contribution of EE for educational inequality predominantly requires monitoring the use of highly intensive forms of tutoring, and how this use differs by parental background.

Finally, we emphasize that our study analyzes the effects of EE exclusively towards test achievement. As also highlighted by Gortazar et al. (2024), EE could positively influence educational aspirations. Such potential effects could, in turn, impact measures of educational achievement such as dropout rates. Similarly, we have not assessed to what extent EE could have negative side-effects, such as less informal or outside play and less participation in sports activities, which can have negative health effects. Hence, EE also entails opportunity costs and these have to be taken into account as well to assess its social welfare effects.

References

Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *The Review of Economics and Statistics 60*(1), 47–57.

Bhatt, M. P., J. Guryan, S. A. Khan, M. LaForest-Tucker, and B. Mishra (2024). Can technology facilitate scale? Evidence from a randomized evaluation of high dosage tutoring. Technical report, National Bureau of Economic Research.

Bray, M. (2021). Shadow education in europe: Growing prevalence, underlying forces, and policy implications. *ECNU Review of Education 4*(3), 442–475.

Bray, T. (1999). *The shadow education system: Private tutoring and its implications for planners*. UNESCO International Institute for Educational Planning.

Byun, S.-y., H. J. Chung, and D. P. Baker (2018). Global patterns of the use of shadow education: Student, family, and national influences. In *Research in the Sociology of Education*, Volume 20, pp. 71–105. Emerald Publishing Limited.

Carneiro, P. and J. J. Heckman (2002). The evidence on credit constraints in post-secondary schooling. *The Economic Journal 112*(482), 705–734.

Clarke, D. and K. Tapia-Schythe (2021). Implementing the panel event study. *The Stata Journal 21*(4), 853–884.

Cullen, J. B., S. D. Levitt, E. Robertson, and S. Sadoff (2013). What can be done to improve struggling high schools? *Journal of Economic Perspectives 27*(2), 133–152.

De Donder, P. and F. Martinez-Mora (2017). The political economy of higher education admission standards and participation gap. *Journal of Public Economics 154*, 1–9.

Dietrichson, J., M. Bøg, T. Filges, and A.-M. Klint Jørgensen (2017). Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Review of Educational Research 87*(2), 243–282.

Fortune Business Insights (2021). Private tutoring market size, share and covid-19 impact analysis, by subjects (academic and non-academic), application (up-to k12 and post k-12), mode (offline and online, and regional forecast 2021-2028). `https://www.fortunebusinessinsights.com/private-tutoring-market-104753`. Accessed: 2023-02-15.

Fryer Jr, R. G. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In *Handbook of economic field experiments*, Volume 2, pp. 95–322. Elsevier.

Gortazar, L., C. Hupkau, and A. Roldán-Monés (2024). Online tutoring works: Experimental evidence from a program with vulnerable children. *Journal of Public Economics 232*, 105082.

Guryan, J., J. Ludwig, M. P. Bhatt, P. J. Cook, J. M. Davis, K. Dodge, G. Farkas, R. G. Fryer Jr, S. Mayer, H. Pollack, et al. (2023). Not too late: Improving academic outcomes among adolescents. *American Economic Review 113*(3), 738–765.

Haelermans, C., T. Huijgen, M. Jacobs, M. Levels, R. van der Velden, L. van Vugt, and S. van Wetten (2020). Using data to advance educational research, policy, and practice: Design, content, and research potential of the Netherlands Cohort Study on Education. *European Sociological Review 36*(4), 643–662.

Heinrich, C. J., R. H. Meyer, and G. Whitten (2010). Supplemental education services under no child left behind: Who signs up, and what do they gain? *Educational Evaluation and Policy Analysis 32*(2), 273–298.

Ku, I., H. Lee, and J.-E. Kim (2022). Does shadow education contribute to inequality? *Asia Pacific Journal of Education*, 1–19.

Luo, J. and C. K. Y. Chan (2022). Influences of shadow education on the ecology of education–A review of the literature. *Educational Research Review 36*, 100450.

Nickow, A., P. Oreopoulos, and V. Quan (2024). The promise of tutoring for PreK–12 learning: A systematic review and meta-analysis of the experimental evidence. *American Educational Research Journal 61*(1), 74–107.

OECD (2016). *PISA 2015 Results (Volume II): Policies and Practices for Successful Schools*. OECD Publishing: Paris.

Robinson, C. D. and S. Loeb (2021). High-impact tutoring: State of the research and priorities for future learning. *National Student Support Accelerator 21*(284), 1–53.

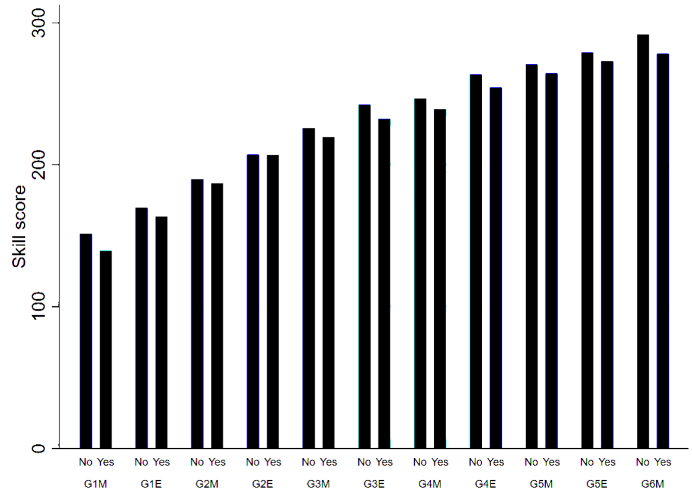
Ryu, D. and C. Kang (2013). Do private tutoring expenditures raise academic performance? Evidence from middle school students in s outh k orea. *Asian Economic Journal 27*(1), 59–83.

Statistics Netherlands (2021). Uitgaven van huishoudens aan onderwijsondersteuning. `https://www.cbs.nl/nl-nl/maatwerk/2021/10/uitgaven-van-huishoudens-aan-onderwijsondersteuning`. Accessed: 2023-02-15.

Zumbuehl, M., S. Hof, and S. C. Wolter (2024). Private tutoring and academic achievement in a selective education system. *Education Economics*, 1–18.

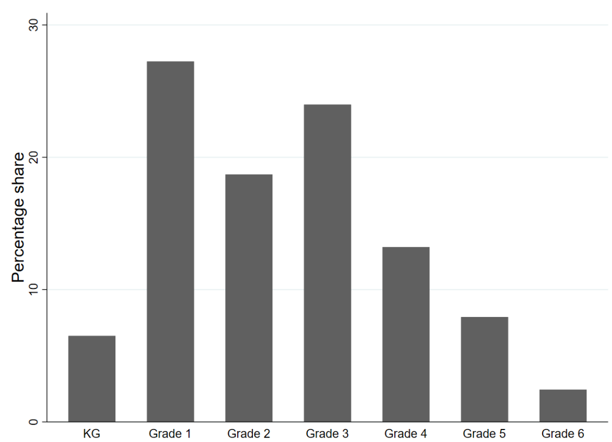
Appendix (for online publication)

Figure A1: Skill scores by grade and extended education status (use in current grade)



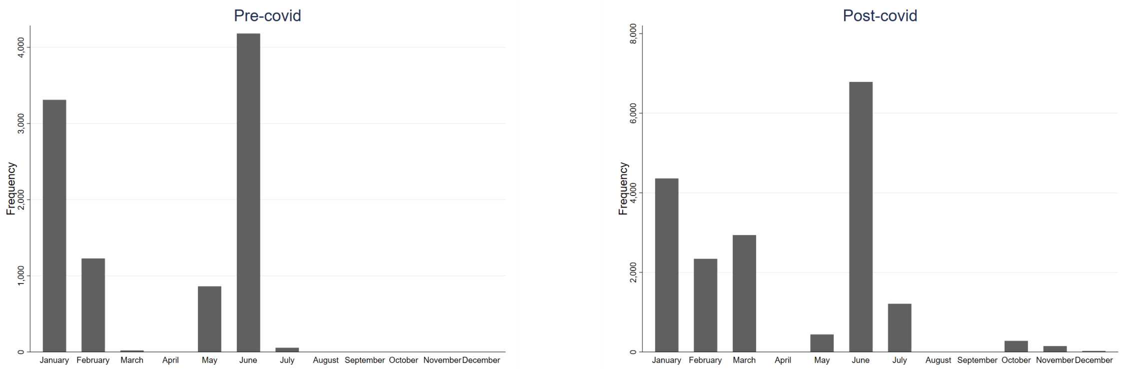
Notes: The figure shows skill scores by grade, for students that used extended education at the moment of testing and students that did not. Skill sores are averaged across subjects and testing periods (winter or spring) for each grade.

Figure A2: Timing of start of extended education by grade



Notes: The figure shows the distribution of grades in which extended education started.

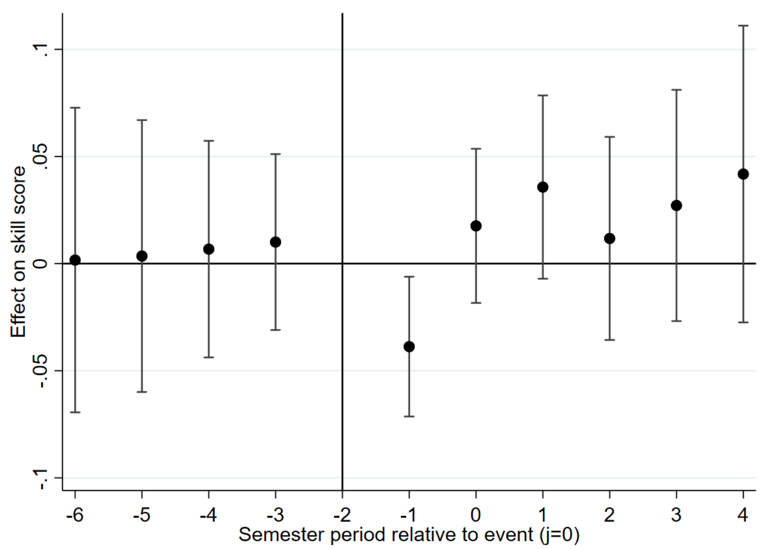
Figure A3: Distribution of testing months



(a) Pre-covid (b) Post-covid

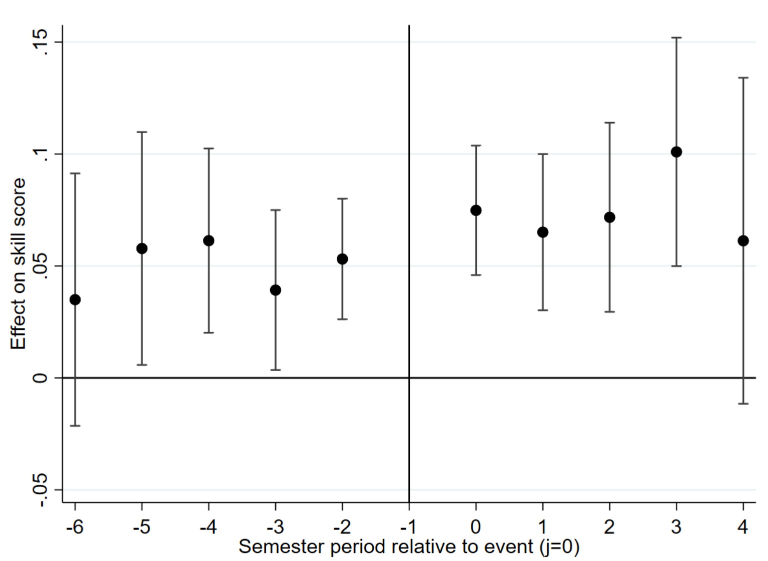
Notes: The figure shows the distribution of months in which tests are administered, separately in the period from 2017 until 2019 (left-hand panel; "Pre-covid") and from 2020 until 2022 (right-hand panel; "Post-covid").

Figure A4: Event study analysis: alternative baseline



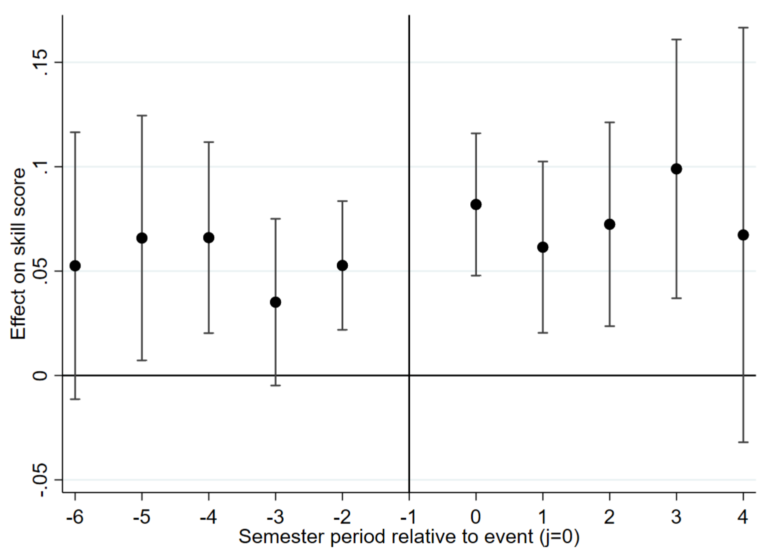
Notes: The figure shows the results of estimating Model 1 for the full sample of students, when putting the baseline period at -2.

Figure A5: Event study analysis: recoding unknown starting months



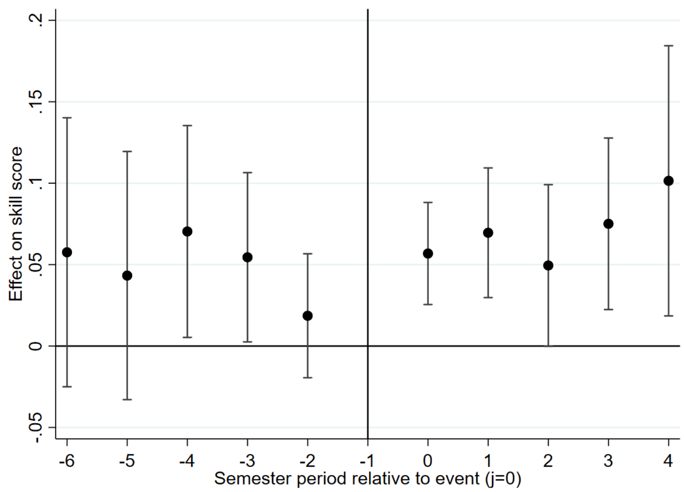
Notes: The figure shows the results of estimating Model 1 for the full sample of students, when recoding missing starting months to January rather than August.

Figure A6: Event study analysis: excluding unknown starting months



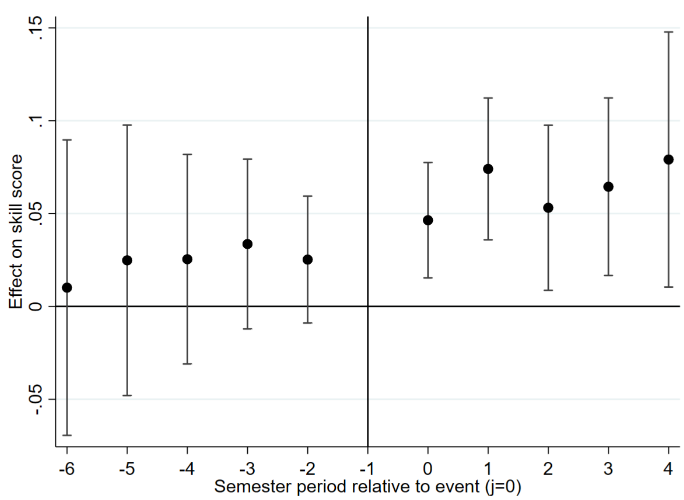
Notes: The figure shows the results of estimating Model 1 for the full sample of students, when excluding those with missing starting months.

Figure A7: Event study analysis: excluding those who stopped extended education



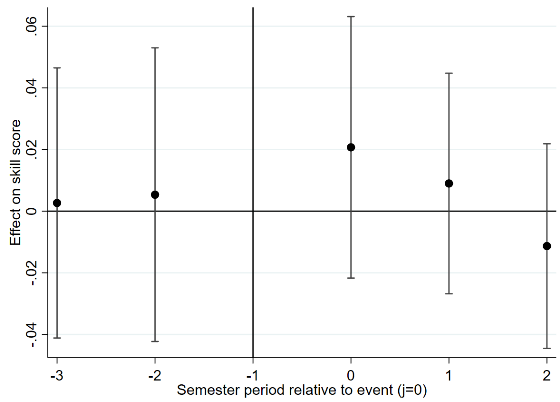
Notes: The figure shows the results of estimating Model 1 and using all tests, when excluding those that reported stopping extended education somewhere within the post-treatment window.

Figure A8: Event study analysis: excluding those with gap in testing history



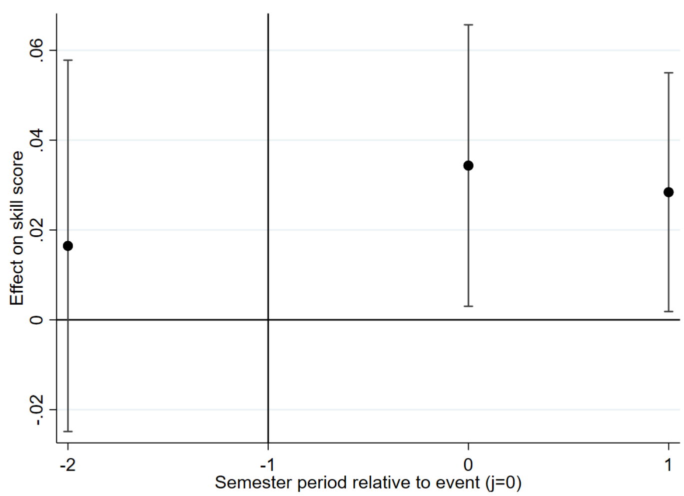
Notes: The figure shows the results of estimating Model 1 using all tests, when excluding those that have gaps in their testing data (this excludes 139 individuals).

Figure A9: Event study analysis: balanced panel (three periods pre and post)



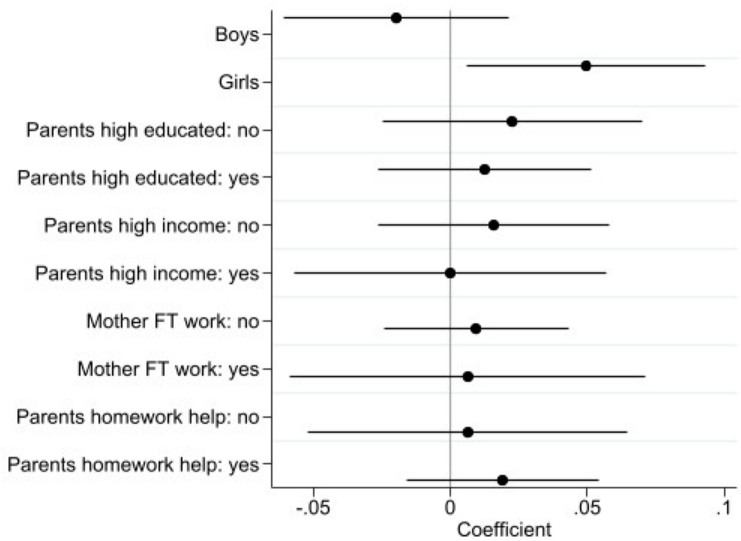
Notes: The figure shows the results of estimating Model 1 when we only include individuals for which there are at least three periods before and after treatment (applies to 158 treated students).

Figure A10: Event study analysis:balanced panel (two periods pre and post)



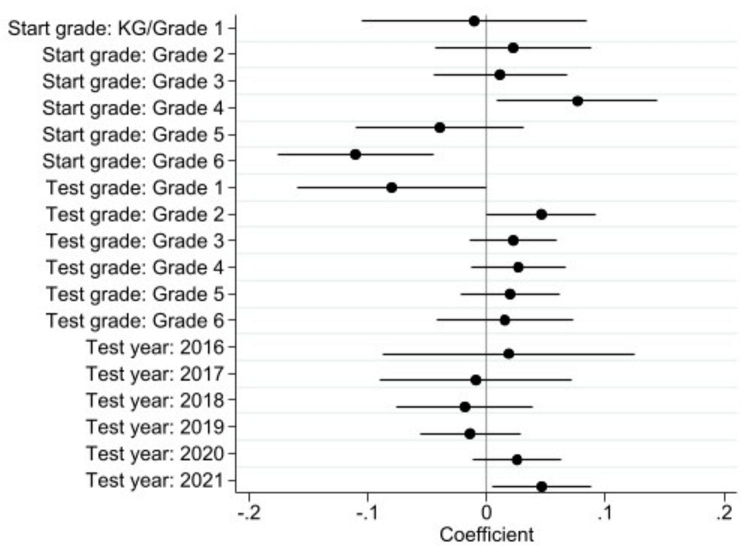
Notes: The figure shows the results of estimating Model 1 when we only include individuals for which there are at least two periods before and after treatment (applies to 273 treated students).

Figure A11: Effects of extended education: effect heterogeneity by background



Notes: The figure shows estimates from the two-way fixed effects model (Model 2) for extended education, by specific subgroups in the student population

Figure A12: Effects of extended education: effect heterogeneity by timing



Notes: The figure shows estimates from the two-way fixed effects model (Model 2) for extended education, depending on the timing of the start grade of extended education, the test grade or the test year.

Table A1: Coefficients from the Event Study Design

	All	Reading	Spelling	Arithmetic
L.6	0.045	0.065*	0.056	0.044
	(0.036)	(0.036)	(0.062)	(0.037)
L.5	0.036	-0.023	0.094	0.020
	(0.032)	(0.035)	(0.058)	(0.032)
L.4	0.049*	0.049	0.039	0.076***
	(0.026)	(0.030)	(0.046)	(0.029)
L.3	0.051**	0.016	0.102**	0.033
	(0.021)	(0.028)	(0.041)	(0.024)
L.2	0.040**	0.025	0.049*	0.038*
	(0.017)	(0.023)	(0.030)	(0.020)
F.0	0.058***	0.070***	0.035	0.051***
	(0.014)	(0.020)	(0.027)	(0.018)
F.1	0.075***	0.087***	0.095***	0.048**
	(0.017)	(0.023)	(0.029)	(0.019)
F.2	0.052**	0.028	0.062*	0.045**
	(0.020)	(0.023)	(0.036)	(0.023)
F.3	0.067***	0.038	0.090**	0.051**
	(0.023)	(0.025)	(0.040)	(0.026)
F.4	0.085***	0.016	0.170***	0.041
	(0.031)	(0.037)	(0.055)	(0.030)
N	28,159	7,641	10,021	10,321
Joint sig lags	0.174	0.071	0.108	0.145
Joint sig leads	0.000	0.008	0.005	0.049
Joint sig all	0.003	0.003	0.008	0.113
Joint sig lags: no T=-1	0.979	0.101	0.485	0.342
Joint sig leads: no T=-1	0.397	0.036	0.043	0.922
Joint sig all: no T=-1	0.808	0.012	0.094	0.725

Notes: The table shows the different lags (L) and leads (F) estimated within the event study design (estimates as portrayed in Figure 3). Standard errors are clustered at the student level.
$*p < 0.10, **p < 0.05, ***p < 0.01$

Table A2: Alternative definitions of tutoring

	All	Reading	Spelling	Arithmetic
Tutoring	-0.001	0.061**	-0.027	-0.048*
(Nickow et al., 2024)	(0.028)	(0.030)	(0.050)	(0.025)
High impact tutoring	0.144***	0.119**	0.253**	0.074
(Robinson and Loeb, 2021)	(0.046)	(0.055)	(0.127)	(0.062)

Notes: The table shows results for more exclusive definitions of extended education. "Tutoring" considers only offline education in groups under 10. "High impact tutoring" only considers one-on-one education, and at least three hours per week. Results for fixed effect regression. Standard errors are clustered at the student level.
$*p < 0.10, **p < 0.05, ***p < 0.01$

Table A3: Heterogeneity in Ashenfelter's dip

	Female	High educ	High inc.	Married	Hw help	Work hrs	Start grade
Baseline	-0.063***	-0.026	-0.054***	-0.081***	-0.071***	-0.045	-0.040**
	(0.020)	(0.023)	(0.020)	(0.027)	(0.018)	(0.034)	(0.017)
Interaction	0.029	-0.033	0.002	0.050*	0.016**	0.000	-0.002***
	(0.025)	(0.026)	(0.031)	(0.029)	(0.006)	(0.001)	(0.001)

Notes: The table shows estimates of a dummy variable for T=-1, and its interactions with student characteristics. Results in each column are from separate regressions. Standard errors are clustered at the student level.
$*p < 0.10, **p < 0.05, ***p < 0.01$

Table A4: Effects of EE on achievement: pre-covid

	All	Reading	Spelling	Arithmetic
EE effect	0.020	0.036	0.028	0.001
	(0.023)	(0.027)	(0.044)	(0.025)

Notes: The table shows estimated treatment effects of extended education on student achievement, using only tests made before March 2020. Standard errors are clustered at the student level.
$*p < 0.10, **p < 0.05, ***p < 0.01$

Table A5: Assessing recall error

	All	<=4 yrs	<=3 yrs	<=2 yrs	<=1 yrs
EE effect	0.015	0.015	0.017	0.016	0.021
	(0.015)	(0.015)	(0.016)	(0.018)	(0.021)

Notes: The table shows results for the main analysis, and for when we exclude cases in which the start of EE was more than 4 years, more than 3 years, more than two years or more than 1 year ago. Results for fixed effect regression. Standard errors are clustered at the student level.
$*p < 0.10, **p < 0.05, ***p < 0.01$

Table A6: Effect of EE on parental support

	All types	General tutoring	Targeted tutoring	Online
EE effect	0.248***	0.225	0.122	0.344***
	(0.090)	(0.180)	(0.150)	(0.128)

Notes: The table shows estimated effects of the start of extended education on parental help with homework (measured in hours per week). Estimations are limited to those students that started EE within the last year of the survey, and control for homework support in the previous year. Standard errors are clustered at the student level.
$*p < 0.10, **p < 0.05, ***p < 0.01$

Table A7: Text analysis on motivation for Extended Education

Tutoring	
To address achievement gaps	52% (17/33)
To boost confidence	12% (4/33)
To compensate for lack of school quality	9% (3/33)
Other reasons (e.g. maximize potential)	27% (9/33)
Specific learning needs	
Dyslexia	23% (23/101)
Giftedness	19% (19/101)
Child lags behind in certain school subject	15% (15/101)
Child needs extra challenges	14% (14/101)
To maximize child's potential	13% (13/101)
Other reasons (e.g. problems with concentration or confidence)	17% (17/101)
Online education	
To address achievement gaps	43% (36/84)
To provide an extra challenge	20% (17/84)
To maximize child's potential	15% (13/84)
Child enjoys learning with educational software	13% (11/84)
Other reasons	8% (7/84)

Notes: A randomly selected 50% of those who had indicated that they ever used extended education for their child were asked to provide their strongest motivation for its use (open question).