TI 2019-053/V
Tinbergen Institute Discussion Paper

Do early-ending conditional cash transfer programs crowd out school enrollment?

Martin Wiegand[1]

[1] Vrije Universiteit Amsterdam

Do early-ending conditional cash transfer programs crowd out school enrollment?

Martin Wiegand*

18.07.2019

Abstract

This paper explores how a conditional cash transfer program influences students' schooling decisions when program payments stop in the middle of the school career. To that end, I examine Mexico's PROGRESA, which covered students only until the end of middle school (at age 15) in its early years. The experimental setup permits to study the program's impact on the probability to continue with high school after middle school. Despite initial randomization, the program itself has likely rendered the respective samples of middle school graduates in the treatment and the control group incomparable. To account for this, I employ a newly developed semiparametric technique that uses a combination of machine learning methods in conjunction with doubly-robust estimation. I find that exposure to PROGRESA during middle school reduced the probability to transfer to high school by 10 to 14 percentage points. Possible explanations for this effect include parents' loss aversion, motivation crowding, anchoring, and classroom peer effects.

JEL Classication: I22, I25, O15, J24, D04, D91, C52.

Keywords: education, conditional cash transfer, PROGRESA, machine learning, doubly-robust estimation, loss aversion, motivation crowding, anchoring, classroom peer effects, Mexico.

*Vrije Universiteit Amsterdam, Netherlands, m.wiegand@vu.nl.

1 Introduction

Around 20 years after their first appearance, conditional cash transfer (CCT) programs for education—initiatives that provide financial incentives for poor households to send their children to school—have never been more popular. Praised for their potential to increase school enrollment while reducing poverty, they are now widespread in Latin America and gain quick traction in Africa and Asia. However, a lot of programs do not cover the entirety of a student's school-days and instead stop with welfare payments after elementary school or middle school. Even the largest and most well-known CCT programs only started to cover high school students long after their introduction.[1]

In this paper, I investigate how CCT programs can affect school enrollment after payments break off, using Mexico's PROGRESA as a case study. PROGRESA is particularly well suited for this research for a number of reasons. One is its program design: for the first five years, the program covered students only until the end of middle school (approximately at age 15). Another reason is the richness of the data collected for evaluation, as well as the experimental setup, which was achieved by the deference of the program in some randomly chosen localities serving as the control group. These features make it possible to estimate the effect of PROGRESA on high school enrollment. I focus on two quantities of interest. The first one is the overall treatment effect of PROGRESA on going to high school. It measures the difference in probability of high school enrollment between treatment and control group, considering all adolescents who had finished primary school just before the program started (around the age of 12). The resulting number is easy to interpret, but may largely be driven by the program impact on middle school enrollment. The second quantity of interest is the effect on the probability to make the transition from middle school to high school. For this, I consider all those middle school graduates who would have completed middle school even in the absence of the program. I then compute how their likelihood of continuing to high school differs depending on having been exposed to the program or not. This conditional treatment effect is a more direct measure of program aftereffects, and thus receives the main focus

[1] This includes the four largest CCTs (in terms of beneficiaries) at the time of writing, namely Brazil's BOLSA FAMILIA, Mexico's PROSPERA (formerly known as PROGRESA), the Philippines' PANTAWID, and Colombia's FAMILIAS EN ACCIÓN. Further examples of currently running CCTs that do not cover high school are Indonesia's PROGRAM KELUARGA HARAPAN, Ghana's LIVELIHOOD EMPOWERMENT AGAINST POVERTY program, Pakistan's PUNJAB FEMALE SCHOOL STIPEND PROGRAMME, Nigeria's NATIONAL CASH TRANSFER PROGRAMME, and Burkina Faso's NAHOURI CASH TRANSFERS PILOT PROJECT.

of this paper.

There are many studies showing that CCT programs increase school enrollment while payments are in place (see review by Bastagli et al. 2016), but to my knowledge there is none about the aftereffects on enrollment. The direction of these effects is not obvious: neoclassical theory suggests that past payments free up resources, rendering continued schooling more likely. On the other hand, a number of theories from psychology and behavioral economics, such as loss aversion, motivation crowding, anchoring, and classroom peer effects, could explain why the probability to continue school might actually decrease due to earlier payments. Studying the aftereffects of CCT programs is necessary to understand their full impact on the education distribution, and is highly relevant for the design of future programs: a policy maker with limited funds needs to worry less about early break-offs if CCT programs continue to have a positive effect on enrollment. If, on the other hand, it turns out that such programs actively discourage students from continued education, they may pose a trade-off between different levels of secondary education, and raise the question how such discouragement may be averted. The study may also prove interesting beyond its policy relevance, shedding light on the interaction of financial incentives and the behavior and social norms revolving around education.

The contribution of this paper is twofold. First, it shows that having been paid for schooling in the past may reduce the probability to continue once the payments stop. It appears that paying students up to a point actively discourages them to stay in school afterwards. This is likely caused by a shift in the perceived value of education. The negative effect does not spill over to adolescents from non-poor households in treatment locations, who were not eligible for the program. Conversely, for these adolescents the program may even have had a positive high school enrollment effect.

The second contribution is to demonstrate a way to estimate program aftereffects conditional on relevant pre-treatment characteristics. Despite the random selection of PROGRESA communities into treatment and control group, education payments have likely changed the composition of middle school graduates between these groups after two years. Some students in the treatment group may not have finished middle school in the absence of payments. So, to estimate the program effect on the probability to transfer to high school, conditional on having made it through middle school, a simple comparison of means is not sufficient. It is worth noting that this resulting imbalance between the treatment and control group is an inherent property of the evaluation of conditional aftereffects. There is no ideal experiment that might serve as a benchmark. Nonetheless, aftereffects may be interesting in a number of contexts, e.g. when studying withdrawal in medical trials or job search after a time-limited unemployment benefit program ends. The approach taken in this paper is to make an assumption of unconfounded treatment, conditional on a large number of baseline characteristics. This is done to correct not for selection into treatment and control group (which are randomized), but for the decision to drop out as a result of group membership.

To obtain causal estimates, I employ a newly developed procedure to estimate treatment effects, called double machine learning (DML), by Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018). It is a doubly-robust estimation technique (see e.g. Bang and Robins 2005), i.e., it makes use of predictions of both the propensity score and outcomes, and is robust to misspecifications of either of them. A variety of machine learning methods as well as sample-splitting are used to learn and predict the relationships of treatment status and outcomes with potential confounders. DML permits to capture complex functional relationships and to control for large sets of covariates—possibly containing more elements than there are observations—without having to know either in advance. This approach relies on far fewer assumptions than conventional propensity score methods, and makes results more credible in comparison. Beside correcting for imbalances between treatment and control group, the method can also correct for imbalances resulting from attrition. This extension of the DML method, using the assumption that outcomes are missing at random, is demonstrated in this paper.

The paper proceeds as follows. In section 2, I summarize the related literature and offer a number of explanations for aftereffects of CCT programs on school enrollment. Section 3 sums up the relevant details about the program and the data used. In section 4, I explain the identification strategy. In section 5, the estimation procedure including the machine learning methods used is being explained and results are presented. The findings are further expanded on in section 6, with an analysis of possible channels and a discussion of covariate stability. Section 7 concludes.

2 Literature

This paper contributes to the literature on CCTs and school enrollment. For Progresa, it has been shown that adolescents of treatment families stay in school with higher probability than those in the control group, and that this effect is particularly pronounced for those in the age for middle school (Schultz 2004, Behrman et al. 2005, and Behrman et al. 2009). Todd and Wolpin (2006) and Attanasio et al. (2012) develop and estimate structural models of decision making about schooling, enabling them to evaluate the costs and benefits of alternative program specifications to the ones of Progresa. Both studies conclude that a shift of program resources from primary school age children to middle school age adolescents would have led to a higher increase of total completed years of schooling. Perhaps closest to my paper, Behrman et al. (2005) and Dubois et al. (2012) estimate students' transition probabilities of Progresa students from one grade to another. The study by Behrman et al. (2005) is the only one that considers the probability to go to high school, however without controlling for compositional changes as a result of prior program exposure, and only for small subsets of the data. Dubois et al. (2012) aim to disentangle the effects of Progresa on grade repetition and continuation. They circumvent the selection problem by looking only at the first year of implementation, and find among other things that middle school students in the treatment group are more likely to repeat a grade. The authors speculate that this may reflect the incentive to stay in middle school longer due to the limited program coverage.

To my knowledge, no study so far has concentrated on the schooling impacts of Progresa—or any other CCT program—after payments stop. On the face of it, it may not be obvious why the decision to continue school should depend on having been paid to go to school before. In the remainder of this section, I offer a number of explanations of how this may come about, each one supported by theoretical or empirical literature.

Easing financial constraints: One of the goals of CCT programs is helping poor households to finance children's education. If financial constraints are in fact the main driver of educational underinvestment, then easing these constraints by making cash payments should result in more schooling. For instance, a family may have saved just enough to allow their child to finish middle school. Giving transfer payments until that point may then enable the household to save more, which in turn might allow the child to go to high school. The study by de Janvry et al. (2006) supports the argument that CCT payments can help smoothing out spending on education. It finds that PROGRESA takes a safety net function, in that it protects children from the impacts of shocks on school enrollment. The smoothing of education spending may not only work intertemporally within households, but also between households. Angelucci et al. (2010) show that PROGRESA raises middle school enrollment only for children with large family networks, in which transfer payments go from better-off to worse-off family members to ensure their children's school enrollment.

While CCT programs allow to save more money for future education and thus might facilitate high school enrollment, a number of insights from psychology and behavioral economics point in the opposite direction. In the following, I highlight four reasons why CCT programs could discourage further schooling after they end: loss aversion, motivation crowding, anchoring, and classroom peer effects.

Loss aversion: Loss aversion is a central feature of prospect theory (Kahneman and Tversky 1979, Tversky and Kahneman 1991). It means that from a psychological point of reference, losses loom larger than gains of equal size. For intertemporal choice problems, this means that people require a larger payment to postpone present consumption than the amount they are willing to pay to have future consumption now (Loewenstein 1988). In the context of CCT programs, the choice between working (more consumption now) and continuing school (more consumption later) may depend on whether going to school is framed as a loss or a forgone gain of current consumption. Since reference points are often derived from past levels of consumption, families who have received PROGRESA payments may frame the choice as having either less or the same current consumption as before. On the other hand, families who never received PROGRESA payments may perceive the choice as having either the same or higher levels of consumption, making them more likely to choose more education. An early study documenting similar behavior is Weiss et al. (1980), on the effect of the Seattle-Denver income maintenance experiment on education. The authors find that reducing the (relatively low) direct costs of schooling, by offering subsidies on schooling expenditure, led to a large increase in enrollment among young adults. At the same time, significantly reducing the opportunity cost of going to school, by increasing the income tax rate for low incomes, had no such effect. This finding is consistent with loss aversion, if the direct costs are perceived as losses while the opportunity costs are viewed as foregone gains (Thaler 1980).

Motivation crowding: Another possibly relevant theory from behavioral economics is motivation crowding theory (Frey and Jegen 2001, Fehr and Falk 2002). It acknowledges that people's actions are often motivated by hope for social approval, a desire to be moral, or intrinsic interest. When monetary incentives are added, they can replace those motives. A famous example is given by Gneezy and Rustichini (2000), who show that introducing a fine for parents who are late to fetch their children from kindergarten makes them arrive even later. The explanation is that being late, which used to be the violation of an ethical norm before, is being reframed into a good that can be bought for a reasonable price. Importantly, removing the fine did not make the parents arrive earlier again. In the same way, PROGRESA may put a price tag on the moral obligation to let children go to school. In distinction to the experiment by Gneezy and Rustichini, the price of non-conformance is high enough to comply with the program. But quantifying the value of sending children to school may reduce the pressure to let them continue after the payments stop. The crowding out effect may also spill over from parents to students, who might view school as necessary labor rather than an opportunity to learn. The negative effect tangible rewards can have on students' intrinsic motivation to learn has been demonstrated in a number of psychological studies (see Deci et al. 1999 for a meta-analysis).

Anchoring: If financial constraints were in fact the only reason for educational underinvestment, there would be no reason to make transfer payments conditional on school attendance. Instead, an unconditional cash transfer could achieve the same result without the need to monitor compliance, and freed from the often raised criticism that CCT programs are paternalistic. One reason for conditionality is that children as well as parents may be poorly informed about the returns to education, or about the natural talent required to complete school (Fiszbein and Schady 2010). For instance, Nguyen (2008) shows that households in Madagascar lack information about returns to education but change decisions rationally when this information is updated. Jensen (2010) shows that eight-graders in the Dominican Republic massively underestimate the rate of return to secondary school. And Dizon-Ross (2018) finds that parents in Malawi hold inaccurate beliefs about their children's ability, the more so when they have low education themselves, and that they misallocate resources to education accordingly. In this light, making cash transfer programs conditional is a way to nudge students into a higher level of educational attainment, thus overcoming not only financial obstacles but also bad decisions due to incomplete information. But in doing so, CCT programs also convey a signal about the value of education: if the government is willing to pay for it, it must be worth pursuing. Conversely, the drop in payments after middle school may suggest that subsidizing poor students to go to high school is not worth it—be it due to low marginal returns to schooling at this level, or because students from poor families are deemed unlikely to succeed there. This particular form of priming effect where a numerical reference point (the program payout) affects the assessment of an unknown value (the value of going to high school) is called anchoring (Tversky and Kahneman 1974). By first anchoring the value of schooling to the Progresa payouts and then reducing it to zero, the government could unintentionally make further education appear less desirable.

Classroom peer effects: If a CCT program works as intended, some students keep attending school who would not have done so in the absence of the program. Presumably, this leads to larger class sizes and a higher share of disadvantaged students, which may affect motivation and learning of the students who would have gone to school without the program. For instance, the literature on classroom peer effects suggests that higher shares of disadvantaged students lead to more misbehavior in class, lower teaching quality, and negative performance spillovers (e.g. Carrell and Hoekstra 2010, Lavy et al. 2012). Thus, by the end of middle school, some students may have lost their motivation or aptitude to continue with high school. However, the changes in the composition of students may well have heterogeneous effects on students with different background or ability, and may even increase motivation for some students (say, through increased competition for grades or out of a need to distinguish themselves from the disadvantaged students). Thus, classroom peer effects may affect high school enrollment in both directions.

All these channels are possible explanations of direct program effects. Most CCT programs only target the poorest households, but they may affect students from other households nonetheless. For the case of Progresa, such spillover effects have been documented: Angelucci and De Giorgi (2009) argue that due to inter-household risk-sharing, food consumption increases even for non-eligible households in Progresa treatment villages. And Attanasio et al. (2012) find, for their sample of boys between 10 and 16, that school enrollment was substantially higher for the non-eligible adolescents in the treatment group than for those in the control group. If households within a village share program resources, one would expect that spillover effects on high school enrollment take the same direction as for the eligible students. If enrollment rises as a result of increased savings, this effect might well spread across household networks to non-eligible students. If the program has aftereffects through a change in the composition of the classroom, or a change of social norms towards going to school, it seems most intuitive to expect that these channels would move most students—eligible or not—in the same direction.

3 Program and Data Description

Progresa is a multi-component antipoverty program that was started in 1997. Its original goal was to improve prior antipoverty programs in Mexico along a number of dimensions, such as increasing targeting efficiency and reducing administrative costs (Gantner 2009). At first, a limited number of rural localities were selected for inclusion. Localities had to have between 50 and 2,500 inhabitants, access to health and education services, and had to be considered highly deprived based on available census data. Households from the selected localities were then classified as poor or not poor based on baseline survey data. Only households classified as poor were eligible.

One declared target was to increase school attendance of adolescents from poor families. The education component of Progresa included bimonthly cash transfers to mothers of every child enrolled in grades 3 to 9 who attended at least 85 percent of classes. This includes the last four years of primary school (*primaria*) and all of middle school (*secondaria*) but not high school (*preparatoria* or *bachillerato*). Payments increased with the age of the child to adjust to the increasing opportunity costs of schooling due to higher child wages. However, according to Schultz (2004), these payments were still lower than the average value of full-time child labor. In 2001, the program was renamed Oportunidades and extended to urban areas, and the schooling grants were extended to include the high school level.

For evaluation purposes, localities were randomized into a treatment and a control group. Payments for eligible households in the treatment group started in May 1998, while payments for eligible households in the control group only started in December 1999. The evaluation sample includes 320 treatment localities and 186 control localities. Survey data was collected biannually for all households of the evaluation set from 1997 to 2000. Two surveys were administered before the program started in the treatment group (in October 1997 and March 1998), three between the start of the program for the treatment and the

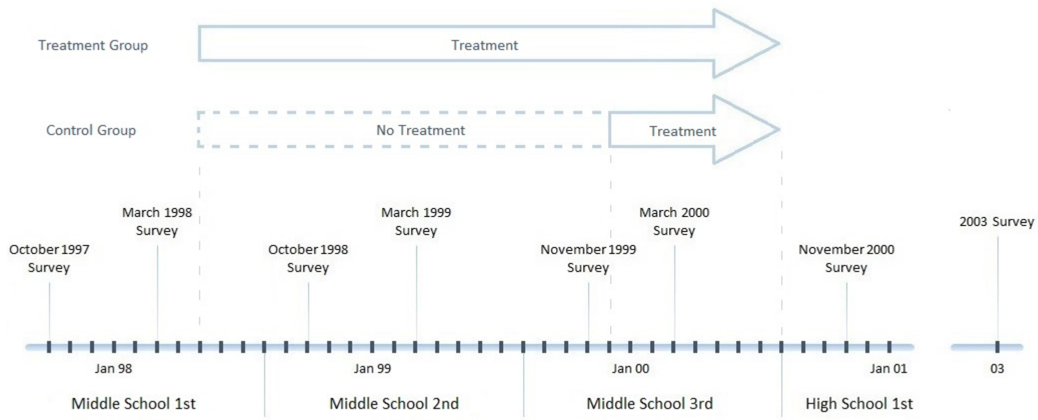

Figure 1: Timeline of relevant events around PROGRESA

control group (in October 1998, March 1999, and November 1999), and three after the program had started for the control group (in March 2000, November 2000, and winter 2003).

To identify the effect that PROGRESA had on the transition to high school, I consider those students who were expected to start high school in the academic year of 2000/01. By the end of the term in July 2000, the eligible students in the treatment group had benefited from PROGRESA for more than two years. Those in the control group had only been exposed to the program for the last semester of middle school, when the decision to continue school afterwards had likely been made already.[2] Figure 1 depicts a timeline with all survey dates. The two arrows indicate how long the students from the cohort under discussion were exposed to the program, for the treatment and the control group, respectively.

I consider three outcome variables related to high school attendance. The first is whether the student went to high school at the time of the November 2000 survey, i.e., right after finishing middle school. The second variable is whether the student had ever been to high school by the time of the 2003 survey. The third variable is whether the student had completed high school in

[2]One could also consider the students who were expected to start high school in the academic year 1999/2000. In July 2000, the eligible students in the treatment group of this older cohort had been exposed to the program for one year, and the students in the control group not at all. Unfortunately, however, for this cohort it is not possible to unambiguously determine the set of students having finished middle school in 1999, and whether these students continued to high school afterwards. This is due to the fact that some of the relevant questions do not appear in the corresponding survey rounds. Therefore, I do not consider this cohort.

2003 or was enrolled in the last grade, thus would supposedly have graduated by 2004. The last two variables are particularly useful to check the medium-term impact on high school enrollment. After all, it could be that any differences in high school enrollment at the end of middle school fade out after a while. This may happen, for instance, if parents and students do eventually overcome any behavioral biases induced by the program and start realizing their full education potential. Or it might be that those in the treatment group had actually formed expectations about a future inclusion of high school students in the program and thus simply postponed enrollment by a little.[3]

In addition to high school outcomes I also consider two outcome variables related to middle school completion: whether a student graduated from middle school in academic year 1999/2000, and whether a student either graduated or was in the second last grade of middle school in academic year 1999/2000, thus would likely have graduated by 2001. The latter variable accounts for students who had a gap between primary and middle school, or who—voluntarily or not—repeated a grade in middle school. While these outcomes are not the main focus of this paper, they serve as a way to verify prior results on the effectiveness of Progresa, and help to put the findings on high school enrollment into perspective. It is worth emphasizing that the average treatment effects on middle school completion and high school enrollment do not allow to make inferences on the average transition probability for middle school graduates, given the potential effect heterogeneity and differences between treatment and control group students at the end of middle school.

Tables 1 and 2 show descriptive statistics by eligibility status as well as treatment and control group. Table 1 summarizes the sample of all adolescents who had graduated from primary school in 1997 and who were between 11 and 14 years old at that time. [4] Thus, assuming a regular school career, these students could have started high school in 2000. I refer to this sample of students as the *unconditional sample*. Table 2 summarizes only those adolescents of the same

[3]Even though the program was extended to include high school students in 2001, it is unlikely that students anticipated this and delayed high school for that reason. The program itself had initially only been promised for three years, and its continuation after the general election in 2000 was uncertain (Schultz 2004, Skoufias 2005).

[4]The variable "had graduated from primary school in 1997" is in fact being constructed from multiple survey questions. The age restriction is being used as an additional fail safe against including people who completed primary school long before 1997. The interval 11 to 14 years is chosen to include students who started school at the regular age and repeated up to two grades. All the calculations in this paper were also done for adolescents between 11 and 16 years as a robustness check. The results do not qualitatively differ from the ones presented in this paper. They are available on request.

Table 1: Descriptive statistics of outcome variables, unconditional sample

	Treatment group				Control group				t-test	
	N	M	SE	NA	N	M	SE	NA	Δ	p
Adolescents from eligible (poor) households										
Started high school in 2000	674	0.132	0.013	308	357	0.162	0.020	168	-0.030	0.327
Some high school by 2003	552	0.246	0.018	430	276	0.322	0.028	249	-0.076	0.104
Graduated or about to graduate from high school in 2003	552	0.178	0.016	430	276	0.228	0.025	249	-0.051	0.209
Graduated from middle school in 2000	713	0.418	0.018	269	373	0.434	0.026	152	-0.016	0.666
Graduated or about to graduate from middle school in 2000	713	0.663	0.018	269	373	0.576	0.026	152	0.087	0.020
Adolescents from non-eligible (non-poor) households										
Started high school in 2000	364	0.261	0.023	216	239	0.134	0.022	122	0.127	0.001
Some high school by 2003	331	0.347	0.026	249	194	0.258	0.031	167	0.090	0.091
Graduated or about to graduate from high school in 2003	331	0.296	0.025	249	194	0.180	0.028	167	0.116	0.012
Graduated from middle school in 2000	388	0.508	0.025	192	264	0.439	0.031	97	0.068	0.171
Graduated or about to graduate from middle school in 2000	388	0.644	0.024	192	264	0.602	0.030	97	0.042	0.345

N = number of observed outcomes, M = mean, SE = standard error, NA = number of unobserved outcomes, Δ = difference in means, p = p-value for a two-sided t-test of equal means, clustered at the village level. The sample includes all adolescents who graduated from primary school and were of age 11-14 in 1997.

Table 2: Descriptive statistics of outcome variables, conditional sample

	Treatment group				Control group				t-test	
	N	M	SE	NA	N	M	SE	NA	Δ	p
Unrestricted conditional sample										
Adolescents from eligible (poor) households										
Started high school in 2000	319	0.310	0.026	0	187	0.417	0.036	0	-0.107	0.097
Some high school by 2003	265	0.332	0.029	54	164	0.409	0.039	23	-0.076	0.241
Graduated or about to graduate from high school in 2003	265	0.268	0.027	54	164	0.341	0.037	23	-0.074	0.246
Adolescents from non-eligible (non-poor) households										
Started high school in 2000	195	0.415	0.035	0	152	0.355	0.039	0	0.060	0.326
Some high school by 2003	161	0.441	0.039	34	124	0.355	0.043	28	0.086	0.249
Graduated or about to graduate from high school in 2003	161	0.379	0.038	34	124	0.266	0.040	28	0.113	0.107
Restricted conditional sample										
Adolescents from eligible households										
Started high school in 2000	213	0.282	0.031	0	120	0.417	0.045	0	-0.135	0.065
Some high school by 2003	178	0.303	0.035	35	103	0.417	0.049	17	-0.114	0.147
Completed high school by 2004	178	0.253	0.033	35	103	0.340	0.047	17	-0.087	0.254
Adolescents from non-eligible households										
Started high school in 2000	123	0.447	0.045	0	100	0.340	0.048	0	0.107	0.165
Some high school by 2003	105	0.505	0.049	18	81	0.346	0.053	19	0.159	0.065
Completed high school by 2004	105	0.427	0.049	18	81	0.235	0.047	19	0.194	0.017

N = number of observed outcomes, M = mean, SE = standard error, NA = number of unobserved outcomes, Δ = difference in means, p = p-value for a two-sided t-test of equal means, clustered at the village level. The unrestricted conditional sample includes all adolescents whose highest completed grade in November 2000 was the last grade of middle school, who were of age 14-17 then, and who were enrolled in school during the academic year 1999/2000. The restricted conditional sample has the further restriction that these adolescents had to be enrolled in the second last grade of middle school during the academic year 1999/2000.

cohorts who finished the last year of middle school in 2000. I refer to this sample of students as the *conditional sample*, since it will be used to compute program treatment effects conditional on having graduated from middle school. Middle school completion is not included as a question in any of the questionnaires, but it can be constructed by taking all adolescents who reported the last year of middle school as their highest completed grade in the November 2000 survey, and who also reported being enrolled in school both in the November 1999 survey and in the March 2000 survey. There are two small caveats with this. First, the conditions do not rule out students who actually graduated from middle school before 2000 and then, in school year 1999/2000, attempted another grade of further education, which was not completed or simply not reported. Second, PROGRESA may have led some of the treated students to repeat the last grade on purpose to remain eligible to the program (see Dubois et al. 2012), and while it may not be wrong to include these students, it is conceivable that they influence the results significantly. So to exclude such cases, a further restriction is to consider only those students who reported the second year of middle school as their highest completed grade in the November 1999 survey. This restricted sample is defined more concisely, but comes at the cost of a loss of potentially relevant observations. Adolescents from eligible (poor) and ineligible (non-poor) households are regarded separately. The estimations of treatment effects are conducted for each of these two groups, to obtain direct program effects and spillover effects, respectively. The tables also report the number of missing observations (*NA*) for each variable and experimental group.[5]

The descriptive statistics indicate that adolescents in the treatment group went to high school with lower probability than those in the control group, for both the unconditional sample and the conditional samples. These differences in probability cannot be interpreted as causal effects. It is to be expected that the composition of students who finish middle school differs between treatment and control group, and that missing outcomes are not missing completely at random. I construct a large number of exogenous characteristics to balance the two groups. They originate from the two pre-treatment surveys, and include such things as demographic and socioeconomic characteristics of the household,

[5] Missing outcomes are mostly a result of sample attrition. The panel also contains a few observation with inconsistent characteristics. These inconsistencies include a sex change, age discrepancies, and diminishing highest school degree. Those observations with inconsistencies between the pre-treatment surveys are dropped. Those with discrepancies only in one of the later surveys are deemed reliable with respect to their pre-treatment characteristics and only have outcomes set to missing.

parents' level of education, parents' assessment of the student's ability and expectations about future educational outcomes, parents' assessment of teacher and school quality, village characteristics, average local education level, and travel times to a number of educational institutions. Using high-dimensional econometric techniques allows me to include a large number of potentially relevant characteristics without having to know in advance which of them are actually correlated with treatment status and high school enrollment. A list of all the considered characteristics is included in section A.1 of the appendix.

4 Identification Strategy

The identification strategy is laid out here with the estimation of direct treatment effects in mind—with adolescents from eligible households as the basic population—but it works identically for the estimation of spillover effects. For each of N students, let $W = (Y, D, X)$, where D is an indicator variable for living in a treatment locality and Y an outcome variable, e.g. an indicator for going to high school. $Y(1)$ and $Y(0)$ denote potential outcomes, so that $Y = DY(1) + (1 - D)Y(0)$. X is a p-dimensional vector of exogenous control variables. All expectations are taken over the distribution of W.

In what follows, I distinguish between unconditional and conditional treatment effects, by which I mean the treatment effects for the correspondent samples, respectively. For the unconditional sample, the statistic of interest is the average treatment effect (ATE),

$$\text{ATE} = \mathrm{E}\left[Y(1) - Y(0)\right]. \tag{1}$$

For the conditional sample, the focus lies on those students who would have finished middle school even without Progresa. This is because the program itself has likely added some students to the pool of middle school graduates in the treatment group. These students do not have a counterpart in the control group, so that for them the treatment effect is not identifiable. On the other hand, it is inconceivable that a student who finishes middle school in the absence of payments would not have done so in their presence. To use the parlance of the literature on local average treatment effects: the analysis is concentrated on the *always-takers*, who by virtue of the experimental setup should be fully represented in both groups. It aims to leave out the *compliers*, as their counterfactual is not observed, as well as the *defiers*, who are nonexistent by assumption. One way to eliminate the compliers and defiers from the sample is to use a trimming technique, dropping students with no overlap in the distribution of covariates or propensity scores. This should eliminate both compliers and defiers from the sample for the ATE. Another way to exclude the compliers is to consider the average treatment effect of the non-treated (ATN),

$$\mathrm{ATN} = \mathrm{E}\left[Y\left(1\right) - Y\left(0\right) \left| D = 0\right.\right], \tag{2}$$

which is based on the distribution of students in the control group. As I am interested in the entire distribution of treatment and control group, the discussion focuses on the ATE. I do, however, compute estimates of both the ATE and the ATN and find that they are very close and statistically indistinguishable in all cases.

If the students under consideration were sampled into treatment and control group at random and if missing outcomes were missing completely at random, the ATE would simply be identified by the difference in sample means between treatment and control group, or average predictive effect (APE),

$$\mathrm{APE} = \mathbb{E}\left[Y \left| D = 1\right.\right] - \mathbb{E}\left[Y \left| D = 0\right.\right]. \tag{3}$$

The APEs of the program are in fact equivalent to the differences in means (Δ) in Tables 1 and 2.

There are two reasons why the APE may not be an unbiased estimator of the ATE, despite the initial randomization of households. The first reason is sample selection, a concern mainly for the conditional sample: it may be that some students in the treatment group only finished middle school because of Progresa. These students would be comparatively less likely to continue to high school, and thus create the false impression that the program has a (more) negative effect on high school enrollment.

The second reason for possible bias is attrition. In the unconditional sample, around 33% of the adolescents identified in the two pre-treatment surveys have missing outcomes from the November 2000 survey, and around 45% of them have missing outcomes from the 2003 survey. In the unrestricted conditional sample, around 16% of the adolescents have missing outcomes from the 2003 survey. Attrition becomes a problem when it does not occur completely at random. For instance, it is conceivable that independently of their treatment status, the students who do not go to high school are more likely to drop out of the sample. Considering only those who stay would then lead the estimate of the ATE to be biased towards 0.

I address these two concerns separately, starting with sample selection. For the estimation of conditional treatment effects, I rely on the assumption that treatment is independent of outcomes conditional on pre-treatment control variables X,

$$Y(1), Y(0) \perp D \mid X. \tag{4}$$

Under this assumption, Rosenbaum and Rubin (1983) famously showed that it is sufficient to condition on the propensity score instead of the whole vector of controls. There are, however, some limitations commonly associated with this approach. The researcher needs to know exactly which variables to condition on, as well as the functional form of the probability model. Economic intuition may be helpful for model selection up to a point. But despite best efforts, seemingly relevant features may nonetheless lead to overfitted propensity scores, while seemingly unrelated variables may hold a lot of predictive power through correlations with important unobserved features. In addition, the established methods require low model complexity for identification—i.e., $p \ll N$—even in cases where a large number of confounders is plausible. Consequently, there is little insurance against misspecification of the probability model, which calls the unconfoundedness assumption (4) and the propensity score method into question.

For this paper, I use the specification and estimation strategy taken in Chernozhukov et al. (2018). To formalize the relationship between D, Y, and X, consider the model

$$Y = g_0(D, X) + U, \quad \mathrm{E}[U \mid D, X] = 0, \tag{5}$$

$$D = m_0(X) + V, \quad \mathrm{E}[V \mid X] = 0. \tag{6}$$

This specification is quite general in that it allows for heterogeneous treatment effects and does not require D and X to be additively separable in the regression function $g_0(D, X)$. $m_0(X)$ is the propensity score, i.e., the conditional probability to be in the treatment group. The subscript 0 indicates true parameters. The ATE is given by

$$\theta_0 \coloneqq \mathrm{E}[g_0(1, X) - g_0(0, X)], \tag{7}$$

and the ATN by

$$\gamma_0 \coloneqq \mathrm{E}\left[g_0\left(1, X\right) - g_0\left(0, X\right) | D = 0\right]. \tag{8}$$

Belloni et al. (2014) and Belloni et al. (2017) point out that in a high-dimensional parameter space, directly estimating equation (5) using sophisticated machine learning methods is ill-advised. While doing so may result in a great fit of Y, this approach neglects how treatment assignment is affected by covariates, potentially resulting in a large regularization bias. One way to overcome this bias—and the approach taken in this paper—is to use machine learning in conjunction with doubly-robust estimation, or double machine learning (Farrell 2015, Belloni et al. 2017, Chernozhukov et al. 2018).

The idea is to estimate the nuisance functions $\eta_0 = (g_0\left(D, X\right), m_0\left(X\right))$ separately using machine learning methods. θ_0 is then identified by plugging these estimates into a set of orthogonal moment conditions, $\mathrm{E}\left[\psi\left(W; \theta_0, \eta_0\right)\right] = 0$. The underlying score functions $\psi\left(W; \theta, \eta\right)$ for ATE and ATN are explained in section A.2 in the appendix. Another crucial part of the DML method is cross-fitting, a sample-splitting technique to ensure that the same observations used to estimate nuisance functions $g_0\left(D, X\right)$ and $m_0\left(X\right)$ are not also used to make predictions thereof. This is done to prevent bias induced by overfitting, which is likely to occur for most machine learning techniques even after careful calibration of hyperparameters. The possibility to aggregate or choose the best out of multiple machine learning methods guarantees estimability for a wide range of data generating processes.[6] Details on the cross-fitting procedure are given in section A.3 of the appendix.

Having discussed how the problem of nonrandom sample selection is approached, I now turn to nonrandom attrition. Let R be an indicator variable for remaining in the sample, thus taking the value 1 if Y is non-missing and 0 otherwise, and $W = (Y, D, R, X)$. I assume that outcomes are missing at random, meaning that attrition is independent of outcomes conditional on treatment status D and control variables X,

$$Y\left(1\right), Y\left(0\right) \perp R \,|\, (D, X). \tag{9}$$

So, while attrition on its own may be predictive of outcomes, this predictive

[6] An alternative approach to deal with regularization bias is discussed in Athey et al. (2018). It does not require estimability of the propensity score, but in turn limits the complexity of the regression function by assuming strong sparsity.

power comes entirely from observable variables. The approach is similar to the one taken in Behrman et al. (2009) on the medium-term effects of Progresa, where attrition from the 2003 survey is also assumed to be random conditional on a (small) number of observables and treatment status. I propose the following extension of the model above to accommodate this assumption:

$$Y = g_0(D, X) + U, \quad \mathrm{E}[U \,|D, R, X] = 0, \tag{10}$$

$$D = m_0(X) + V, \quad \mathrm{E}[V \,|X] = 0, \tag{11}$$

$$R = r_0(D, X) + Z, \quad \mathrm{E}[Z \,|D, X] = 0. \tag{12}$$

$r_0(D, X)$ is the conditional probability that student i's outcome is observed. R does not enter the regression function g_0, but $r_0(D, X)$ is needed to account for possible differences in the distribution of (D, X) between students with observed and unobserved outcomes.

As long as the nuisance functions are well approximated by any of the machine learning methods used, the resulting DML estimators $\hat{\theta}_0$ and $\hat{\gamma}_0$ are $\sqrt{N}$-consistent, approximately unbiased and asymptotically normally distributed. For the model without attrition, this is stated in Theorem 5.1 of Chernozhukov et al. (2018). An equivalent version of this theorem for the model with attrition is given in section A.4 of the appendix of this paper.

5 Estimation

I estimate the ATE and the ATN using 10-fold cross-fitting with 100 repetitions. The sample is being split such that all students from the same location end up in the same fold. For the separate estimation of the nuisance functions $\eta_0(X)$, I use six different machine learning methods. The first three machine learning techniques are regularized logistic regression techniques, namely the Lasso (with ℓ_1 penalty), Ridge (with ℓ_2 penalty), and elastic net (with both ℓ_1 and ℓ_2 penalty). Furthermore, I use two tree-based techniques—namely the random forest and extreme gradient boosting—and support vector machines (SVM).[7] In addition, I include a technique that combines the best machine

[7] In addition to these methods, I also considered neural networks with one hidden layer, as well as different ensemble learners that would combine the aforementioned methods. Both the neural network and the ensemble methods turned out to be computationally expensive to tune, while showing relatively poor predictive performance. For that reason, I chose to leave them out eventually.

learning methods for each nuisance function, i.e., the ones that produce the smallest out-of-sample mean squared error (or: Brier score). Before the actual estimation, the hyperparameters for each machine learning method are tuned to maximize out-of-sample predictive power. This is done via repeated cross-validation, using 10 folds and 10 repetitions. Details on hyperparameter tuning, data preparation for each method, and handling of missing feature values are given in section A.5 of the appendix.

To exclude extreme values for the propensity score and to guarantee overlap between treatment and control group, I apply the trimming procedure developed in Crump et al. (2009) and Imbens and Rubin (2015). It produces an interval $(\alpha, 1-\alpha)$ such that all observations with propensity scores outside this interval are discarded. The number of trimmed observations varies by sample and cross fitting iteration, but it is 0 for the large majority of iterations for each sample. The maximum fraction of discarded observations in an iteration is 5.2% for the conditional samples and 0.3% for the unconditional sample.

For the unconditional sample, I compute the ATE for the three high school indicators and the two middle school indicators discussed above as outcomes.[8] For the conditional sample, I compute both the ATE and ATN for the three high school indicators. For the variance estimation of the DML estimators, it is necessary to account for clustering, since the treatment status of the Progresa experiment does not vary within villages. Since the number of observations per village varies substantially, cluster-robust standard errors may not be consistent, as is argued in Mackinnon and Webb (2017). This can be overcome by using a wild cluster bootstrap instead. I obtain standard errors in this way using 100,000 bootstrap replications and the distribution for the bootstrap multiplier suggested by Mammen (1993). The ATE results from combining the nuisance function estimates of the respective best-fitting machine learning methods (i.e., my preferred specification) are reported in Table 3. More detailed results, with estimates for each of the machine learning methods used and including the ATN, are depicted in Tables 8 to 15 in section A.7 of the appendix.

[8] In fact, since the unconditional sample is randomized, it would be sufficient for an unbiased estimate of the ATE to correct only for attrition. But accounting for covariates may increase precision even in the absence of sample selection, particularly if treatment and control group are not completely balanced, as is suggested by Behrman and Todd (1999) for the Progresa baseline data.

Table 3: ATE estimates of middle and high school education

	eligible (poor)			non-eligible (non-poor)		
Dependent variable	uncond.	cond. unres.	cond. res.	uncond.	cond. unres.	cond. res.
Started high school in 2000	-0.033	-0.125**	-0.145**	0.129***	0.066	0.093
	(0.030)	(0.055)	(0.069)	(0.035)	(0.055)	(0.073)
Some high school by 2003	-0.085**	-0.101*	-0.126	0.100**	0.109	0.162**
	(0.043)	(0.059)	(0.078)	(0.044)	(0.072)	(0.080)
Graduated or about to graduate from high school in 2003	-0.050	-0.100*	-0.098	0.107***	0.149**	0.193**
	(0.035)	(0.056)	(0.076)	(0.038)	(0.064)	(0.078)
Graduated from middle school by 2000	-0.029			0.069		
	(0.032)			(0.044)		
Graduated or about to graduate from middle school in 2000	0.088***			0.026		
	(0.032)			(0.037)		
Observations	1,507	506	333	941	347	223

*: significant at the 10% level. **: significant at the 5% level. ***: significant at the 1% level.
The average treatment effect (ATE) together with standard errors (in parentheses) is estimated for the unconditional sample (adolescents who graduated from primary school in 1997) and for the unrestricted and restricted conditional samples (adolescents who graduated from middle school in 2000; in addition, the restricted sample has only those having been to the second-last grade of middle school in 1999), and separately for adolescents from eligible and non-eligible households. Treatment effects are estimated for five outcome variables: whether the student continued with high school in school year 2000/01, whether the student had attended some high school by the end of 2003, whether the student had finished or was about to finish high school by the end of 2003, whether the student had finished middle school by the end of 2000, and whether the student had finished or was in the last grade of middle school by the end of 2000.
Point estimates are obtained via the respective orthogonal moment conditions (see appendix section A.2). Standard errors are obtained via wild cluster bootstrap. Results are based on 10-fold cross-fitting, using 100 sample splits. Estimates from the sample splits are combined using the median method, based on equations (18) and (19). Only results from combining the best-fitting machine learning methods for each nuisance function are reported here. Estimates of ATE and ATN for all the machine learning methods are reported in the section A.7 of the appendix.

6 Discussion and Possible Channels

6.1 Discussion of Results

I start by discussing the eligible students, since they are the main focus of this paper. Looking at the first column of Table 3, it appears that the program did not have a statistically or economically significant effect on timely high school enrollment for the overall student population. However, it seems that by 2003, the program made it less likely by about 8.5 percentage points for students to have enrolled in high school at some point. Looking at the third outcome, it is unclear whether this translated into lower high school graduation rates for treated students. It is a somewhat unexpected result that students in the treatment group who had just completed primary school when the program started would not have *higher* eventual high school continuation rates. This is especially so since PROGRESA seems to have had a positive effect on middle school completion: the probability to have graduated from or to be in the last grade of middle school in 2000 went up by about 8.8 percentage points.[9] Therefore, it must be that the program had a negative effect on the continuation decisions of middle school graduates. This hypothesis is confirmed when looking at the second and third columns. The ATE for high school in 2000 is -12.5 and -14.5 percentage points for the unrestricted and restricted sample, respectively. The effect is slightly smaller when looking at high school participation until 2003, with -10.1 and -12.6 percentage points, and at high school graduation or near graduation by 2003, with -10.0 and -9.8 percentage points.

For the non-eligible students, the fourth column of Table 3 shows that being in the treatment group increased high school enrollment and graduation by up to 13 percentage points. This result is partly driven by the program's known spillover effect of middle school graduation on non-eligible students (Attanasio et al., 2012), which is also confirmed here. One would not expect that it is also driven by the program's direct effect on middle school graduates. Looking at the fifth and sixth column, though, this seems to be the case: the ATE for all high school outcomes is positive, and it is statistically significant for high

[9]However, the effect is close to zero when only considering middle school graduation by 2000. The difference between the effects on graduation by 2000 and 2001 could arise because some adolescents may have enrolled in middle school in response to the program after it was launched in 1998, while the program did not affect the completion rate of those already enrolled at that time. Another explanation is that some students in the treatment group purposefully repeated a grade to receive payments longer, as is conjectured by Dubois et al. (2012).

school graduation and (for the restricted sample) for having done some high school by 2003. This result is unexpected, since it implies a sort of reverse spillover effect: those who saw their peers getting paid became more likely to continue schooling, while those actually getting paid lost interest. There are nonetheless ways to rationalize this result. One explanation may be peer effects: an increase in eligible low ability students in middle school as a result of the program may reduce learning outcomes for other low ability students, while it may strengthen the relative position of high ability students (of which many may be non-eligible) and thus heighten their self-esteem and motivation. An influx of poor students in middle school may also increase the need of non-poor students and their parents to distinguish themselves through further education. It is also conceivable that seeing their peers getting paid triggers the ineligible students' will to demonstrate their capability despite being at a relative disadvantage.

6.2 Possible Channels

The precise channels for the observed treatment effects are impossible to learn with certainty from the data at hand. Nonetheless, I can examine two of the possible causes mentioned in section 2, namely classroom peer effects and loss aversion, and check whether they constitute credible explanations for the observed effects. Starting with classroom peer effects: if it is true that PROGRESA worsens the pool of middle school students, one would expect this to show in measures of performance. Unfortunately, such measures are not available until the 2003 survey. However, parents' assessments are available from the pre-treatment survey of March 1998. I consider three binary variables: whether (according the the parents) the student is good at school, whether the student is apt enough to go to high school or further, and whether the student is apt enough to go to university. While these assessments are certainly very noisy signals of a student's ability, there is no obvious reason to believe they should not at least convey some information thereon. Table 4 shows the differences in averages of the assessment variables between the treatment and control group, both for eligible and non-eligible students. Judging by this table, the hypothesis that the eligible students who finished middle school in the treatment group should on average be less apt than those in the control group is not supported. If anything, the numbers suggest the opposite. Thus, the theory of negative classroom peer effects is not supported by the data.

Next, in order to examine loss aversion as a possible explanation, I check

Table 4: Comparing parents' assessment of student's ability

	Treatment group			Control group			t-test	
	N	M	SE	N	M	SE	Δ	p
Adolescents from eligible households								
Student is good at school	264	0.614	0.030	156	0.577	0.040	0.037	0.459
Student can make it at least to high school	315	0.352	0.027	181	0.287	0.034	0.065	0.138
Student can make it to university	355	0.117	0.018	181	0.061	0.018	0.057	0.040
Adolescents from non-eligible households								
Student is good at school	160	0.613	0.039	132	0.621	0.042	-0.009	0.879
Student can make it at least to high school	187	0.380	0.036	151	0.411	0.040	-0.031	0.564
Student can make it to university	187	0.187	0.029	151	0.172	0.031	0.015	0.723

N = number of observed outcomes, M = mean, SE = standard error, Δ = difference in means, p = p-value for a two-sided t-test of equal means, clustered at the village level. Based on the unrestricted conditional sample.

whether financial concerns are responsible for the differences in high school enrollment between treatment and control group. To that end, I look at a question asking for the reasons why students did not go to school in the November 2000 survey. Of the possible answers to this question, three point to financial constraints, namely: (1) there is not enough money to send the student to school, (2) the student is needed for work, and (3) the student is needed at home. I lump these together in an indicator variable that is 1 if one of these three reasons were given and 0 otherwise. In the same fashion, I create another indicator that lumps together all the other reasons why a student might not go to high school. The most frequent reasons here include: (1) the student does not like school, (2) the student attends a vocational school, (3) the student is already grown up, and (4) the school is too far away. Yet another variable related to financial constraints may be actual household expenditures. Therefore, I also construct a variable of monthly per capita expenditure, using prices and quantities of goods as indicated in the section on consumption and expenditure from the November 2000 survey.

I use these three variables—student does not go to high school due to financial reasons, student does not go to high school for other than financial reasons, and monthly per capita expenditure—as outcomes and apply the DML model on the conditional samples. This way it should be possible to see whether or not the program effects can be explained through intermediate effects on household finances. In particular, if loss aversion is the main driver of the negative treatment effect for the eligible students, this might show in a higher share of students not going to school for money reasons and higher per capita expenditures in the treatment group. Table 5 sums up the results, with more detailed outputs in Tables 16 to 19 in section A.7 of the appendix.

The results show that not going to high school due to financial constraints as well as expenditure seem to be nearly unaffected by the program. On the other hand, the program does seem to disincline eligible students from high school for other than financial reasons, whereas it has the opposite effect on non-eligible students. This result indicates that loss aversion may not be the main explanation for the negative treatment effect on eligible students. Instead, it seems more likely that the program impacts enrollment through other factors such as social norms or students' motivation.

Table 5: ATE estimates of further outcomes

	eligible		non-eligible	
Dependent variable	unrestr.	restr.	unrestr.	restr.
Not going to high school due to financial constraints	0.035	0.031	0.002	-0.010
	(0.053)	(0.064)	(0.057)	(0.072)
Not going to high school for other than financial reasons	0.073*	0.113**	-0.087*	-0.085
	(0.042)	(0.051)	(0.046)	(0.060)
Log monthly per capita expenditure	0.059	0.019	-0.079	-0.113
	(0.063)	(0.077)	(0.081)	(0.101)
Observations	506	333	347	223

*: significant at the 10% level. **: significant at the 5% level. ***: significant at the 1% level.
The average treatment effect (ATE) together with standard errors (in parentheses) is estimated for the unrestricted and restricted conditional samples, and separately for students from eligible and non-eligible households. The unrestricted sample consists of all adolescents who completed middle school in 2000 and were of age 14-17 then. The restricted sample consists of only those adolescents who in addition were found to be enrolled in the second-last grade of middle school in 1999. Treatment effects are estimated for three outcome variables: whether the adolescent did not continue with high school in school year 2000/01 for money-related reasons, whether the adolescent did not continue with high school in school year 2000/01 for reasons other than money, and log monthly per capita expenditure in the adolescent's household.
Point estimates are obtained via the respective orthogonal moment conditions (see appendix section A.2). Standard errors are obtained via wild cluster bootstrap. Results are based on 10-fold cross-fitting, using 100 sample splits. Estimates from the sample splits are combined using the median method, based on equations (18) and (19). Only results from combining the best-fitting machine learning methods for each nuisance function are reported here. Estimates of ATE and ATN for all the machine learning methods are reported in section A.7 of the appendix.

6.3 Coefficient Stability

The estimation of treatment effects using the DML method takes into account a large number of observed characteristics, as well as unobserved characteristics that are correlated with any combination of the observed ones. Nonetheless, the identification is arguably not impervious to any unobserved characteristics. (Due to the lack of an ideal experiment that could be emulated, this holds for any estimation strategy for conditional aftereffects.) Therefore, it seems worth assessing in how far selection bias may remain an issue. One approach to doing so is offered by Oster (2017), who developed an estimator for the omitted variable bias. The idea is to compare the main estimator of interest to the estimate of a simple regression of the outcome on the treatment variable. Ceteris paribus, the closer the former is to the latter, the less the observed covariates seem to matter, and the less bias is expected from omitting unobserved ones. Similarly, the better the model with observed covariates explains the outcome and the worse the model without covariates does, the less scope for omitted variable bias is left. In section A.6 of the appendix, I compute a version of Oster's estimator. It suggests that for all results with significant treatment effects, an inclusion of unobserved factors in the outcome model would not change the ATE estimator enough to cancel out the treatment effect—even if the influence of those factors was up to five times that of the observable factors. Moreover, adding observable characteristics to the simple regression moves the ATE estimate further away from zero. So if the correlations of unobserved portion and observed portion with the treatment variable have the same sign, if anything, one would expect the true effects to be larger in absolute terms than the estimates. In summary, it appears that unobserved factors are unlikely to invalidate the findings.

7 Conclusion

The positive effects of CCT programs like Progresa on school enrollment have been demonstrated in numerous studies. However, surprisingly, their aftereffects have not been explored so far. With this paper, I try to fill this gap by estimating Progresa's impact on the probability to continue school after program payments stop. The main finding is that for the eligible students, the program has large and significant negative aftereffects. There are a number of possible explanations. Financial incentives may crowd out the social norm of sending children to school or reduce the intrinsic motivation to attend school regularly once they stop being in place. Establishing program payments and then reducing them to zero again may convey the false signal that education is not worth it at the later levels. Payments may shift parents' income reference point such that the sudden drop needs to be compensated by the child's wage income. And a change in the composition of students induced by the program may lead to negative classroom peer effects. Though conclusive evidence in favor of one over the other explanations is lacking, it seems that loss aversion and classroom peer effects are not much supported by the data, leaving motivation crowding and anchoring as the remaining candidates.

The paper also looks at possible spillovers to the students who were not eligible to the program but lived in treatment villages. Curiously, it seems that—if anything—these students are more likely to finish high school as a result of their peers getting paid. This could be explained by a heightened desire of the non-poor students to separate themselves from the poor through more education, or by a surge in self-esteem as a result of changes in the classroom composition.

Of course, the findings of this paper are confined to the relatively short time the program had been in effect by the year 2000. For younger cohorts who start middle school with the program already in place, one may expect positive unconditional treatment effects, via the intermediate positive effect the program has on middle school enrollment. On the other hand, the conditional treatment effects might be even more extreme due to a longer exposure to the program.

The main result is remarkable, as it constitutes a textbook case of unintended consequences. It encourages to look further into how motivation and social norms change through financial incentives. It raises the question of whether a potential motivation crowding effect carries on to higher education, vocational training, or the labor market, and whether it shows not only in participation but also performance. Moreover, the finding should be considered in the program design of future CCT programs. Even in cases where coverage on all school levels is not feasible due to budgetary constraints, there may be ways to counter the adverse program effect. This could, for instance, be done by systematically informing students and parents about the marginal rate of return to continued education. Another way may be to let go of the conditionality of payments altogether, particularly to counter possible crowding-out effects and anchoring. Further research that explores these channels may help to mitigate the negative side effects of CCT programs.

References

Angelucci, M. and G. De Giorgi (2009). Indirect effects of an aid program: How do cash transfers affect ineligibles' consumption? *American Economic Review 99*(1), 486–508.

Angelucci, M., G. D. Giorgi, M. A. Rangel, and I. Rasul (2010). Family networks and school enrolment: Evidence from a randomized social experiment. *Journal of Public Economics 94*(3), 197 – 221.

Athey, S., G. W. Imbens, and S. Wager (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society Series B 80*(4), 597–623.

Attanasio, O. P., C. Meghir, and A. Santiago (2012). Education choices in Mexico: Using a structural model and a randomized experiment to evaluate PROGRESA. *The Review of Economic Studies 79*(1), 37–66.

Bang, H. and J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61*(4), 962–973.

Bastagli, F., J. Hagen-Zanker, L. Harman, V. Barca, G. Sturge, T. Schmidt, and L. Pellerano (2016). Cash transfers: What does the evidence say? A rigorous review of programme impact and of the role of design and implementation features. Ods report, Overseas Development Institute.

Behrman, J. R., S. W. Parker, and P. E. Todd (2009). 7 medium-term impacts of the Oportunidades conditional cash transfer program on rural youth in Mexico. *Poverty, Inequality, and Policy in Latin America*, 219.

Behrman, J. R., P. Sengupta, and P. Todd (2005). Progressing through PROGRESA: An impact assessment of a school subsidy experiment in rural Mexico. *Economic Development and Cultural Change 54*(1), 237–275.

Behrman, J. R. and P. E. Todd (1999). Randomness in the experimental samples of PROGRESA (education, health, and nutrition program). *International Food Policy Research Institute, Washington, DC*.

Belloni, A., V. Chernozhukov, I. Fernandez-Val, and C. Hansen (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica 85*(1), 233–298.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*(2), 608–650.

Carrell, S. E. and M. L. Hoekstra (2010, January). Externalities in the classroom: How children exposed to domestic violence affect everyone's kids. *American Economic Journal: Applied Economics 2*(1), 211–28.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika 96*(1), 187–199.

de Janvry, A., F. Finan, E. Sadoulet, and R. Vakis (2006). Can conditional cash transfer programs serve as safety nets in keeping children at school and from working when exposed to shocks? *Journal of Development Economics 79*(2), 349 – 373. Special Issue in honor of Pranab Bardhan.

Deci, E., R. Koestner, and R. Ryan (1999, 12). A meta-analytic review of experiments examining the effect of extrinsic rewards on intrinsic motivation. *Psychological bulletin 125*, 627–68; discussion 692.

Dizon-Ross, R. (2018, May). ParentsÂŽ beliefs about their childrenÂŽs academic ability: Implications for educational investments. Working Paper 24610, National Bureau of Economic Research.

Dubois, P., A. de Janvry, and E. Sadoulet (2012). Effects on school enrollment and performance of a conditional cash transfer program in mexico. *Journal of Labor Economics 30*(3), 555–589.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics 189*(1), 1 – 23.

Fehr, E. and A. Falk (2002). Psychological Foundations of Incentives. IEW - Working Papers 095, Institute for Empirical Research in Economics - University of Zurich.

Fiszbein, A. and N. R. Schady (2010). *Conditional Cash Transfers.* The World Bank.

Frey, B. S. and R. Jegen (2001). Motivation crowding theory. *Journal of Economic Surveys 15*(5), 589–611.

Gantner, L. (2009). *PROGRESA: An Integrated Approach to Poverty Alleviation in Mexico*, pp. 211–220. Cornell University Press.

Gneezy, U. and A. Rustichini (2000). A fine is a price. *The Journal of Legal Studies 29*(1), 1–17.

Imbens, G. W. and D. B. Rubin (2015). *Trimming to Improve Balance in Covariate Distributions*, pp. 359–374. Cambridge University Press.

Jensen, R. (2010). The (perceived) returns to education and the demand for schooling. *The Quarterly Journal of Economics 125*(2), 515–548.

Kahneman, D. and A. Tversky (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica 47*(2), 263–291.

Lavy, V., M. D. Paserman, and A. Schlosser (2012). Inside the black box of ability peer effects: Evidence from variation in the proportion of low achievers in the classroom. *The Economic Journal 122*(559), 208–237.

Loewenstein, G. F. (1988). Frames of mind in intertemporal choice. *Manage. Sci. 34*(2), 200–214.

Mackinnon, J. G. and M. D. Webb (2017). Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics 32*(2), 233–254.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics 21*(1), 255–285.

Nguyen, T. (2008). Information, role models and perceived returns to education: Experimental evidence from Madagascar. *Unpublished manuscript*.

Oster, E. (2017). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 1–18.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Schultz, T. P. (2004). School subsidies for the poor: evaluating the Mexican Progresa poverty program. *Journal of Development Economics 74*(1), 199 – 250. New Research on Education in Developing Economies.

Skoufias, E. (2005, 02). PROGRESA and its impacts on the welfare of rural households in Mexico. Research report of the international food policy research institute.

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization 1*(1), 39 – 60.

Todd, P. E. and K. I. Wolpin (2006). Assessing the impact of a school subsidy program in Mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility. *The American Economic Review 96*(5), 1384–1417.

Tversky, A. and D. Kahneman (1974). Judgment under uncertainty: Heuristics and biases. *Science 185*(4157), 1124–1131.

Tversky, A. and D. Kahneman (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics 106*(4), 1039–1061.

Weiss, Y., A. Hall, and F. Dong (1980). The effect of price and income on investment in schooling. *Journal of Human Resources 15*(4), 611–640.

A Appendix

A.1 List of Pre-Treatment Characteristics

Variable description	Type
Student and household characteristics	
Student is female	binary
Age of student in 1997	continuous
Degree of poverty index in 1997 (by 1997 criteria)	continuous
Degree of poverty index in 1997 (by 2003 criteria)	continuous
Very poor, poor, marginally non-poor, or clearly non-poor in 1997 (by 1997 criteria)	categorical
Household size	count
Number of household members below age 15	count
Father lives in the household	binary
Mother lives in the household	binary
Father is literate	binary
Mother is literate	binary
Father went to school	binary
Mother went to school	binary
Father finished at least primary school	binary
Mother finished at least primary school	binary
Father finished at least middle school	binary
Mother finished at least middle school	binary
Student attended school in October 1997	binary
Student attended school in March 1998	binary
Parents' assessments and opinions	
Parents say student is good at school in 1998	binary
Parents say student is able to finish middle school	binary
Parents say student is able to continue after middle school	binary
Parents say student can finish high school	binary
Parents say student is able to finish university	binary
Desired level of schooling for girls is at least middle school	binary
Desired level of schooling for girls is more than middle school	binary
Desired level of schooling for girls is at least high school	binary

Variable description	Type
Desired level of schooling for girls is university	binary
Desired level of schooling for boys is at least middle school	binary
Desired level of schooling for boys is more than middle school	binary
Desired level of schooling for boys is at least high school	binary
Desired level of schooling for boys is university	binary
Children eat breakfast before school	binary
Reason why children don't eat breakfast before school	categorical
Parent talked to teacher this year	binary
Reason for talk with teacher	categorical
Parent participates in parent / guardian association of school	binary
Parent participates in school work	binary
In school, there are problems with lack of discipline	binary
In school, there are problems with lack of interest of the teachers	binary
In school, there are problems with poor communication between teachers and parents	binary
In school, there are problems with poor teacher attendance	binary
The teacher is usually prepared	binary
The teacher is usually fulfilled	binary
The teacher is usually on time	binary
The teacher is usually patient with the children	binary
Age from which girls can help younger siblings	continuous
Age from which boys can help younger siblings	continuous
Age from which girls can help with work	continuous
Age from which boys can help with work	continuous
Age from which girls can work to earn money	continuous
Age from which boys can work to earn money	continuous
Household expenditures	
Weekly expenditures for public transport to school	continuous
Weekly expenditures for public transport for other trips	continuous
Weekly expenditures for cigarettes and tobacco	continuous
Weekly expenditures for alcoholic beverages	continuous
Weekly expenditures for nonalcoholic beverages	continuous
Monthly expenditures for hygiene items	continuous
Monthly expenditures for medicine	continuous

Variable description	Type
Monthly expenditures for medical consultations	continuous
Biannual expenditures for household articles	continuous
Biannual expenditures for toys	continuous
Biannual expenditures for girls' clothes	continuous
Biannual expenditures for boys' clothes	continuous
Biannual expenditures for women's clothes	continuous
Biannual expenditures for men's clothes	continuous
Biannual expenditures for girls' shoes	continuous
Biannual expenditures for boys' shoes	continuous
Biannual expenditures for women's shoes	continuous
Biannual expenditures for men's shoes	continuous
Biannual expenditures for school supplies	continuous
Biannual expenditures for school contributions	continuous
If family had more money, they would spend it on food	rank
If family had more money, they would spend it on housing repairs	rank
If family had more money, they would spend it on clothing or shoes	rank
If family had more money, they would spend it on debt settlement	rank
If family had more money, they would spend it on animals	rank
If family had more money, they would spend it on seeds or plants	rank
If family had more money, they would spend it on work tools	rank
If family had more money, they would spend it on medicine	rank
If family had more money, they would spend it on school supplies	rank
If family had more money, they would save it	rank
Location characteristics	
Marginality index	continuous
Degree of marginality very high (1) or high (0) in 1997	binary
Village is indigenous	binary
Village has a municipal delegate	binary
Village has a municipal subdelegate	binary
Village has a commissioner of agricultural land	binary
Village has a commissioner of communal goods	binary
Village has a municipal development committee	binary
Village has a health committee	binary
Village has a education committee	binary

Variable description	Type
Village has a agricultural committee	binary
Village has a Diconsa store officer	binary
Village has a production cooperative	binary
Village has religious organizations	binary
Village has political organizations	binary
Village has a school parent association	binary
Village has community assemblies	binary
Village has NGOs	binary
Village has a communal work system (tequio)	binary
Source of water	categorical
Type of garbage disposal	categorical
Electricity available everywhere	binary
Electricity at least partly available	binary
Public drainage at least partly available	binary
Public phone available	binary
Number of preschools	count
Number of primary schools	count
Number of distance middle schools	count
Most important sector in this village	categorical
Second most important sector in this village	categorical
Third most important sector in this village	categorical
Child labor takes place in this village	binary
Average daily salary paid to children	continuous
Number of inhabitants	count
Number of poor inhabitants	count
Number of primary school graduates between 11 and 14	count
Share among inhabitants of primary school graduates between 11 and 14	share
Number of poor primary school graduates between 11 and 14	count
Share among poor inhabitants of primary school graduates between 11 and 14	share
Number of primary school graduates between 11 and 14 enrolled in 1997	count
Share among inhabitants of primary school graduates between 11 and 14 enrolled in 1997	share
Number of poor primary school graduates between 11 and 14 enrolled in 1997	count
Share among poor inhabitants of prim. school graduates between 11 and 14 enrolled in 1997	share

Variable description	Type
Number of inhabitants who completed at least primary school	count
Share among inhabitants who completed at least primary school	share
Share among inhabitants between 15 and 20 who completed at least primary school	share
Share among inhabitants between 21 and 30 who completed at least primary school	share
Number of poor inhabitants who completed at least primary school	count
Share among poor inhabitants who completed at least primary school	share
Share among poor inhabitants between 15 and 20 who completed at least primary school	share
Share among poor inhabitants between 21 and 30 who completed at least primary school	share
Number of inhabitants who completed at least secondary school	count
Share among inhabitants who completed at least secondary school	share
Share among inhabitants between 15 and 20 who completed at least secondary school	share
Share among inhabitants between 21 and 30 who completed at least secondary school	share
Number of poor inhabitants who completed at least secondary school	count
Share among poor inhabitants who completed at least secondary school	share
Share among poor inhabitants between 15 and 20 who completed at least secondary school	share
Share among poor inhabitants between 21 and 30 who completed at least secondary school	share
Number of inhabitants who completed at least high school	count
Share among inhabitants who completed at least high school	share
Share among inhabitants between 15 and 20 who completed at least high school	share
Share among inhabitants between 21 and 30 who completed at least high school	share
Number of poor inhabitants who completed at least high school	count
Share among poor inhabitants who completed at least high school	share
Share among poor inhabitants between 15 and 20 who completed at least high school	share
Share among poor inhabitants between 21 and 30 who completed at least high school	share
Travel time (minutes) to nearest private middle school	continuous
Travel time (minutes) to nearest public middle school	continuous
Travel time (minutes) to nearest distance middle school	continuous

Variable description	Type
Travel time (minutes) to nearest middle school	continuous
Travel time (minutes) to nearest private high school	continuous
Travel time (minutes) to nearest public high school	continuous
Travel time (minutes) to nearest high school	continuous
Travel time (minutes) to nearest national college of technical professional education	continuous
Travel time (minutes) to nearest agricultural technological center	continuous
Travel time (minutes) to nearest industrial technology and services center	continuous
Travel time (minutes) to nearest agricultural college	continuous
Travel time (minutes) to nearest industrial and services college	continuous

A.2 Score functions

For the model without attrition, Chernozhukov et al. (2018) show that under a number of regularity conditions—particularly concerning the speed at which the nuisance functions η converge to their true values η_0—and using Neyman-orthogonal moment conditions as well as cross-fitting, their estimators of the ATE and the ATT (average treatment effect of the treated) are $\sqrt{N}$-consistent and asymptotically normal. The authors state that a crude requirement for the nuisance functions is that they converge at rate $o\left(N^{-1/4}\right)$. This rate is shown to be achievable for a variety of data generating processes in conjunction with specific machine learning methods. Given the possibility to aggregate or choose the best out of multiple machine learning methods, this guarantees estimability for a wide range of problems.

A Neyman-orthogonal score function for the ATE is

$$\begin{aligned}\psi_\theta\left(W;\theta,\eta\right) = g\left(1,X\right) - g\left(0,X\right) + \frac{D\left(Y-g\left(1,X\right)\right)}{m\left(X\right)} \\ - \frac{\left(1-D\right)\left(Y-g\left(0,X\right)\right)}{1-m\left(X\right)} - \theta,\end{aligned} \tag{13}$$

with data $W=(Y,D,X)$ and nuisance functions $\eta\left(X\right)=\left(g\left(1,X\right),g\left(0,X\right),m\left(X\right)\right)$. The true value of η is $\eta_0\left(X\right)=\left(g_0\left(1,X\right),g_0\left(0,X\right),m_0\left(X\right)\right)$. A Neyman-orthogonal score function for the ATN is

$$\begin{aligned}\psi_\gamma\left(W;\gamma,\eta\right) = \frac{D\left(1-m\left(X\right)\right)\left(Y-g\left(1,X\right)\right)}{m\left(X\right)\left(1-p_D\right)} - \frac{\left(1-D\right)\left(Y-g\left(0,X\right)\right)}{1-p_D} \\ + \frac{\left(1-D\right)\left(g\left(1,X\right)-g\left(0,X\right)\right)}{1-p_D} - \gamma\frac{1-D}{1-p_D},\end{aligned} \tag{14}$$

with nuisance functions $\eta\left(X\right)=\left(g\left(1,X\right),g\left(0,X\right),m\left(X\right),p_D\right)$. Here, the true value of η is $\eta_0\left(X\right)=\left(g_0\left(1,X\right),g_0\left(0,X\right),m_0\left(X\right),\mathrm{E}\left[D\right]\right)$.

For the model with attrition, the moment conditions need to be adapted to the fact that not all outcomes are observed. This entails that treated observations with observed outcomes are weighted by the inverse conditional probability of being observed and treated. Non-treated observations with observed outcomes are weighted by the inverse conditional probability of being observed and non-treated.

Let $q_0\left(D,X\right) \coloneqq Dr_0\left(1,X\right)m_0\left(X\right)+\left(1-D\right)r_0\left(0,X\right)\left(1-m_0\left(X\right)\right)$. Then, the corresponding score function for the ATE is

$$\begin{aligned}\varphi_\theta\left(W;\theta,\eta\right) = g\left(1,X\right) - g\left(0,X\right) + \frac{RD\left(Y-g\left(1,X\right)\right)}{q\left(1,X\right)} \\ - \frac{R\left(1-D\right)\left(Y-g\left(0,X\right)\right)}{q\left(0,X\right)} - \theta,\end{aligned} \tag{15}$$

with data $W = (Y, D, R, X)$ and nuisance functions $\eta(X) = (g(1, X), g(0, X), q(1, X), q(0, X))$ whose true value is $\eta_0(X) = (g_0(1, X), g_0(0, X), r_0(1, X) m_0(X), r_0(0, X)(1 - m_0(X)))$.
For the ATN, the corresponding score function is

$$\begin{aligned} \varphi_\gamma(W; \gamma, \eta) = & \frac{RD(Y - g(1, X))(1 - m(X))}{q(1, X)(1 - p_D)} - \frac{R(1 - D)(Y - g(0, X))}{r(0, X)(1 - p_D)} \\ & + \frac{(1 - D)(g(1, X) - g(0, X))}{1 - p_D} - \gamma \frac{1 - D}{1 - p_D}, \end{aligned} \tag{16}$$

with nuisance functions $\eta(X) = (g(1, X), g(0, X), m(X), r(0, X), q(1, X), p_D)$, whose true value is $\eta_0(X) = (g_0(1, X), g_0(0, X), m_0(X), r_0(0, X), r_0(1, X) m_0(X), \mathrm{E}[D])$.

A.3 Repeated cross-fitting

The cross-fitting procedure works the same way for all score functions; I use ψ_θ as the example here. For a fixed integer K, the sample is randomly split into folds $I_1, \ldots, I_K$ of roughly equal size. For each $k \in \{1, \ldots, K\}$, the nuisance functions are estimated using only the observations outside of I_k. The resulting functional estimates are then used to predict $\eta_0(X)$ in fold I_k. The predictions over all folds are in turn used to obtain the point estimate of θ, through the equation

$$\mathbb{E}\left[\psi_\theta\left(W; \hat{\theta}, \hat{\eta}\right)\right] = 0. \tag{17}$$

The sample-splitting procedure itself also introduces additional uncertainty. Therefore, the above procedure is repeated a number of times B with different random splits. The final estimator is then put together via the median method suggested by Chernozhukov et al. (2018). The final estimate point estimate is the median of estimates for each split,

$$\hat{\theta}^{\text{median}} = \text{median}\left\{\hat{\theta}_b\right\}_{b=1}^{B}. \tag{18}$$

The final variance estimator takes into account the variation introduced by sample splitting:

$$\hat{\sigma}^{2,\text{median}} = \text{median}\left\{\hat{\sigma}_b^2 + \left(\hat{\theta}_b - \hat{\theta}^{\text{median}}\right)^2\right\}_{b=1}^{B}. \tag{19}$$

In this paper, $\hat{\sigma}_b^2$ is obtained via a wild cluster bootstrap over the values of $\psi_\theta\left(W; \hat{\theta}_b, \hat{\eta}_b\right)$.

A.4 Inference for the model with attrition

The following theorem parallels Theorem 5.1 in Chernozhukov et al. (2018), stating that the DML estimators of the ATE and the ATN for the model with attrition are approximately unbiased and asymptotically normal.

Expectation and probability operators as well as norms are always with respect to a probability measure P of the data $W = (Y, D, R, X)$. I use $\|\cdot\|_q$ to denote the $L^q(P)$ norm, and for nuisance functions $\eta = (\ell_1, \ldots, \ell_l)$, denote $\|\eta\|_q := \max_{1 \le j \le l} \|\ell_j\|_q$. Let $(\delta_n)_{n=1}^{\infty}$ and $(\Delta_n)_{n=1}^{\infty}$ be sequences of positive constants approaching 0, and let ε, c, C, C', and q be positive constants, with $q > 2$.

Theorem. Assume that the following conditions hold: (a) equations (10)–(12) hold; (b) $\|Y\|_q \le C$; (c) $\Pr(\varepsilon \le q_0(D, X) \le 1 - \varepsilon) = 1$; (d) $\Pr(\varepsilon \le r_0(D, X)) = 1$; (e) $\|RU\|_2 \ge c$; (f) $\left\|\mathrm{E}\left[U^2|X\right]\right\|_\infty \le C$; (g) for subset I of $[N]$ of size n, $\eta = \eta\left((W_i)_{i \in I}\right) \in \mathcal{T}_N$ with P-probability no less than $1 - \Delta_N$, where the realization set $\mathcal{T}_N$ is a shrinking neighborhood of η_0 containing all the nuisance parameter estimates η that obey the following conditions: $\|\eta - \eta_0\|_q \le C$, $\|\eta - \eta_0\|_2 \le \delta_N$, $\max\left\{\left\|m - \frac{1}{2}\right\|_\infty, \left\|q - \frac{1}{2}\right\|_\infty\right\} \le \frac{1}{2} - \varepsilon$, $\|r\|_\infty \ge \varepsilon$, and $\|g - g_0\|_2 \times \left(\|m - m_0\|_2 + \|r - r_0\|_2 + \|q - q_0\|_2\right) \le \delta_N N^{-1/2}$. Then, the DML estimators for the ATE and ATN constructed above, $\hat{\theta}_0$ and $\hat{\gamma}_0$, obey $\sqrt{N}\left(\hat{\theta}_0 - \theta_0\right) \rightsquigarrow \mathcal{N}\left(0, \sigma_\theta^2\right)$ with $\sigma_\theta^2 = \mathrm{E}\left[\varphi_\theta^2(W; \theta_0, \eta_0)\right]$, and $\sqrt{N}(\hat{\gamma}_0 - \gamma_0) \rightsquigarrow \mathcal{N}\left(0, \sigma_\gamma^2\right)$ with $\sigma_\gamma^2 = \mathrm{E}\left[\varphi_\gamma^2(W; \gamma_0, \eta_0)\right]$.

Proof: The proof follows along the same lines as the one given in Chernozhukov et al. (2018, pp. 65–68) for the model without attrition, and is therefore being omitted here. It is available on request.

A.5 Data preparation and hyperparameter tuning

For the random forest, boosted trees, and SVM, the dictionary of considered controls encompasses all the variables listed in Appendix Table 1, with categorical variables expanded into dummy variables. For the regularized regression techniques, I use an extended set of candidate variables. Next to the variables already mentioned, it includes squared and cubed terms of all numerical variables, and cubic B-splines with five interior knots of three continuous variables—two household poverty indices and a village-level poverty index. Last of all, it includes interactions of all the previously mentioned variables with a subset of 28 variables that are deemed particularly relevant; these include student characteristics, household demographics, parents' education and expectations, local wages paid to children, and poverty levels. Missing values are treated as follows: for categorical variables, a new *missing* category is created. For numerical variables, missing entries are assigned the average of all non-missing entries, and an additional *missing* dummy is created. Extreme gradient boosting is the only method which does not require missing value imputation. After dropping duplicates and perfectly collinear variables, the basic dictionary of variables for the eligible students in the conditional sample includes 222 variables, whereas the extended dictionary includes 16,395 variables. The sets of variables for the non-eligible students, as well as for eligible and non-eligible students in the unconditional sample, are very similar in magnitude.

To obtain the best possible prediction model for each machine learning method and nuisance function, a number of hyperparameters need to be selected. For the regularized regression techniques, these include the ℓ_1 and ℓ_2 regularization parameters. For SVM, they are the cost as well as the parameter γ for the radial basis kernel. For boosting with logistic regression trees, the parameters are number of boosting iterations, learning rate, maximal tree depth, minimum loss reduction, subsample ratio of training set observations, subsample ratio of variables, and minimum sum of instance weight per leaf. For random forests, since overfitting is not a concern, I go without hyperparameter tuning and simply choose a large enough number of trees (1,000) and a leaf size of 1.

For each nuisance function and method, I create a grid with likely values for the hyperparameters and run a repeated cross-validation. To that end, the dataset is split in the same way as for the cross-fitting procedure, i.e., in 10 folds, with roughly equal ratios of treated observations, and with no overlapping locations. For each hyperparameter vector from the grid, the model is tuned in 9 folds and predictions made in the remaining fold. This is done separately for all 10 folds, and repeated a total of 10 times for different random splits. In the end, the hyperparameter vector with the lowest average out-of-sample mean squared error is selected.

A.6 Omitted variable bias estimation

In this section, I compute an estimator for the omitted variable bias by Oster (2017). Its basis is an outcome model of the form

$$Y = D\theta + Z_1 + Z_2 + \epsilon, \tag{20}$$

with an observed part $Z_1 = X\beta$ and an unobserved part Z_2. This model is more restrictive than the heterogeneous treatment effect model, most notably since both treatment status D and observables X enter the outcome model linearly. This means that the results presented here should merely be seen as back-of-the-envelope calculations. Let $\hat{\theta}$ be the point estimate of the estimation excluding the unobservable part, and let $\mathring{\theta}$ be the point estimate of a simple regression of Y on D. Let $\mathring{R}$, $\hat{R}$, and $R_{\max}$ respectively denote the R-squared of the simple regression model, of the model with observed characteristics, and of the (hypothetical) outcome model with all observed and unobserved pre-treatment variables included. Furthermore, let δ denote the relative importance of selection on the observed and the unobserved part of the model, i.e.,

$$\delta \frac{\operatorname{cov}(Z_1, D)}{\operatorname{var}(Z_1)} = \frac{\operatorname{cov}(Z_2, D)}{\operatorname{var}(Z_2)}.$$

The omitted variable bias Π is then estimated as

$$\hat{\Pi} = \delta \left(\mathring{\theta} - \hat{\theta}\right) \left(R_{\max} - \hat{R}\right) / \left(\hat{R} - \mathring{R}\right). \tag{21}$$

Among the components in equation (21), $\hat{\theta}$, $\mathring{\theta}$, $\hat{R}$, and $\mathring{R}$ are observed, while $R_{\max}$ and δ are not. $R_{\max}$ represents the maximal R-squared achievable in a prediction of Y using pre-treatment information, a value bounded from above by 1. It is reasonable to expect $R_{\max}$ to be lower than 1, for the following two reasons. First, all outcomes are measured years after the time before treatment. Therefore, idiosyncratic shocks can occur after the beginning of the treatment, which affect outcomes but are by definition unpredictable using pre-treatment information. Second, any measurement error in Y reduces predictability. For this exercise, I chose multiples of $\hat{R}$ as possible values for $R_{\max}$, with multipliers 2, 3, 4, and 5. I then calculate the value of δ for which the true treatment effect would be zero, or $\hat{\Pi} = \hat{\theta}$. A value for δ that is far from 0 is unlikely—given the effort to include all covariates that are highly correlated with D—and in turn means a null result is unlikely. Oster argues that $\delta = 1$ is an appropriate cutoff value, as this implies that unobservables are as important as observables. I adopt this argument, with the qualification that $\operatorname{cov}(Z_1, D)$ and $\operatorname{cov}(Z_2, D)$ may have opposite signs. This means that for any given $R_{\max}$, $\theta = 0$ is rejected if $|\delta| > 1$.

Table 7 shows the δ implied by $\theta = 0$, where θ denotes the true average treatment effect, for $R_{\max} = \min(i, 1)$ with $i = 2, \ldots, 5$ and all samples and school-related outcomes discussed in the paper. It can be seen that for all results that are at least significant at the 10% level, $|\delta| > 1$ even if the unobserved variables explain three times as much as the observed ones, i.e., if $R_{\max} = 4\hat{R}$. A small caveat here is that the values of $\mathring{\theta}$ do not take account of observations with missing outcomes, whereas the values of $\hat{\theta}$ do. Thus, for samples with some outcomes missing, the estimates of δ are likely still too close to 0. For all those samples without missing observations, $|\delta| > 1$ even if $R_{\max} = 5\hat{R}$. In conclusion, given the very low predictability of treatment status, even a significant amount of unobserved pre-treatment information is likely not going to invalidate the main results.

Table 7: Relative importance of selection on the observed and the unobserved

	eligible				non-eligible			
$R_{\max} = \min(\cdot, 1)$	$2\hat{R}$	$3\hat{R}$	$4\hat{R}$	$5\hat{R}$	$2\hat{R}$	$3\hat{R}$	$4\hat{R}$	$5\hat{R}$
Unconditional sample								
Going to high school in November 2000	-12.42	-6.21	-4.14	-3.11	-58.0	-29.0	-19.3	-14.5
Some high school by 2003	-8.57	-4.28	-2.86	-2.14	-8.48	-4.24	-2.83	-2.12
Graduated or about to graduate from high school in 2003	125.3	62.7	41.8	31.3	9.77	4.88	3.26	2.44
Graduated from middle school by 2000	-2.28	-1.14	-0.76	-0.71	-143.0	-71.5	-47.7	-35.7
Graduated or about to graduate from middle school in 2000	-63.9	-31.94	-21.3	-18.4	1.58	0.79	0.53	0.47
Conditional unrestricted sample								
Going to high school in November 2000	-6.29	-3.15	-2.10	-1.57	-10.2	-5.11	-3.41	-2.56
Some high school by 2003	-3.80	-1.90	-1.27	-0.95	-3.94	-1.97	-1.31	-0.98
Graduated or about to graduate from high school in 2003	-3.45	-1.73	-1.15	-0.86	-3.18	-1.59	-1.06	-0.79
Conditional restricted sample								
Going to high school in November 2000	-10.6	-5.29	-3.53	-2.64	5.91	2.96	1.97	1.48
Some high school by 2003	-2.22	-1.11	-0.74	-0.56	-51.4	-25.7	-17.1	-12.9
Graduated or about to graduate from high school in 2003	-2.47	-1.24	-0.82	-0.62	249.6	124.8	83.2	62.4

A.7 Result Tables

Table 8: Detailed estimates of conditional ATE and ATN on high school education for adolescents from eligible households; unrestricted sample

	Lasso	Ridge	Elastic net	Random forest	Boosted trees	SVM	Best Method
(1) Dependent variable: going to high school in November 2000.							
ATE	-0.124**	-0.107*	-0.127**	-0.123**	-0.107	-0.127*	-0.125**
	(0.055)	(0.064)	(0.054)	(0.062)	(0.065)	(0.077)	(0.055)
ATN	-0.111**	-0.106	-0.110**	-0.116*	-0.113*	-0.109	-0.107*
	(0.056)	(0.065)	(0.055)	(0.061)	(0.064)	(0.084)	(0.057)
(2) Dependent variable: some high school by 2003.							
ATE	-0.101*	-0.076	-0.101*	-0.100	-0.070	-0.077	-0.101*
	(0.060)	(0.065)	(0.059)	(0.065)	(0.089)	(0.065)	(0.059)
ATN	-0.061	-0.076	-0.064	-0.068	-0.041	-0.046	-0.067
	(0.068)	(0.065)	(0.067)	(0.064)	(0.082)	(0.069)	(0.066)
(3) Dependent variable: graduated or about to graduate from high school in 2003.							
ATE	-0.092	-0.074	-0.099*	-0.114*	-0.058	-0.075	-0.100*
	(0.057)	(0.063)	(0.056)	(0.064)	(0.082)	(0.063)	(0.056)
ATN	-0.082	-0.073	-0.086	-0.091	-0.051	-0.065	-0.091
	(0.059)	(0.064)	(0.059)	(0.064)	(0.079)	(0.068)	(0.058)

*: significant at the 10% level. **: significant at the 5% level. ***: significant at the 1% level.
Average treatment effects (ATE) and average treatment effects of the non-treated (ATN) as well as standard errors (in parentheses) are estimated for adolescents from eligible households in the unrestricted conditional sample. That includes adolescents whose highest completed grade by November 2000 was the third grade of middle school, and who had been enrolled in school during the entire academic year 1999/2000. Column labels denote the method used to estimate the nuisance functions.
Point estimates for the first outcome are obtained via the orthogonal estimating equations (13) and (14), while their respective modifications for attrition (15) and (16) are employed for the other two outcomes, as these are unobserved for some adolescents. Standard errors are obtained via wild cluster bootstrap. Results are based on 10-fold cross-fitting, using 100 sample splits. Estimates from the sample splits are combined using the median method, based on equations (18) and (19).

Table 9: Detailed estimates of conditional ATE and ATN on high school education for adolescents from eligible households; restricted sample

	Lasso	Ridge	Elastic net	Random forest	Boosted trees	SVM	Best Method
(1) Dependent variable: going to high school in November 2000.							
ATE	-0.141**	-0.135*	-0.146**	-0.166**	-0.171**	-0.147**	-0.145**
	(0.069)	(0.073)	(0.069)	(0.076)	(0.074)	(0.070)	(0.069)
ATN	-0.124*	-0.135*	-0.139**	-0.155**	-0.165**	-0.140**	-0.135*
	(0.070)	(0.073)	(0.069)	(0.079)	(0.082)	(0.070)	(0.070)
(2) Dependent variable: some high school by 2003.							
ATE	-0.119	-0.113	-0.122	-0.172**	-0.106	-0.144*	-0.126
	(0.078)	(0.078)	(0.078)	(0.084)	(0.158)	(0.082)	(0.078)
ATN	-0.120	-0.114	-0.127	-0.125	-0.219	-0.145*	-0.126
	(0.078)	(0.079)	(0.078)	(0.087)	(0.144)	(0.081)	(0.077)
(3) Dependent variable: graduated or about to graduate from high school in 2003.							
ATE	-0.092	-0.086	-0.091	-0.146*	-0.046	-0.119	-0.098
	(0.076)	(0.077)	(0.077)	(0.085)	(0.174)	(0.081)	(0.076)
ATN	-0.092	-0.086	-0.092	-0.096	-0.162	-0.114	-0.099
	(0.076)	(0.077)	(0.077)	(0.086)	(0.153)	(0.078)	(0.076)

*: significant at the 10% level. **: significant at the 5% level. ***: significant at the 1% level.
Average treatment effects (ATE) and average treatment effects of the non-treated (ATN) as well as standard errors (in parentheses) are estimated for adolescents from eligible households in the restricted conditional sample. That includes adolescents whose highest completed grade by November 1999 was the second grade of middle school, by November 2000 the third grade of middle school, and who had been enrolled in school during the entire academic year 1999/2000. Column labels denote the method used to estimate the nuisance functions.
Point estimates for the first outcome are obtained via the orthogonal estimating equations (13) and (14), while their respective modifications for attrition (15) and (16) are employed for the other two outcomes, as these are unobserved for some adolescents. Standard errors are obtained via wild cluster bootstrap. Results are based on 10-fold cross-fitting, using 100 sample splits. Estimates from the sample splits are combined using the median method, based on equations (18) and (19).

Table 10: Detailed estimates of conditional ATE and ATN on high school education for adolescents from non-eligible households; unrestricted sample

	Lasso	Ridge	Elastic net	Random forest	Boosted trees	SVM	Best Method
(1) Dependent variable: going to high school in November 2000.							
ATE	0.064	0.023	0.059	0.090	0.104*	0.116	0.066
	(0.056)	(0.062)	(0.055)	(0.064)	(0.063)	(0.075)	(0.055)
ATN	0.078	0.060	0.077	0.098*	0.113*	0.115	0.089
	(0.054)	(0.061)	(0.054)	(0.056)	(0.062)	(0.078)	(0.056)
(2) Dependent variable: some high school by 2003.							
ATE	0.122*	0.086	0.120*	0.087	0.103	0.089	0.109
	(0.071)	(0.075)	(0.071)	(0.079)	(0.089)	(0.080)	(0.072)
ATN	0.124*	0.086	0.121*	0.117	0.135	0.093	0.118
	(0.070)	(0.074)	(0.070)	(0.075)	(0.091)	(0.074)	(0.074)
(3) Dependent variable: graduated or about to graduate from high school in 2003.							
ATE	0.169***	0.105	0.165***	0.128*	0.162*	0.142**	0.149**
	(0.064)	(0.067)	(0.064)	(0.070)	(0.086)	(0.072)	(0.064)
ATN	0.167**	0.104	0.161**	0.132**	0.162	0.123*	0.141**
	(0.066)	(0.069)	(0.065)	(0.067)	(0.109)	(0.070)	(0.071)

*: significant at the 10% level. **: significant at the 5% level. ***: significant at the 1% level.
Average treatment effects (ATE) and average treatment effects of the non-treated (ATN) as well as standard errors (in parentheses) are estimated for adolescents from non-eligible households in the unrestricted conditional sample. That includes adolescents whose highest completed grade by November 2000 was the third grade of middle school, and who had been enrolled in school during the entire academic year 1999/2000. Column labels denote the method used to estimate the nuisance functions.
Point estimates for the first outcome are obtained via the orthogonal estimating equations (13) and (14), while their respective modifications for attrition (15) and (16) are employed for the other two outcomes, as these are unobserved for some adolescents. Standard errors are obtained via wild cluster bootstrap. Results are based on 10-fold cross-fitting, using 100 sample splits. Estimates from the sample splits are combined using the median method, based on equations (18) and (19).

Table 11: Detailed estimates of conditional ATE and ATN on high school education for adolescents from non-eligible households; restricted sample

	Lasso	Ridge	Elastic net	Random forest	Boosted trees	SVM	Best Method
(1) Dependent variable: going to high school in November 2000.							
ATE	0.106	0.108	0.102	0.093	0.007	0.090	0.093
	(0.073)	(0.076)	(0.072)	(0.074)	(0.117)	(0.078)	(0.073)
ATN	0.103	0.108	0.110	0.112	0.001	0.131	0.110
	(0.074)	(0.077)	(0.074)	(0.075)	(0.139)	(0.081)	(0.075)
(2) Dependent variable: some high school by 2003.							
ATE	0.181**	0.161*	0.169**	0.152*	0.133	0.157*	0.162**
	(0.084)	(0.085)	(0.083)	(0.083)	(0.195)	(0.087)	(0.080)
ATN	0.182**	0.161*	0.177**	0.186**	0.204	0.151*	0.184**
	(0.089)	(0.086)	(0.088)	(0.089)	(0.160)	(0.090)	(0.084)
(3) Dependent variable: graduated or about to graduate from high school in 2003.							
ATE	0.216***	0.194**	0.209***	0.182**	0.180	0.193**	0.193**
	(0.079)	(0.080)	(0.079)	(0.078)	(0.203)	(0.080)	(0.078)
ATN	0.212**	0.195**	0.213**	0.200**	0.220	0.177**	0.191**
	(0.082)	(0.081)	(0.084)	(0.084)	(0.169)	(0.088)	(0.086)

*: significant at the 10% level. **: significant at the 5% level. ***: significant at the 1% level.
Average treatment effects (ATE) and average treatment effects of the non-treated (ATN) as well as standard errors (in parentheses) are estimated for adolescents from non-eligible households in the restricted conditional sample. That includes adolescents whose highest completed grade by November 1999 was the second grade of middle school, by November 2000 the third grade of middle school, and who had been enrolled in school during the entire academic year 1999/2000. Column labels denote the method used to estimate the nuisance functions.
Point estimates for the first outcome are obtained via the orthogonal estimating equations (13) and (14), while their respective modifications for attrition (15) and (16) are employed for the other two outcomes, as these are unobserved for some adolescents. Standard errors are obtained via wild cluster bootstrap. Results are based on 10-fold cross-fitting, using 100 sample splits. Estimates from the sample splits are combined using the median method, based on equations (18) and (19).

Table 12: Detailed estimates of unconditional ATE on high school education for adolescents from eligible households

	Lasso	Ridge	Elastic net	Random forest	Boosted trees	SVM	Best Method
(1) Dependent variable: going to high school in November 2000.							
ATE	-0.033	-0.030	-0.033	-0.042	-0.031	-0.009	-0.033
	(0.030)	(0.031)	(0.030)	(0.031)	(0.050)	(0.033)	(0.030)
(2) Dependent variable: some high school by 2003.							
ATE	-0.091**	-0.075	-0.089**	-0.080*	-0.067	-0.074	-0.085**
	(0.044)	(0.047)	(0.044)	(0.043)	(0.041)	(0.056)	(0.043)
(3) Dependent variable: graduated or about to graduate from high school in 2003.							
ATE	-0.046	-0.050	-0.045	-0.062	-0.046	-0.041	-0.050
	(0.036)	(0.040)	(0.036)	(0.039)	(0.034)	(0.047)	(0.035)

*: significant at the 10% level. **: significant at the 5% level. ***: significant at the 1% level.
Average treatment effects (ATE) and average treatment effects of the non-treated (ATN) as well as standard errors (in parentheses) are estimated for adolescents from eligible households in the unconditional sample. That includes adolescents who had graduated from primary school in 1997. Column labels denote the method used to estimate the nuisance functions.
Point estimates are obtained via the orthogonal estimating equations that account for attrition, (15) and (16). Standard errors are obtained via wild cluster bootstrap. Results are based on 10-fold cross-fitting, using 100 sample splits. Estimates from the sample splits are combined using the median method, based on equations (18) and (19).

Table 13: Detailed estimates of unconditional ATE on high school education for adolescents from non-eligible households

	Lasso	Ridge	Elastic net	Random forest	Boosted trees	SVM	Best Method
(1) Dependent variable: going to high school in November 2000.							
ATE	0.124***	0.123***	0.124***	0.127***	0.145*	0.139***	0.129***
	(0.034)	(0.036)	(0.034)	(0.034)	(0.077)	(0.038)	(0.035)
(2) Dependent variable: some high school by 2003.							
ATE	0.093*	0.115**	0.089*	0.094**	0.087	0.091	0.100**
	(0.048)	(0.050)	(0.047)	(0.046)	(0.068)	(0.061)	(0.044)
(3) Dependent variable: graduated or about to graduate from high school in 2003.							
ATE	0.101**	0.135***	0.098**	0.105***	0.108*	0.101*	0.107***
	(0.042)	(0.042)	(0.041)	(0.040)	(0.059)	(0.056)	(0.038)

*: significant at the 10% level. **: significant at the 5% level. ***: significant at the 1% level.
Average treatment effects (ATE) and average treatment effects of the non-treated (ATN) as well as standard errors (in parentheses) are estimated for adolescents from non-eligible households in the unconditional sample. That includes adolescents who had graduated from primary school in 1997.
Column labels denote the method used to estimate the nuisance functions.
Point estimates are obtained via the orthogonal estimating equations that account for attrition, (15) and (16). Standard errors are obtained via wild cluster bootstrap. Results are based on 10-fold cross-fitting, using 100 sample splits. Estimates from the sample splits are combined using the median method, based on equations (18) and (19).

Table 14: Detailed estimates of unconditional ATE on middle school education for adolescents from eligible households

	Lasso	Ridge	Elastic net	Random forest	Boosted trees	SVM	Best Method
(1) Dependent variable: graduated from middle school by 2000.							
ATE	-0.014	-0.025	-0.013	-0.039	-0.072	-0.043	-0.029
	(0.032)	(0.037)	(0.032)	(0.036)	(0.064)	(0.049)	(0.032)
(2) Dependent variable: graduated or about to graduate from middle school in 2000.							
ATE	0.089***	0.087**	0.090***	0.068*	0.031	0.052	0.088***
	(0.031)	(0.037)	(0.031)	(0.035)	(0.069)	(0.045)	(0.032)

*: significant at the 10% level. **: significant at the 5% level. ***: significant at the 1% level.
Average treatment effects (ATE) and average treatment effects of the non-treated (ATN) as well as standard errors (in parentheses) are estimated for adolescents from eligible households in the unconditional sample. That includes adolescents who had graduated from primary school in 1997. Column labels denote the method used to estimate the nuisance functions.
Point estimates are obtained via the orthogonal estimating equations that account for attrition, (15) and (16). Standard errors are obtained via wild cluster bootstrap. Results are based on 10-fold cross-fitting, using 100 sample splits. Estimates from the sample splits are combined using the median method, based on equations (18) and (19).

Table 15: Detailed estimates of unconditional ATE on middle school education for adolescents from non-eligible households

	Lasso	Ridge	Elastic net	Random forest	Boosted trees	SVM	Best Method
(1) Dependent variable: graduated from middle school by 2000.							
ATE	0.063	0.066	0.068	0.079*	0.114	0.094	0.069
	(0.044)	(0.050)	(0.044)	(0.043)	(0.088)	(0.051)	(0.044)
(2) Dependent variable: graduated or about to graduate from middle school in 2000.							
ATE	0.023	0.042	0.024	0.055	0.092	0.086*	0.026
	(0.036)	(0.045)	(0.036)	(0.039)	(0.081)	(0.049)	(0.037)

*: significant at the 10% level. **: significant at the 5% level. ***: significant at the 1% level.
Average treatment effects (ATE) and average treatment effects of the non-treated (ATN) as well as standard errors (in parentheses) are estimated for adolescents from non-eligible households in the unconditional sample. That includes adolescents who had graduated from primary school in 1997. Column labels denote the method used to estimate the nuisance functions.
Point estimates are obtained via the orthogonal estimating equations that account for attrition, (15) and (16). Standard errors are obtained via wild cluster bootstrap. Results are based on 10-fold cross-fitting, using 100 sample splits. Estimates from the sample splits are combined using the median method, based on equations (18) and (19).

Table 16: Detailed estimates of conditional ATE and ATN on further outcomes for adolescents from eligible households; unrestricted sample

	Lasso	Ridge	Elastic net	Random forest	Boosted trees	SVM	Best Method
(1) Dependent variable: not going to high school due to financial constraints.							
ATE	0.045	0.032	0.048	0.031	0.031	0.050	0.035
	(0.054)	(0.057)	(0.053)	(0.058)	(0.061)	(0.081)	(0.053)
ATN	0.036	0.031	0.036	0.002	0.017	0.015	0.019
	(0.057)	(0.057)	(0.056)	(0.060)	(0.073)	(0.093)	(0.055)
(2) Dependent variable: not going to high school for other than financial reasons.							
ATE	0.071*	0.075*	0.073*	0.079*	0.075	0.075	0.073*
	(0.043)	(0.042)	(0.042)	(0.045)	(0.047)	(0.064)	(0.042)
ATN	0.074*	0.075*	0.074*	0.089*	0.077	0.098	0.074*
	(0.044)	(0.044)	(0.044)	(0.051)	(0.052)	(0.083)	(0.044)
(3) Dependent variable: log monthly per capita expenditure.							
ATE	0.051	0.059	0.051	0.051	0.043	0.060	0.059
	(0.066)	(0.068)	(0.066)	(0.071)	(0.070)	(0.066)	(0.063)
ATN	0.052	0.058	0.052	0.043	0.045	0.060	0.051
	(0.063)	(0.068)	(0.063)	(0.068)	(0.073)	(0.066)	(0.063)

*: significant at the 10% level. **: significant at the 5% level. ***: significant at the 1% level.
Average treatment effects (ATE) and average treatment effects of the non-treated (ATN) as well as standard errors (in parentheses) are estimated for adolescents from eligible households in the unrestricted conditional sample. That includes adolescents whose highest completed grade by November 2000 was the third grade of middle school, and who had been enrolled in school during the entire academic year 1999/2000. Column labels denote the method used to estimate the nuisance functions.
Point estimates are obtained via the orthogonal estimating equations (13) and (14). Standard errors are obtained via wild cluster bootstrap. Results are based on 10-fold cross-fitting, using 100 sample splits. Estimates from the sample splits are combined using the median method, based on equations (18) and (19).

Table 17: Detailed estimates of conditional ATE and ATN on further outcomes for adolescents from eligible households; restricted sample

	Lasso	Ridge	Elastic net	Random forest	Boosted trees	SVM	Best Method
(1) Dependent variable: not going to high school due to financial constraints.							
ATE	0.031	0.038	0.031	0.040	0.069	0.038	0.031
	(0.064)	(0.065)	(0.064)	(0.068)	(0.072)	(0.065)	(0.064)
ATN	0.044	0.039	0.044	0.035	0.059	0.038	0.044
	(0.064)	(0.065)	(0.064)	(0.074)	(0.093)	(0.065)	(0.064)
(2) Dependent variable: not going to high school for other than financial reasons.							
ATE	0.109**	0.097*	0.114**	0.106*	0.082	0.104*	0.113**
	(0.052)	(0.054)	(0.051)	(0.055)	(0.059)	(0.054)	(0.051)
ATN	0.103*	0.096*	0.103*	0.090	0.040	0.087	0.103*
	(0.056)	(0.055)	(0.056)	(0.063)	(0.085)	(0.058)	(0.056)
(3) Dependent variable: log monthly per capita expenditure.							
ATE	0.019	0.017	0.016	0.020	0.031	0.018	0.019
	(0.078)	(0.080)	(0.079)	(0.086)	(0.082)	(0.080)	(0.077)
ATN	0.023	0.017	0.023	-0.016	0.026	0.018	0.016
	(0.076)	(0.079)	(0.076)	(0.082)	(0.082)	(0.079)	(0.075)

*: significant at the 10% level. **: significant at the 5% level. ***: significant at the 1% level.
Average treatment effects (ATE) and average treatment effects of the non-treated (ATN) as well as standard errors (in parentheses) are estimated for adolescents from eligible households in the restricted conditional sample. That includes adolescents whose highest completed grade by November 1999 was the second grade of middle school, by November 2000 the third grade of middle school, and who had been enrolled in school during the entire academic year 1999/2000. Column labels denote the method used to estimate the nuisance functions.
Point estimates are obtained via the orthogonal estimating equations (13) and (14). Standard errors are obtained via wild cluster bootstrap. Results are based on 10-fold cross-fitting, using 100 sample splits. Estimates from the sample splits are combined using the median method, based on equations (18) and (19).

Table 18: Detailed estimates of conditional ATE and ATN on further outcomes for adolescents from non-eligible households; unrestricted sample

	Lasso	Ridge	Elastic net	Random forest	Boosted trees	SVM	Best Method
(1) Dependent variable: not going to high school due to financial constraints.							
ATE	0.000	0.017	0.002	-0.013	-0.020	-0.001	0.002
	(0.059)	(0.061)	(0.058)	(0.061)	(0.063)	(0.077)	(0.057)
ATN	-0.021	0.017	-0.015	-0.031	-0.046	0.002	-0.014
	(0.057)	(0.061)	(0.056)	(0.053)	(0.059)	(0.079)	(0.055)
(2) Dependent variable: not going to high school for other than financial reasons.							
ATE	-0.087*	-0.078*	-0.087*	-0.076	-0.064	-0.106*	-0.087*
	(0.046)	(0.047)	(0.046)	(0.050)	(0.052)	(0.061)	(0.046)
ATN	-0.084*	-0.076	-0.085*	-0.075	-0.080	-0.120**	-0.085*
	(0.047)	(0.048)	(0.047)	(0.048)	(0.052)	(0.054)	(0.047)
(3) Dependent variable: log monthly per capita expenditure.							
ATE	-0.090	-0.107	-0.090	-0.067	-0.076	-0.105	-0.079
	(0.086)	(0.088)	(0.085)	(0.091)	(0.092)	(0.087)	(0.081)
ATN	-0.113	-0.107	-0.112	-0.084	-0.082	-0.107	-0.104
	(0.086)	(0.089)	(0.086)	(0.085)	(0.094)	(0.088)	(0.085)

*: significant at the 10% level. **: significant at the 5% level. ***: significant at the 1% level.
Average treatment effects (ATE) and average treatment effects of the non-treated (ATN) as well as standard errors (in parentheses) are estimated for adolescents from non-eligible households in the unrestricted conditional sample. That includes adolescents whose highest completed grade by November 2000 was the third grade of middle school, and who had been enrolled in school during the entire academic year 1999/2000. Column labels denote the method used to estimate the nuisance functions.
Point estimates are obtained via the orthogonal estimating equations (13) and (14). Standard errors are obtained via wild cluster bootstrap. Results are based on 10-fold cross-fitting, using 100 sample splits. Estimates from the sample splits are combined using the median method, based on equations (18) and (19).

Table 19: Detailed estimates of conditional ATE and ATN on further outcomes for adolescents from non-eligible households; restricted sample

	Lasso	Ridge	Elastic net	Random forest	Boosted trees	SVM	Best Method
(1) Dependent variable: not going to high school due to financial constraints.							
ATE	-0.018	-0.020	-0.013	-0.009	0.090	-0.012	-0.010
	(0.070)	(0.073)	(0.070)	(0.074)	(0.118)	(0.076)	(0.072)
ATN	-0.015	-0.021	-0.019	-0.038	0.052	-0.039	-0.037
	(0.068)	(0.073)	(0.070)	(0.069)	(0.144)	(0.077)	(0.069)
(2) Dependent variable: not going to high school for other than financial reasons.							
ATE	-0.088	-0.088	-0.088	-0.091	-0.095	-0.086	-0.085
	(0.061)	(0.061)	(0.060)	(0.062)	(0.098)	(0.063)	(0.060)
ATN	-0.087	-0.087	-0.088	-0.083	-0.080	-0.091	-0.085
	(0.061)	(0.061)	(0.061)	(0.060)	(0.108)	(0.064)	(0.062)
(3) Dependent variable: log monthly per capita expenditure.							
ATE	-0.133	-0.160	-0.138	-0.103	-0.132	-0.162	-0.113
	(0.098)	(0.101)	(0.098)	(0.102)	(0.166)	(0.101)	(0.101)
ATN	-0.145	-0.158	-0.148	-0.139	-0.148	-0.159	-0.139
	(0.097)	(0.104)	(0.097)	(0.098)	(0.205)	(0.105)	(0.096)

*: significant at the 10% level. **: significant at the 5% level. ***: significant at the 1% level.
Average treatment effects (ATE) and average treatment effects of the non-treated (ATN) as well as standard errors (in parentheses) are estimated for adolescents from non-eligible households in the restricted conditional sample. That includes adolescents whose highest completed grade by November 1999 was the second grade of middle school, by November 2000 the third grade of middle school, and who had been enrolled in school during the entire academic year 1999/2000. Column labels denote the method used to estimate the nuisance functions.
Point estimates are obtained via the orthogonal estimating equations (13) and (14). Standard errors are obtained via wild cluster bootstrap. Results are based on 10-fold cross-fitting, using 100 sample splits. Estimates from the sample splits are combined using the median method, based on equations (18) and (19).