EBERHARD KARLS
UNIVERSITY OF TÜBINGEN

Master Seminar in Econometrics

Heterogeneous Treatment Effect Estimation under Group-invariant Effects[1]

Roman Rakov
M.Sc. Economics and Finance (6294779)
roman.rakov@student.uni-tuebingen.de

Date: July 22, 2024

School of Business and Economics
Department of Statistics, Econometrics and Quantitative Methods

[1]A huge thank you to Marian Rümmele and Michael Knaus for answering questions and sharing their experience. The replication code can be found in the GitHub repository

Contents

List of Figures

List of Tables

1 Introduction

Machine learning (ML) methods in causal inference complement the existing econometric toolbox for program evaluation across multiple dimensions (Athey and Imbens, 2017). For one, they provide a data-driven approach to variable and model selection, while also offering a more comprehensive evaluation through flexible estimation methods.

This paper considers double machine learning (DML) (Chernozhukov et al., 2018) for program evaluation, highlighting its application in estimating heterogeneous treatment effects, including traditional subgroup effects, best linear prediction of effect heterogeneity, and non-parametric effect heterogeneity. DML can also be used to estimate optimal treatment assignment rules and exhibits favorable statistical properties. These methods can be combined with standard tools such as t-tests, OLS, kernel regression, and series regression (Knaus, 2022).

This paper starts by covering the necessary machinery of DML-based methods and proposes the demeaned augmented inverse probability weighting (AIPW) estimator as a potential solution for settings involving unobserved group-invariant effects (Wooldridge, 2010, ch. 10), and applies these methods in a standard labor economics setting.

It then discusses the potential issues of estimating heterogeneous treatment effects using conditional-variance-weighted estimators like fixed effects and conducts two Monte Carlo simulations to assess their finite sample properties. The paper concludes with a discussion of latent treatment effect heterogeneity and its implications for conditional average treatment effect (CATE) estimation.

Overall, the estimated average program effects align with the replicated study (Attanasio, 2011). DML-based methods suggest that while the effects of program participation are stronger for women, they are not significantly different compared to men. Treatment effects are higher for those with more education and less time spent working prior to the training. However, discovered treatment effect heterogeneity is limited due to the modest sample size.

Simulation results conclude that under identical effective treatments across groups, fixed effects and demeaned AIPW are unbiased and consistent estimators with favorable statistical properties. However, estimating CATE under multiple effective treatments grouped as "treated" or "not treated" can lack external validity and puzzle the identification process. In such cases, it remains unclear whether the observed effect heterogeneity under binary treatment reflects genuine differences in treatment effects or variations in the effective treatments themselves.

In the case of a binary treatment indicator, the underlying treatment heterogeneity should be explicitly discussed in applications, especially when interpreting heterogeneous effects.

2 Estimands of Interest

We describe the parameters of interest using Rubin's (1974) potential outcomes framework. The dummy variable W_i indicates a binary treatment, for example participation in a training program. Let $Y_i(1)$ denote the potential outcome (e.g., wage) when individual i ($i = 1, \ldots, N$) receives the treatment ($W_i = 1$). Conversely, $Y_i(0)$ denotes the potential outcome when individual i does not receive the treatment ($W_i = 0$). Each individual receives either the treatment or the control, but not both simultaneously. Consequently, only one of the potential outcomes is observed, illustrating the fundamental problem of

causal inference (Holland, 1986). The observed outcome can be written as

$$Y_i = Y_i(1)W_i + Y_i(0)(1 - W_i). \tag{1}$$

And the causal effect of W on Y for individual i is

$$\tau_i = Y_i(1) - Y_i(0).$$

Since the counterfactual potential outcome is missing, the individual treatment effect (ITE) τ_i is never observed. However, its conditional expectations can be identified under plausible assumptions. This paper focuses on conditional average treatment effects (CATEs), which are the expectations of τ_i conditional on exogenous pre-treatment covariates X_i.

Given the various ways τ_i can be aggregated, we distinguish two types of CATEs:

- **Individualized ATE (IATE)**: the finest conditioning level that uses all covariates available to the researcher in a given application:

  $$\tau(x) = E[\tau_i \mid X_i = x] = E[Y_i(1) - Y_i(0) \mid X_i = x].$$

  Aggregating them, for instance, allows for a Classification Analysis (CLAN) to compare covariate values of groups sorted by CATE. This helps identify what describes subgroups with the strongest and weakest estimated treatment effects.

- **Group ATE (GATE)** provide the average effects for pre-specified, usually low-dimensional, groups:

  $$\tau(g) = E[\tau_i \mid G_i = g] = E[Y_i(1) - Y_i(0) \mid G_i = g]$$

  This covers standard subgroup analysis comparing, e.g., effects of men and women, or heterogeneity along pre-specified continuous variables like age or income. These summaries can be valuable for policy analysis, particularly when examining effects within specific recipient groups.

Identifying treatment effects in observational studies is challenging due to non-random treatment assignment. However, it is feasible to identify the individual average treatment effect (IATE) and broader aggregation levels under standard assumptions (Imbens and Rubin, 2015):

Assumption 1

(a) *Measured confounding*: $Y_i(w) \perp\!\!\!\perp W_i \mid X_i = x$, for all $w \in W$ and $x \in X$.

(b) *Common Support*: $0 < P[W_i = w \mid X_i = x] \equiv e_w(x) < 1$, for all $w \in W$ and $x \in X$.

(c) *Stable Unit Treatment Value Assumption (SUTVA)*: $Y_i = Y_i(W_i)$.

The measured confounding assumption requires that X_i contains all confounding variables that jointly affect program assignment and the outcome. Common support states that it must be possible to observe each individual in all programs. SUTVA excludes spillover effects between treated and non-treated and ensures that each unit's potential outcome is uniquely determined by the specific treatment it receives.

These assumptions allow the identification of the average potential outcome (APO) conditional on confounders in three common ways (Knaus, 2022):

$$E[Y_i(w) \mid X_i = x] = E[Y_i \mid W_i = w, X_i = x] \equiv m(w, x) \tag{2.1}$$

$$= E\left[\frac{\mathbb{1}[W_i = w]Y_i}{e_w(x)}\middle| X_i = x\right] \tag{2.2}$$

$$= E\Bigg[\underbrace{m(w, x) + \frac{\mathbb{1}[W_i = w](Y_i - m(w, x))}{e_w(x)}}_{=\Gamma(w,x)}\Bigg| X_i = x\Bigg], \tag{2.3}$$

where $e_w(x) = P[W_i = 1 \mid X_i = x]$ is the conditional treatment probability, also known as the propensity score.

Equation (2.1) shows that the conditional APO is identified as a conditional expectation of the observed outcome. Equation (2.2) shows that it is identified by reweighting the observed outcome with the inverse treatment probability. Finally, equation (2.3) adds the reweighted outcome residual to the conditional outcome representation of equation (2.1). This might initially appear redundant, as we can verify that the reweighted residual has an expectation of zero under measured confounding. However, this identification result is doubly robust, meaning it remains valid even if we replace either $m(w, x)$ or $e_w(x)$ in equation (2.3) by arbitrary functions of x. The doubly robust structure will be discussed later.

From an identification perspective, $\Gamma(w, x)$ defined in equation (2.3) suffices to identify all estimands of interest[2]:

(a) APO: $\gamma_w = E[Y_i(w)] = E[\Gamma(w, X_i)]$

(b) ATE: $\tau_{ATE} = E[Y_i(1) - Y_i(0)] = E[\Gamma(1, X_i) - \Gamma(0, X_i)]$

(d) CATE: $\tau(x) = E[Y_i(1) - Y_i(0) \mid X_i = x] = E[\Gamma(1, X_i) - \Gamma(0, X_i) \mid X_i = x]$

2.1 Different Valleys for CATE Estimation

From a statistical perspective, estimation and inference for CATE are fundamentally more challenging than for average effects. This is because the causal parameter of interest represents either a function or the value of a function at a specific point. Two main approaches to estimation can be distinguished.

The first aims to construct an estimate $\hat{\tau}(x)$ of the true CATE function $\tau(x)$ by decomposing the estimation into a sequence of regression problems. Subsequently, generic machine learning (ML) techniques can be employed to address each of these regression problems.

This approach is referred to as meta-learning in the literature on CATE estimation, as it involves treating ML techniques as a black-box oracle capable of solving any regression problem and building upon this oracle to learn the CATE. Trying to recover the true CATE function is an exciting area of current research, but it also requires especially rich datasets to perform well.

Besides, methods for valid inference are still being developed. One possible approach is to forgo constructing confidence intervals on the CATE and instead prioritize achieving

[2]Step by step identification can be found in ATE estimation: AIPW-Double ML

high accuracy in the learned function, measured by MSE. In this scenario, the CATE problem is essentially treated as a prediction problem. To compensate for the absence of confidence intervals, hypothesis tests can be used as validation metrics to assess the quality of the CATE model as a whole. For a detailed exploration of meta-learning approaches for finer CATE estimation, refer to Chernozhukov (2024). Okasa (2022) provides insights on implementation nuances.

The second approach is the one pursued in this paper. Our goal is to estimate the Best Linear Approximation (BLA) of the CATE function using a predefined set of low-dimensional features, essentially focusing on estimating GATE. It allows to recover classical quantities of interest, including confidence intervals for the BLA at specific points, and simultaneous confidence bands for the BLA across a set of target evaluation points. Despite resulting in somewhat aggregated heterogeneities, this approach is the way to go to achieve statistical validity with a modest dataset.

2.2 Doubly Robust Scores

The construction of the doubly robust scores requires the input of so-called nuisance parameters that are usually of secondary interest and considered as tools to eventually obtain the parameters of interest (Knaus, 2022). In our case, the two nuisance parameters are $m(w, x) = E[Y_i \mid W_i = w, X_i = x]$, which is the conditional outcome mean for the subgroup observed in program w and the propensity score $e_w(x) = P[W_i = 1 \mid X_i = x]$.

Usually, these functions are unknown and need to be estimated. Following Chernozhukov (2024), it can be achieved using K-fold cross-fitting:

(a) randomly divide the sample into K folds of similar size;

(b) leave out fold k and estimate models for the nuisance parameters in the remaining $K - 1$ folds;

(c) use these models to predict $\hat{m}_{-k}(w, x)$ and $\hat{e}_{-k}(x)$ in the left-out fold k;

(d) repeat steps (a) to (c) such that each fold is left out once.

This procedure avoids overfitting in the sense that no observation is used to predict its own nuisance parameters.

The main building block of the following estimators is the doubly robust score of the Average Potential Outcome (APO), which replaces the true nuisance parameters in equation (2.3) by their cross-fitted predictions:

$$\widehat{\Gamma}_{i,w} = \hat{m}(w, x) + \frac{\mathbb{1}[W_i = w](Y_i - \hat{m}(w, x))}{\hat{e}_w(x)}$$

The estimated variances require no adjustment for the fact that we have estimated the nuisance parameters in the first step. The resulting estimators are consistent, asymptotically normal, and semiparametrically efficient under the main assumption that the estimators of the cross-fitted nuisance parameters are consistent and converge sufficiently fast (Chernozhukov, 2024).

The "doubly robustness" property in this context means that the parameters of interest are estimated at the parametric rate $n^{1/2}$ even if the nuisance parameters are estimated at slower rates using machine learning methods that do not require the specification of

an actual parametric model. For us, it is sufficient if both nuisance parameter estimators achieve $n^{1/4}$ or faster.

The rate double robustness is the consequence of the so-called Neyman orthogonality of the doubly robust score.[3] Scores with this orthogonality are immune to small errors in the estimation of nuisance parameters and thus allow them to be estimated via machine learning. For a more detailed treatment, refer to Knaus (2022).

By using the doubly robust (DR) augmented inverse probability weighting (AIPW) estimator, the ATE score is constructed as:

$$\widetilde{Y}_{ATE} = \widetilde{Y}_{\gamma_1} - \widetilde{Y}_{\gamma_0} = \underbrace{\hat{m}(1,X) - \hat{m}(0,X)}_{\text{outcome predictions}} + \underbrace{\frac{W(Y - \hat{m}(1,X))}{\hat{e}(X)} - \frac{(1-W)(Y - \hat{m}(0,X))}{1 - \hat{e}(X)}}_{\text{weighted residuals}}$$

And can be estimated as follows:

1. Form cross-fitted predictions of $\hat{m}(1,X)$, $\hat{m}(0,X)$, and $\hat{e}(X)$ using ML methods.
2. Estimate the outcome models $\widetilde{Y}\gamma_1$ and $\widetilde{Y}\gamma_0$.
3. Compute the pseudo-outcomes for each observation.
4. Estimate the ATE $\hat{\tau}_{ATE}^{AIPW}$ as the mean of pseudo-outcomes $\widetilde{Y}_{ATE}$.
5. Perform a t-test on the mean for hypothesis testing (no adjustments needed).

2.3 Optimal Treatment Assignment

The average effects presented above provide a comprehensive evaluation of programs under the current treatment assignment. However, many applications aim to conclude with recommendations to improve the assignment policy.

Let $\pi(X_i)$ denote a personalized treatment policy that assigns individuals to the treatment (or the absence thereof) based on their characteristics X_i. Formally, the function $\pi(X_i)$, given any instance X_i, maps observable characteristics to a program $\pi : X \to W$.

In the optimal scenario, we assign each individual to the treatment with the highest conditional average potential outcome, $E[Y_i(w) \mid X_i = x]$. However, candidate policy rules are often constrained to ensure interpretability, incorporate program costs or fairness considerations.

We continue to assume binary treatment, although the analysis can be extended to multiple treatments. Consider a stylized example[4] with two assignment rules π_1 and π_2 in Table 1.

i	$Y_i(0)$	$Y_i(1)$	π^1	π^2	$Y_i(\pi^1)$	$Y_i(\pi^2)$
1	$Y_1(0)$	$Y_1(1)$	0	0	$Y_1(0)$	$Y_1(0)$
2	$Y_2(0)$	$Y_2(1)$	1	0	$Y_2(1)$	$Y_2(0)$
3	$Y_3(0)$	$Y_3(1)$	0	1	$Y_3(0)$	$Y_3(1)$
4	$Y_4(0)$	$Y_4(1)$	1	1	$Y_4(1)$	$Y_4(1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 1: Hypothetical policy example

[3] Neyman-orthogonality states that the Gateaux derivative with respect to the nuisance parameters is zero in expectation at the true nuisance parameters.

[4] The example is taken from Causal ML: Offline policy learning.

Given the policy, individuals are assigned to treatment or control based on their potential outcomes. The value functions of the policies then take expectations over the two right columns: $Q(\pi^1) = E[Y(\pi^1)]$ and $Q(\pi^2) = E[Y(\pi^2)]$, essentially asking, "What would the Average Potential Outcome (APO) be if we had implemented the policy?"

The objective function of policy learning is to maximize the value function

$$\pi^* = \arg\max E[Y(\pi(X))] = \arg\max Q(\pi)$$

and the corresponding optimal value is $Q^* \equiv Q(\pi^*)$.

In case one wants to use machine learning for policy learning, there is a particularly useful representation of the objective function:[5]

$$\pi^* = \arg\max E[|\tau(X)| \operatorname{sign}(\tau(X)) \cdot (2\pi(X) - 1)] \tag{3}$$

where $(2\pi(X) - 1) \in \{-1, 1\}$ equals 1 if the policy assigns treatment and -1 if not. The function then measures the advantage of a policy compared to random allocation.

Assuming that X contains all confounders and that the common support holds, the value function is identified as:

$$\begin{aligned} Q(\pi) &= E[Y(\pi(X))] = E[\pi(X)Y(1) + (1 - \pi(X))Y(0)] \\ &= E[\pi(X)m(1, X) + (1 - \pi(X))m(0, X)]. \end{aligned}$$

Then the optimal policy with the highest value function for the set of candidate policy rules is identified as:

$$\begin{aligned} \pi^* &= \arg\max_{\pi} E[Y(\pi(X))] \\ &= \arg\max_{\pi} E[\pi(X)m(1, X) + (1 - \pi(X))m(0, X)]. \end{aligned}$$

Equation (3) suggests treating policy learning as a weighted classification problem. It aims to predict outcomes accurately for people with large CATEs, while those with CATEs close to zero are seen as less important.

Conveniently, the doubly robust APO score can be reused to estimate optimal treatment assignment:

$$\pi^* = \arg\max_{\pi} E[|\widetilde{Y}_{ATE}| \operatorname{sign}(\widetilde{Y}_{ATE}) \cdot (2\pi(X) - 1)]$$

This suggests that we can estimate $\tau(X)$ using the same approach as before, treating $\tau(X)$ as an additional nuisance parameter to estimate and plug into the policy under evaluation. The resulting weighted classification problem can be addressed using methods like classification trees/forests, Logistic Lasso, Support Vector Machines, and others.

Zhou et al. (2018) demonstrate that the Doubly Robust Machine Learning (DML)-based procedure asymptotically minimizes maximum difference between the true and estimated optimal value functions under two conditions. First, the product of the convergence rates of the nuisance parameters achieves $n^{1/2}$. Second, the set of candidate policy rules Π should not be overly complex. Specifically, they show that decision trees with fixed depth provide a suitable class of policy rules.

[5] For details, see Causal ML: Offline policy learning.

2.4 Conditional Independence Violation

Consider a case where $Y_i(w) \not\perp\!\!\!\perp W_i \mid X_i = x$, i.e., the measured confounding assumption does not hold: given the pre-treatment characteristics, potential outcomes and treatment are not independent. Specifically, this failure is due to some group-specific unobserved component C_i, commonly referred to as the "unobserved effect" (Wooldridge, 2010, ch. 10).

We assume that if the group indicator $C_i = c$ were available, we could achieve conditional independence, i.e.,

Assumption 2: $Y_i(w) \perp\!\!\!\perp W_i \mid C_i = c, X_i = x$, for all $w \in W$, $x \in X$ and $c \in C$.

Since treatment W_i is random given C_i and X_i, it holds that

$$E[W_i \mid C_i, X_i, Y_i(1), Y_i(0)] = E[W_i \mid C_i, X_i],$$

i.e., when conditioning additionally on the group indicator C_i, treatment assignment is independent of the potential outcomes $Y_i(1)$ and $Y_i(0)$. This also implies mean independence:

$$E[Y_i(0) \mid C_i, X_i] = E[Y_i(0) \mid C_i, X_i, W_i = 0] \overset{(1)}{=} E[Y_i \mid C_i, X_i, W_i = 0],$$
$$E[Y_i(1) \mid C_i, X_i] = E[Y_i(1) \mid C_i, X_i, W_i = 1] \overset{(1)}{=} E[Y_i \mid C_i, X_i, W_i = 1].$$

We also allow C_i to be arbitrarily correlated with X_i. Under a linear functional form assumption, this setup closely resembles a fixed effects setting for cross-sectional data. Traditional estimation methods to ensure consistency include adding group dummies to the regression, first differencing the data, or subtracting the group mean.

To eliminate unit fixed effects and still be able to use flexible non-parametric estimators like AIPW, we adjust the data by subtracting the respective group mean from each observation i (a so-called "within transformation"). This approach allows us to remain agnostic about the specific functional form of the data while ensuring that the measured confounding assumption is satisfied.

This adjustment can be represented as:

$$\ddot{Y}_i = Y_i - \overline{Y}_c, \quad \ddot{X}_i = X_i - \overline{X}_c$$

with

$$\overline{Y}_c = \frac{1}{N_c} \sum_{i \in C_i} Y_i, \quad \overline{X}_c = \frac{1}{N_c} \sum_{i \in C_i} X_i$$

where $\ddot{Y}_i$ and $\ddot{X}_i$ are the demeaned values, $\overline{Y}_c$ and $\overline{X}_c$ are the averages, and N_c is the number of observations, all within cluster $C_i = c$.

Demeaning effectively controls for unobserved, group-invariant characteristics specific to each cluster C_i, isolating the treatment effect W_i on the outcome Y_i. This approach also avoids the need to explicitly condition on C_i.

For the transformed data it must hold that

$$\ddot{Y}_i(w) \perp\!\!\!\perp W_i \mid \ddot{X}_i = \ddot{x},$$

The *demeaned AIPW*[6] estimator then differences the pseudo-outcomes learned from the adjusted data:

$$\widetilde{Y}^*_{ATE} = \underbrace{\hat{m}(1,\ddot{X}) - \hat{m}(0,\ddot{X})}_{\text{outcome predictions}} + \underbrace{\frac{W(\ddot{Y} - \hat{m}(1,\ddot{X}))}{\hat{e}(\ddot{X})} - \frac{(1-W)(\ddot{Y} - \hat{m}(0,\ddot{X}))}{1-\hat{e}(\ddot{X})}}_{\text{weighted residuals}} \tag{4}$$

Note that the treatment indicator W_i remains unaffected by these transformations. If one assumes a linear conditional expectation function (CEF) and runs a regression on these data, the point estimates would *not* be numerically equivalent to running a fixed effects regression with group-specific dummies. This specific transformation is essential for the AIPW estimator, which needs a binary indicator to learn CATE for treated and control groups separately.

3 Application: Vocational Training for Disadvantaged Youth in Colombia

3.1 The setting

We consider the study by Attanasio et al. (2011), which assesses the impact of a randomized training program for disadvantaged youth introduced in Colombia in 2005.

The study evaluates the Jóvenes en Acción ("Youth in Action") program, which offered subsidized training to impoverished young individuals in urban areas. It was rolled out to multiple cohorts over four years, with the paper focusing on the final cohort that was randomly assigned to receive training.

The program consisted of three months of classroom training and three months of on-the-job training. The vocational skills provided by the courses were very diverse (see Figure A1). The greatest number of courses were offered in administrative occupations (such as sales, secretarial work, and marketing). However, there were also a large number of courses in manual occupations (electricians, cooking assistants), as well as courses in fairly skilled occupations including IT specialists and accountant assistants.

The randomization worked as follows: Each training institution was instructed to select a list of up to 50 percent more applicants than they had capacity for. Since the total number of slots per class was fixed but the extent of oversubscription differed by site and class, the probability of being offered a spot differed between training institutions. Most importantly, applicants were able to sort themselves into different training institution or courses on the basis of tastes, ability, etc.

The follow-up interviews to access the effect of the intervention were carried out 14 months after the conclusion of the program.

3.2 Descriptive statistics

We summarize descriptive statistics to confirm successful randomization and representativeness of the sample for the general population.

The attrition rate between baseline and endline surveys is comparable to that of other labor market studies (19.5%). However, treated men are 0.07 more likely to remain in

[6] Unless otherwise stated, demeaned AIPW will simply be referred to as AIPW to reduce notational clutter.

the sample, which could potentially introduce bias. It is not clear, however, in which direction this bias operates.

Compared to the 2005 National Household Surveys (NHS), the sample alligns with the NHS in mean age (21 years) and gender distribution (55.6% women). However, it shows higher educational attainment and employment rates, with more written employment contracts. In contrast, NHS respondents are more likely to work in the formal sector and have longer job tenure, making the overall program selection effect ambiguous.

If the randomization was successful, the baseline characteristics of control and treatment groups within courses should not be significantly different. Overall, the two samples are balanced: the F-statistic for women is 1.54 and 2.61 for men. The authors marginally reject that the baseline characteristics of treated and control men are the same. The baseline imbalance for men points to a slight positive selection bias.

Following the authors, we control for observable pretreatment characteristics to reduce any remaining baseline imbalances (Attanasio et al., 2011).

3.3 Identifying Program Effects

In the following, estimation results for both the fixed effects and AIPW estimators are presented. A necessary condition for successful identification is that the discussed assumptions are satisfied:

1. *Measured confounding* is expected to hold since self-selection is addressed by (a) controlling for site-by-course fixed effects for the linear model and (b) demeaning the data prior to estimation for the AIPW estimator. This should effectively eliminate unobserved heterogeneity between units.

2. *Common support* is ensured by the appropriate randomization design.

3. *SUTVA* may be violated since participants receive different training programs under a single "treated" category. We follow the authors' rationale but will address this issue later in the paper.

By imposing a linear functional form and a classical strict exogeneity assumption, one can show that, in combination with Assumption 1, the treatment effect for the fixed effects model can be identified as[7]

$$\mathbb{E}[Y_i \mid W_i, X_i, C_i] = \tau W_i + C_i + X_i'\beta.$$

The unknown parameter of interest, τ, can then be obtained as the solution of a least squares problem by adding group dummies into the regression.

3.4 Estimating Program Effects

For the AIPW score, the nuisance parameters are estimated via random forest using the implementation with honest splitting in the grf R-package (Athey et al., 2019) and three-fold cross-fitting.[8] The tuning parameters in each regression are selected by out-of-bag

[7] The complete identification exercise is provided in Section A.1.

[8] Courses with only treated or only control participants are excluded from the analysis. Fixed effects ignore them, but the AIPW estimator learns their CATE, violating the common support assumption.

validation. All regressions apply the full set of pre-treatment characteristics.[9] We run the outcome regressions for each treatment group separately to obtain $\hat{m}(w, x)$. The propensity scores are estimated using a treatment indicator as outcome in the random forest.

In the original paper, estimation results are presented separately for men and women, focusing on the treatment effect heterogeneity between these two subgroups. We aggregated the data to compute the average treatment effect (ATE) for the entire sample, comparing the performance of the fixed effects and AIPW estimators.[10]

Table 2: Comparison of AIPW and FE for ATE estimation

	Days/ month	Hours/ week	Tenure	Wage and salary earnings	Self-employment earnings
Fixed effects	0.446	0.552	-2.048***	28,692***	-2,892
	(0.414)	(0.964)	(0.494)	(7,463)	(3,022)
AIPW score	**0.433**	**0.510**	**-1.974*****	**26,732*****	**-3,034**
	(0.374)	(0.867)	(0.451)	(6,658)	(2,686)
Observations	3,119	3,119	3,119	3,119	3,119

Significance levels: * p < 0.1, ** p < 0.05, *** p < 0.01.

This aggregated comparison serves as a sanity check (Table 2). Note that the values in the table represent treatment effects rather than outcomes. The results from both estimators are reasonably close. Specifically, program participation has a positive and significant effect on salary earnings and a negative impact on tenure, reflecting the time lost during training.

Table 3 presents the core subgroup analysis conducted in Attanasio et al. (2011). For AIPW, it can be performed at low computational cost by regressing the pseudo-outcome on dummy variables for all groups except the reference group. The resulting point estimates and standard errors describe the differences in treatment effects between the subgroups.

A standard subgroup analysis using fixed effects would conclude a significant increase in the amount of time spent working and salary earnings for women (e.g., a 20% increase in women's earnings compared to the control group). The cost of training is reflected in lost tenure, estimated to be slightly over one month and significant. For men, none of the effects are significant at the 5% level, except for a reduction in tenure for program participants by about three months.

The AIPW suggests that while the effects of program participation are stronger for women, they are not significantly different compared to men. There is minimal treatment effect heterogeneity between these subgroups, except for a higher increase in hours worked per week among females (significant at the 10% level).

[9] The regressions control for the following pre-training baseline characteristics: age, education, marital status, employment, paid employment, salary, self-employment earnings, whether working in the formal sector, whether working with a contract, days worked per month, and hours worked per week.

[10] Additionally, the authors estimate participation effects on binary variables using conditional logit. These variables include indicators for employment status, paid employment, formal employment, and having a written contract. These are excluded from the analysis as they are unsuitable for the AIPW estimator.

Table 3: Treatment Effects on multiple outcomes by gender

	Days/ month	Hours/ week	Tenure	Wage and salary earnings	Self-employment earnings
Female					
Control means	14.84	31.82	7.14	177,161	11,970
Fixed effects	1.17*	2.87**	-1.43**	34,668***	1,950
	(0.61)	(1.40)	(0.62)	(9,743)	(3,568)
Observations	1,767	1,767	1,756	1,767	1,767
Male					
Control means	20.06	45.15	11.05	265,292	35,435
Fixed effects	-0.55	-2.27	-2.78***	13,690	-6,731
	(0.62)	(1.47)	(0.90)	(12,819)	(5,839)
Observations	1,464	1,464	1,460	1,464	1,464
Difference					
Fixed effects	1.72	5.14	1.342	19,978	8,681
AIPW score	**1.25**	**3.75***	**0.48**	**21,765**	**9,989**
	(0.89)	(2.07)	(1.14)	(16,332)	(6,164)
Observations	3,119	3,119	3,119	3,119	3,119

Significance levels: * p < 0.1, ** p < 0.05, *** p < 0.01

The differences in results between the two estimators could potentially be attributed to AIPW's use of random forest to non-parametrically control for any remaining imbalances and confounding. However, further exploration is needed.

Notably, interpreting the level of treatment effects becomes increasingly imprecise when regressing AIPW scores on the transformed data. For instance, course-demeaned dummy variables, such as gender, effectively transform into pseudo-continuous variables, altering the interpretation of dummy regressions. While the difference in effects between subgroups (from 0 for male to 1 for female) remains valid, interpreting the effect with `female = 1` would not be correct, as the value of "1" is not within the common support of the demeaned gender indicator.

To approximate the level effects, one could compute the mean value of `female < 0` = -0.45 (male) and `female > 0` = 0.38 (female), which would be -0.5 and 0.5, respectively, in the case of equal gender composition in the whole sample. Then, multiply the dummy coefficient by -0.45 to get the effect for males (16,510) and by 0.38 to get the effect for females (35,405). These values are close to the results provided by fixed effects.

However, there is an additional issue: some courses are exclusively male or female (18% of the total number of courses). By demeaning these, we get an average gender within a course of 0 (both 0-0 and 1-1), thereby pooling the effects of these two groups together. This reduces the credibility of the approximation and complicates the logic, especially for discrete variables like education.

In addition to examining differences in treatment effects by gender, we investigate GATEs for subgroups based on marital status, following a pre-specified hypothesis. Having a hypothesis specified before estimation is necessary to keep the analysis credible.

Apart from the difference in lost tenure, there are no significant treatment effect

Table 4: Treatment Effects on multiple outcomes by marital status (AIPW)

	Salary (1)	Days/ month (2)	Hours/ week (3)	Profit (4)	Tenure (5)
(Intercept)	26,732***	0.43	0.51	-3,336	-1.97***
	(6,660)	(0.37)	(0.87)	(2,654)	(0.45)
Married	1,853	0.96	1.70	6,681	2.09*
	(18,352)	(1.04)	(2.42)	(7,933)	(1.25)

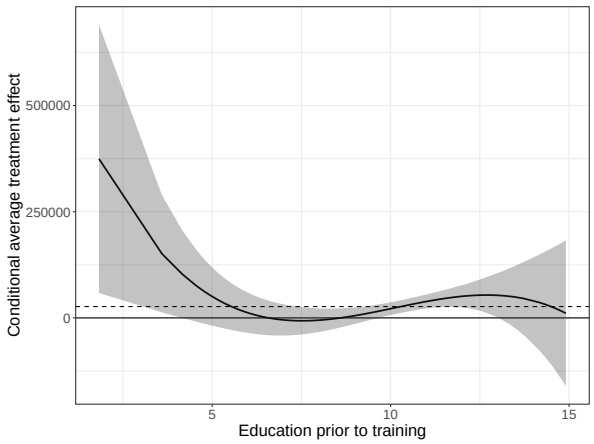
Significance levels: * p < 0.1, ** p < 0.05, *** p < 0.01

heterogeneities between married and non-married participants (Table 4).

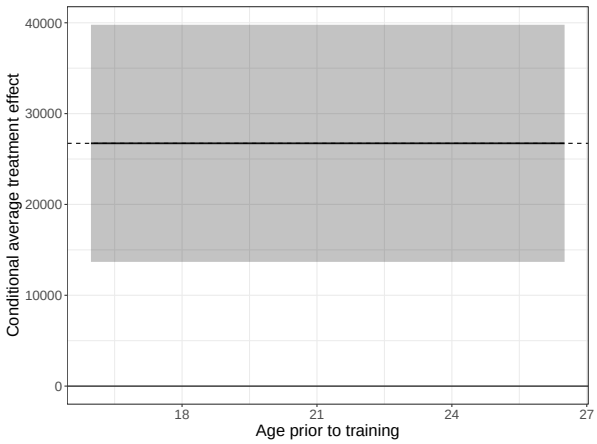
While subgroup analysis is standard in program evaluations, the estimation of non-parametric GATEs using kernel or series regression is rarely conducted. We estimate kernel (reported in the code) and spline regressions using the R packages np (Hayfield and Racine, 2008) and crs (Racine and Nie, 2021), respectively. The kernel regressions apply a second-order Gaussian kernel function and use 0.9 of the cross-validated bandwidth for undersmoothing. The spline regressions use B-splines with cross-validated degree and number of knots (following Knaus, 2022).

Conditional treatment effects are estimated along education, age and the time worked prior to the intervention[11] (see Figure 1). Due to the limited sample size, drawing solid conclusions is challenging. However, we cautiously suggest that there might be a higher treatment effect for those with more education (10+ years). Age does not appear to be a determinant of the treatment effect, which is plausible given the program design (most participants are of similar age). Individuals with less time worked prior to the training seem to experience a higher treatment effect.

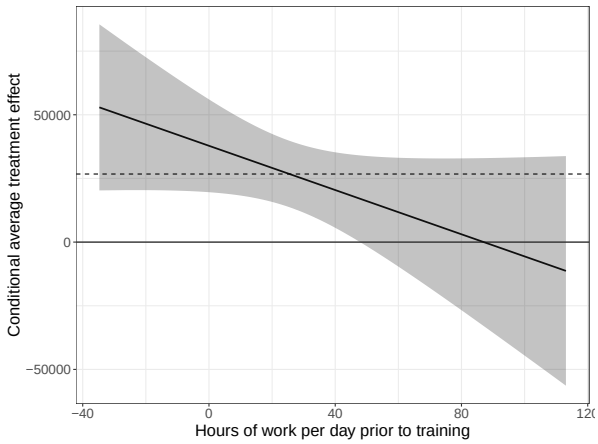
[11]Note that the data were demeaned. To facilitate interpretation, we re-added the population mean of the variable of interest on the X-axis (i.e., mean education in the sample). This is an imperfect approximation.



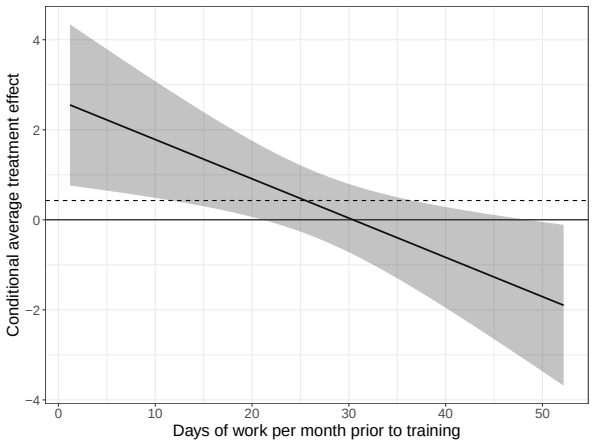
(a) Salary increase across education levels prior to participation



(b) Salary increase across participants' age prior to intervention



(c) Salary increase across hours worked per week prior to participation



(d) Increase in days worked per month across days worked per month before to participation

Figure 1: Effect heterogeneity along X. Dotted line indicates point estimate of the respective ATE. Grey area shows 95% confidence interval.

Finally, we approximate the CATE function by specifying a multivariate OLS regression with all available covariates entered linearly. While it is an approximation, it provides a concise and accessible summary of the effect heterogeneities.

Compared to Tables 3 and 4, the coefficients in Table 5 differ because they represent partial effects with other variables held constant. Similar strategies used to interpret an outcome in OLS regression can be applied to interpret effect regressions.

Overall, considering multiple hypotheses testing, we conclude that we found no evidence of linear effect heterogeneity, or we were unable to identify it due to the modest sample size or weak signal.

3.5 Cost-Benefit Analysis and Optimal Treatment Assignment

Attanasio et al. (2011) calculate the benefits of the program using the estimated salary gains. The program's direct cost is US$750 per person.

Under two scenarios, i.e. permanent and 10% annually depreciated gains, assuming 40 more working years (average age is 22) and discounting at 5 percent annually, they

Table 5: Best linear prediction of CATEs.

	Salary (1)	Days/ month (2)	Hours/ week (3)	Profit (4)	Tenure (5)
(Intercept)	28,035***	0.48	0.85	-2,952	-1.93***
	(6,737)	(0.38)	(0.87)	(2,684)	(0.46)
Female	20,766	1.02	3.06	11,918*	0.40
	(17,130)	(0.93)	(2.17)	(6,172)	(1.16)
Age	-2,138	0.08	-0.17	183	-0.50
	(3,761)	(0.21)	(0.48)	(1,571)	(0.31)
Married	3,397	0.61	1.46	5219	2.46*
	(19,403)	(1.09)	(2.52)	(8,255)	(1.34)
Education	7,231	0.21	0.68	27	-0.17
	(4,743)	(0.28)	(0.64)	(1,804)	(0.40)
Employed	54,278	5.32**	11.10*	7,036	-1.44
	(46,335)	(2.57)	(5.86)	(28,206)	(4.01)
Privately employed	-15,788	-0.84	-1.82	-6,296	0.40
	(39,736)	(2.10)	(4.86)	(19,952)	(3.67)
Salary	-0.07	0.00	0.00	0.11**	0.00
	(0.11)	(0.00)	(0.00)	(0.05)	(0.00)
Formal	-4,135	0.12	2.27	5,839	-0.66
	(37,027)	(1.92)	(4.44)	(11,363)	(2.67)
Contract	28,942	1.36	3.26	-29,047**	0.49
	(39,121)	(2.02)	(4.62)	(13,191)	(2.87)
Days	-610	-0.30***	-0.47*	-989	-0.10
	(1,986)	(0.11)	(0.24)	(1,111)	(0.15)
Hours	-545	0.02	-0.05	103	0.01
	(639)	(0.03)	(0.08)	(271)	(0.05)
Profit	-0.08	-0.00	-0.00	-0.02	-0.00
	(0.15)	(0.00)	(0.00)	(0.11)	(0.00)

Significance levels: * p < 0.1, ** p < 0.05, *** p < 0.01

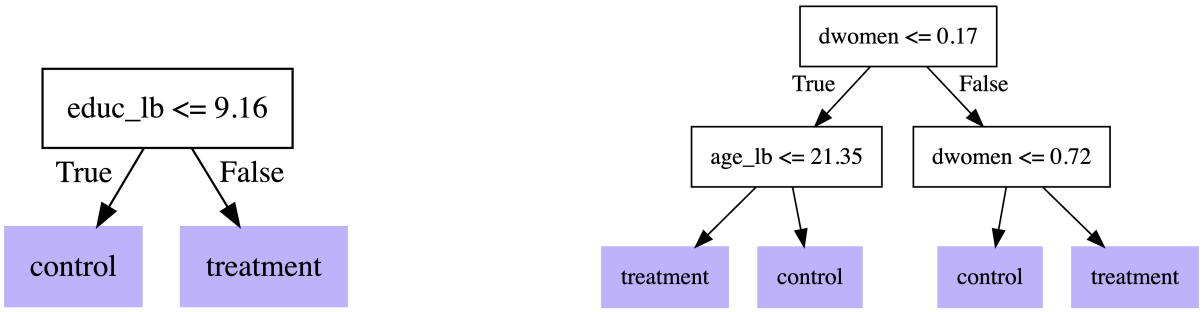
compute the gains separately for male and female [12], showing that the program is highly effective for women under both scenarios (permanent effects and with depreciation). For men, there are no easily measurable gains.

Since no substantial heterogeneity was documented in the previous section, with only occasional effect heterogeneity observed with respect to some variables, one can still apply the DML-based optimal policy algorithm as an exercise, without placing strong interpretation on the results.

The optimal treatment assignment rule is estimated as decision trees of depth one and two that is implemented in the policytree R-package (Sverdrup et al., 2020). We estimate the trees allowing the separation on any pre-treatment characteristic. These variables include gender and might be too sensitive to include in practice. An alternative would be to investigate a set of variables that includes only the objective measures of education and

[12] All conversions to US dollars are made at a rate of US$1 = COP$1,970.00

labour market history, which would be more neutral. Methods for introducing program costs into the policy recommendation are briefly discussed in Section A.2.



(a) Policy tree of size 1

(b) Policy tree of size 2

Figure 2: Optimal treatment assignment decision trees of depth one and two.

Because of demeaning, interpreting policy recommendations becomes challenging. Similar to spline regressions, the population mean is reintroduced to the variables. The policy recommendation selects variables with the highest t-statistics in the best linear predictor of CATE. For the treatment effect on salary, these variables are gender and education (Table 5). However, these results lack credibility, and the issue of performing statistical inference on decision tree quality remains unresolved.

4 Fixed Effects as a biased estimator of ATE

Estimating heterogeneous treatment effects in the Youth in Action study is challenging due to questionable assumptions, modest sample size, and necessary data adjustments. Additionally, there is a subtle issue that may introduce bias in the estimation.

Under the within-groups estimator, the program's treatment effect is a conditional-variance weighted average of effects across multiple courses (Attanasio et al., 2011):

$$\hat{\tau}^{FE} \xrightarrow{p} \frac{\sum_c P_c(1-P_c)\pi_c\tau_c}{\sum_c P_c(1-P_c)}$$

where P_c is the share of treated individuals in group c (i.e., the training course), π_c is the population frequency of group c, and τ_c is the group-specific average treatment effect.

This approach is efficient if the treatment effect is identical across all groups. However, a potential issue arises when treatment effects vary among the groups used for fixed effects. In such cases, fixed effects regressions average the group-specific slopes based on both the group's sample frequency and the conditional variance of the treatment.

In general, it can be shown that fixed effects is a biased and inconsistent estimator of the ATE[13]:

$$\hat{\tau}^{FE} \xrightarrow{p} \tau^{ATE} + \sum_c \pi_c\tau_c \left[\frac{P_c(1-P_c)}{\sum_c P_c(1-P_c)} - 1\right]$$

Gibbons et al. (2019) and Imai et al. (2019) note that this average generally does not coincide with the average treatment effect: the fixed effects estimator up-weights

[13]We take the result of the derivation proposed by Gibbons et al. (2019) and adapt it to the setting in Attanasio et al. (2011). See Section A.3

groups with high variance in treatment within a group and down-weights groups with low variance in treatment.

An example where FE would give unbiased results is a regression using data from a perfectly randomized experiment where treatment has the same variance across groups (i.e. the share of treated is identical). Achieving such perfect conditions is unlikely in observational or experimental settings, though.

Gibbons et al. (2019) replicate Karlan and Zinman's (2008) randomized experiment where treatment is randomized within fixed effects groups, but treatment variances differ across groups. They found the average treatment effect differs from the fixed effects estimate by over 60%.

In the Youth in Action sample, we can calculate the conditional treatment probability distribution across courses by dividing the number of treated individuals within each course by the total number of students in that course (Figure 3).

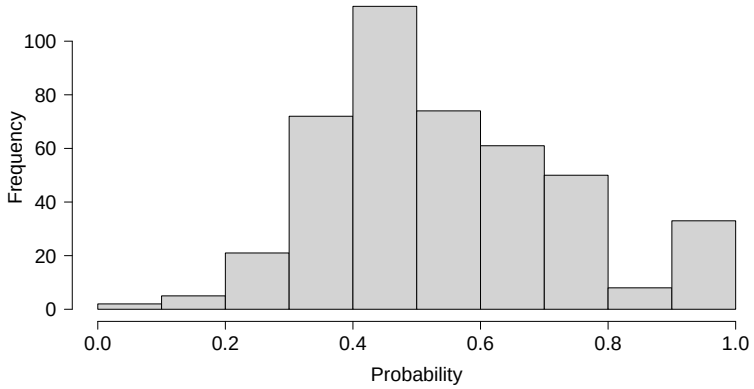


Figure 3: Distribution of propensity scores across courses

Treatment propensity varies between groups, which puts fixed effects at risk of inconsistently estimating the ATE. Meanwhile, the extent of the bias remains unclear.

To demonstrate how potential bias might affect the estimate of τ, we generate two datasets. Both have linear confounding, DGP-specific fixed effects, and a constant treatment effect, which are of opposite signs: $\tau_1 = -1.5$ and $\tau_2 = 1.5$. Most importantly, the treatment indicator has the highest variance for the group with a negative treatment effect ($W_1 \sim$ Bernoulli(0.5)). Therefore, when pooling these DGPs together, we expect to see a downward bias (true $\tau_{ATE} = 0$)[14].

	τ_{ATE}	Bias_{POLS}	Bias_{FE}
Value	0	-0.2971	-0.3200

Table 6: Biasedness of fixed effects under varying treatment shares

As expected, the ATE estimate is negatively biased because the group with an equal share of treated and control observations receives higher weight in the final estimate (Table

[14]Each DGP has $n = 1000$, totaling 2000 when pooled. We conduct 1000 simulations, across which the average bias is computed. For a complete description of the DGPs, see Table A1

6). However, this illustrates an extreme case of opposite group treatment effects.

In fact, the AIPW estimator also produces a variance-weighted estimate. Consistent estimation of ATE under different shares of treated within groups, coupled with heterogeneous treatment effects, is questionable.

In the next section, we simulate multiple settings with increasing complexity to assess the estimators' properties in a controlled environment.

5 Monte Carlo Simulation

We evaluate the finite sample performance of our estimators using a synthetic data generating process (DGP). While an Empirical Monte Carlo Study (EMCS) would be preferred[15], we use a synthetic DGP for its simplicity.

The objective is to reconstruct the main features of the job training design as presented in Attanasio et al. (2011). The main simulation principles are:

(a) The treatment W is randomly assigned. Importantly, the proportion of treated units across groups (clusters) is kept roughly constant (Case 1) and allowed to vary (Case 2) to resemble the study's setting. Thus, we assess the impact of conditional treatment variance on bias.

(b) Heterogeneity is modeled linearly. This choice is motivated by our focus on the Best Linear Approximation (BLA) of the Conditional Average Treatment Effect (CATE); non-linear heterogeneity would not be accurately captured by a linear function.

(c) Attanasio et al. (2011) control for residual imbalances by including pretreatment characteristics in the regression. We then add both linear and non-linear confounding factors that influence the outcome (but not the treatment).

(d) Central to our study is the inclusion of unit fixed effects, which may be arbitrarily correlated with X. Additionally, levels of explanatory variables are allowed to vary across groups.

The general goal is to determine when the estimation works and at which point it breaks. This basic understanding serves as a solid foundation for further research.

5.1 Case 1: No Confounding

We create a dataset with 5 groups C_i (e.g., training courses), each having the same number of observations (balanced groups). The treatment effect is a linear function of the covariates (Table 7). Each observation includes a unit fixed effect, which captures its unique, unobserved characteristics.

Throughout this section, it is convenient to have a hypothetical example to make the numerical results more tangible.

The story could go as follows: The government is interested in evaluating a new teaching method intended to improve student test scores. This study involves five different schools, each with 200 (or 800) students participating. The goal is to understand how

[15]The Empirical Monte Carlo Study (EMCS) informs DGPs using real data to minimize synthetic components. For an example, see Knaus et al. (2021), where they construct a DGP based on covariates and treatment assignments from four medical datasets. This approach remains a topic for future research.

DGP: Case 1	
Covariates	$p = 4$ independent covariates $X_1, \ldots, X_4$, where $X_k \sim N(k, 1) + R_i, k \in \{1, 2, 3\}$ and $X_4 \sim$ Bernoulli(0.5)
Treatment	$W \sim$ Bernoulli(0.5)
CATE function	$\tau(X) = -0.3 - 0.8X_1 + 0.8X_2 + 0.3X_3 - 0.6X_4$
Outcome (Treated)	$Y(1) = \tau(X)W + Y(0)$
Outcome (Control)	$Y(0) = C_i X_3^2 + \varepsilon$, with $\varepsilon \sim N(0, 5)$
Unit Effects	$C_i \sim$ Uniform$(-4, 4)$; $R_i \sim$ Uniform$(-2.5, 2.5)$

Table 7: Data Generating Process (Case 1)

the effectiveness of the new teaching method varies depending on certain characteristics of the students.

The covariates could be the student's initial proficiency level (X_1), parental involvement (X_2), the student's motivation level (X_3), and a binary indicator of participation in after-school programs (X_4).

The new teaching method W is then randomly assigned to 50% of the students across all schools. Due to the relatively small number of groups, the share of treated students is roughly constant across units.

The impact of the new teaching method on test scores $\tau(X)$ varies linearly based on the covariates: for example, its effectiveness decreases with higher initial proficiency ($-0.8X_1$) but increases with greater parental involvement ($+0.8X_2$).

The test score $Y(1)$ is a function of the treatment effect, a school-specific effect, and some random noise. Each school has its unique characteristics C_i, be it teaching quality, school facilities, etc., which are uniformly distributed across a certain range.

Simulation results

Regressing the AIPW score on the covariates allows the intercept to be interpreted as the ATE, and the coefficient estimates for X represent the effect heterogeneity with respect to each variable. The standard errors require no adjustment, as mentioned earlier.

Inference on ATE is less straightforward when estimating pooled OLS and fixed effects regressions. If we assume that the researcher correctly specified the linear model by adding the interaction terms, the equation takes the following form:

$$\begin{aligned} Y_i = &\beta_0 + \beta_1 W_i + \beta_2 X_{1i} + \beta_3 X_{2i} + \beta_4 X_{3i} + \beta_5 X_{4i} \\ &+ \beta_6 X_{1i} \cdot W_i + \beta_7 X_{2i} \cdot W_i + \beta_8 X_{3i} \cdot W_i + \beta_9 X_{4i} \cdot W_i + C_i + \varepsilon_i \end{aligned}$$

And the Average Treatment Effect is then defined as:

$$\tau_{FE}^{ATE} = \beta_1 + \beta_6 \cdot \overline{X}_1 + \beta_7 \cdot \overline{X}_2 + \beta_8 \cdot \overline{X}_3 + \beta_9 \cdot \overline{X}_4,$$

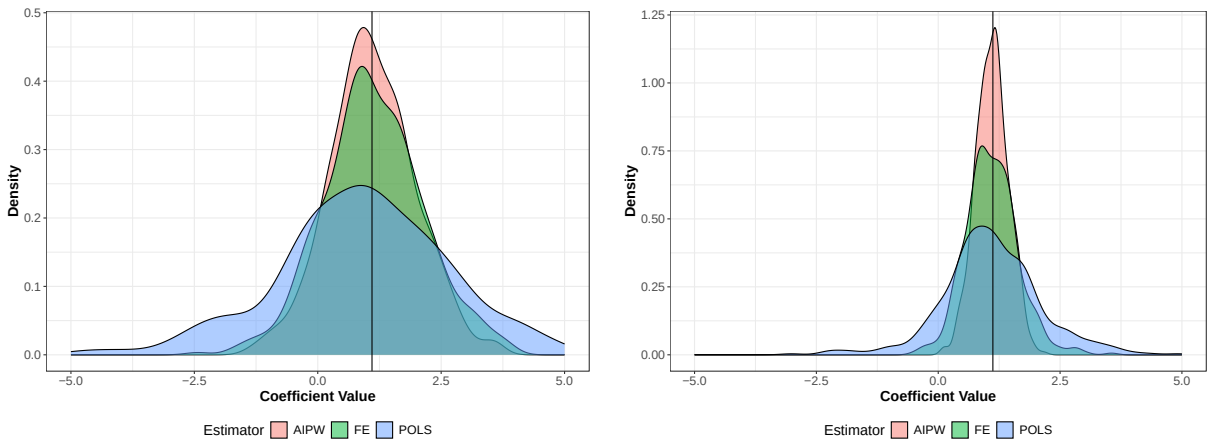
where $\overline{X}_1, \overline{X}_2, \overline{X}_3$, and $\overline{X}_4$ are the average values of the respective variables.

	1,000 observations				4,000 observations			
	MSE	|Bias|	SD	95% CI	MSE	|Bias|	SD	95% CI
POLS	3.67	1.44	1.92	.952	1.02	0.73	1.01	.942
FE	0.98	0.77	1.01	.950	0.26	0.39	0.54	.956
AIPW	0.72	0.67	0.87	.948	0.07	0.21	0.33	.962

Table 8: Estimation performance summary (Case 1)

Since the parameter of interest τ_{FE}^{ATE} is a linear combination of the estimated parameters, standard errors are obtained by the delta method.

Table 8 shows the main performance measures for the estimated average treatment effect. AIPW consistently outperforms both FE and POLS in terms of bias, standard deviation and MSE.[16] As sample size increases, the difference becomes more striking. Non-parametric ML estimation methods, such as random forests (which are the building blocks of AIPW), are known to have slower convergence rates. However, given sufficient data, they learn the CATE function more accurately.[17]



(a) $n = 1000$, 500 simulations (b) $n = 4000$, 500 simulations

Figure 4: Estimated parameter density (Case 1)

For all estimators, the coverage rate is nominal, indicating that the estimators are unbiased (Figure 4) and their standard errors are reliable.

The performance of the estimated effect heterogeneity with respect to X can be found in the appendix (Section A.5). Essentially, the patterns discussed here are closely repeated.

[16] By design, the average treatment effect depends on the values of the covariates (see Table 7). The Mean Squared Error (MSE) is then decomposed into the squared bias, the estimator's variance, and a specific term that captures how the bias of the estimator correlates with the true parameter across different simulations. Further details can be found in Section A.4.

[17] Sample size is increased by adding more observations per unit while keeping the number of units constant.

5.2 Case 2: Nonlinear Confounding

This time, there are 40 clusters, each with only 25 (or 100) observations per cluster. Additionally, there is non-linear outcome confounding,

$$\boldsymbol{\beta X} = 0.1X_1^2 - 0.1X_2 + 0.5\sin(X_3) - 0.5\exp(X_4)$$
$$Y(1) = \tau(X)W + u_i + \boldsymbol{\beta X} + \varepsilon$$

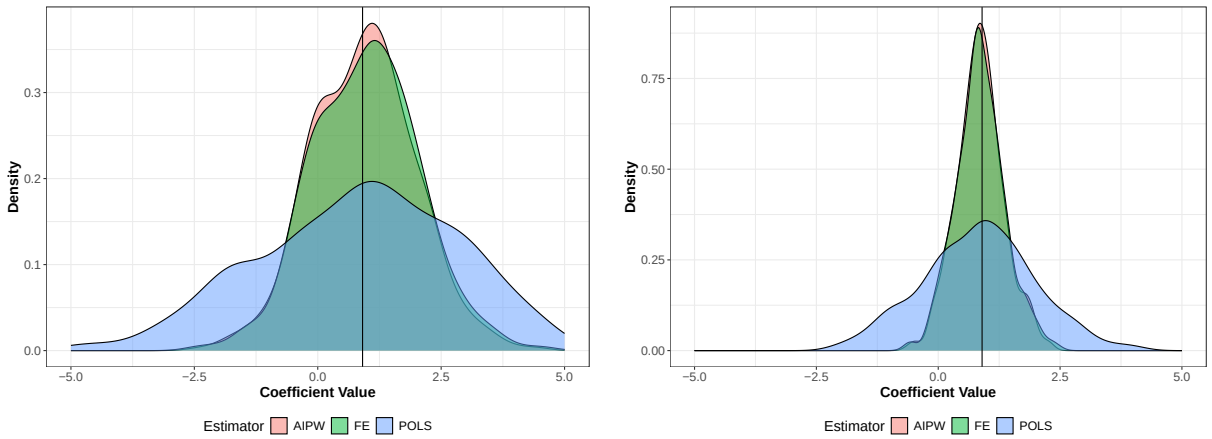
which should not make the treatment effect estimation much harder, since W is still randomly assigned (see Table A4 for complete description).

Following our hypothetical example, a new teaching method could be randomly assigned to smaller entities like classes. In fact, this setting closely resembles the one described in Attanasio et al. (2011), where the shares of treated individuals vary across groups to roughly the same extent (see Figure A3).

	1,000 observations				4,000 observations			
	MSE	|Bias|	SD	95% CI	MSE	|Bias|	SD	95% CI
POLS	5.33	1.81	2.31	.950	1.34	0.92	1.15	.956
FE	1.16	0.86	1.08	.946	0.26	0.39	0.51	.946
AIPW	1.09	0.83	1.05	.950	0.24	0.38	0.49	.950

Table 9: Estimation performance summary (Case 2)

Interestingly, this time both FE and AIPW demonstrate comparable results, while POLS is unbiased but has a large variance (Table 9). Potentially, POLS stays unbiased because the fixed effects component is centered around zero ($C_i \sim \text{Uniform}(-4, 4)$). The good performance of FE as well as re-centering of C_i remain to be explored. The coverage rates for all estimators are nominal.



(a) $n = 1000$, 500 simulations (b) $n = 4000$, 500 simulations

Figure 5: Estimated parameter density (Case 2)

5.3 Case 3: Non-parametric Heterogeneity

The fixed effects regression, if correctly specified, effectively approximates the CATE function linearly. This questions the usefulness of the AIPW estimator.

In fact, the decision to generate the data with linear effect heterogeneity in X greatly facilitated the task for fixed effects. If we were to model complex nonlinear relationships in X, we would likely aim to construct an estimator $\hat{\tau}(x)$ of the true CATE function $\tau(x)$ and not its best linear approximation.

To explore the limitations of the linear model, we modeled the data-generating process with zero ATE and highly nonlinear effect heterogeneity (see Figure 6). The complete DGP can be found in Table A5.

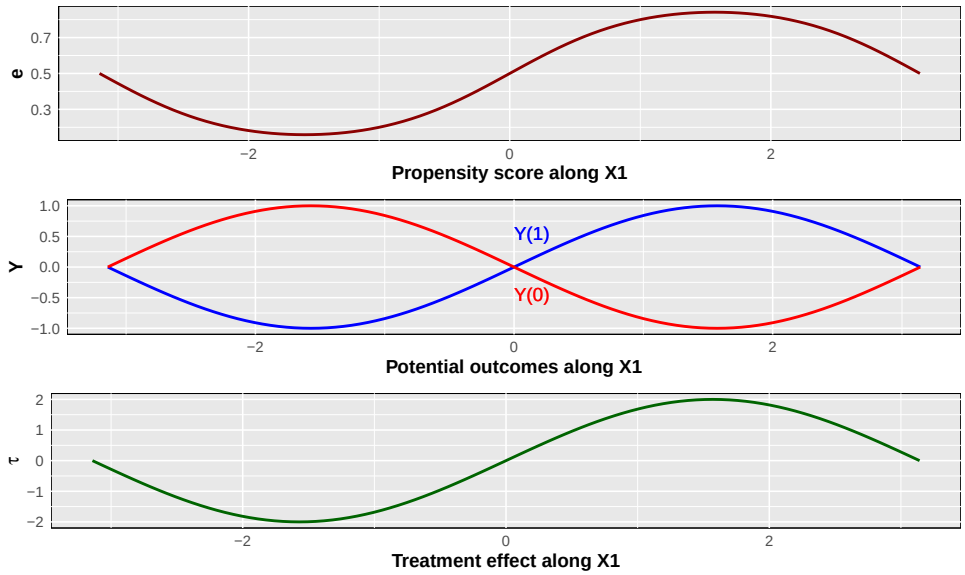
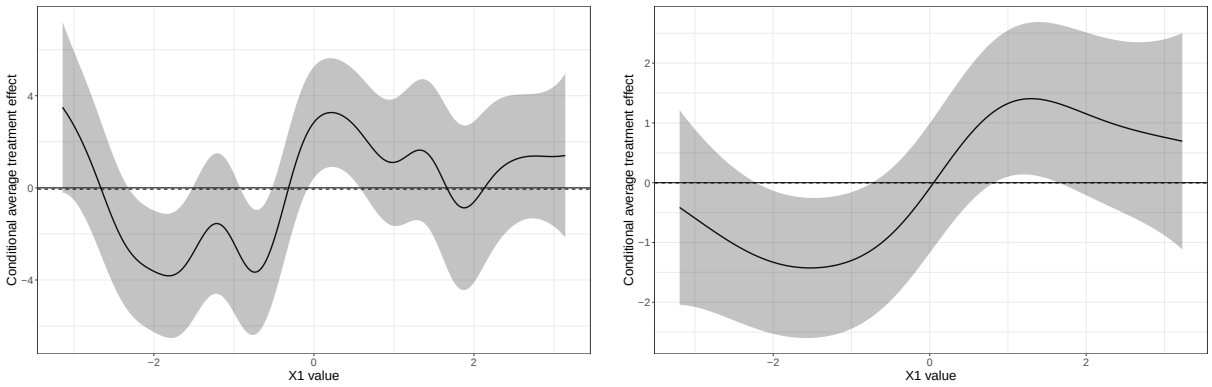


Figure 6: Visual DGP representation (Case 3)

Instead of running an OLS regression (considering non-linear heterogeneity this time), we use the pseudo-outcome $\tilde{Y}_{ATE}$ with X_1 as the only independent variable. The bandwidth for kernel regression is cross-validated, and the estimation is conducted using the np package (Figure 7).



(a) AIPW estimator on non-demeaned data (b) AIPW estimator on group-demeaned data

Figure 7: Kernel regressions for raw and transformed data, $Y = \widetilde{Y}_{ATE}$

Notably, both results employ doubly robust scores for estimation, but (b) uses demeaned data. We observe a general improvement: the sine wave is more pronounced

when unit fixed effects are removed.[18] Importantly, kernel regression with $\widetilde{Y}_{ATE}$ as the dependent variable allowed for a nonlinear CATE estimation.

6 Latent treatment heterogeneity

Throughout the paper, we implicitly assumed that while treatment effects may vary between participants, units, or groups, the treatment itself is homogeneous. In other words, while people's response to an intervention is not constant, the intervention itself must be identical for all.

In many applications, however, binary treatments can be seen as heterogeneous: they may represent diverse underlying effective treatments that influence the outcome of interest.[19]

In Attanasio et al. (2011), participants could select themselves into multiple training courses. This selection bias was supposedly eliminated by using the fixed effects estimator, which explicitly controlled for site-by-course imbalances. However, this does not mask the fact that the training courses were drastically different (Figure A1).

These are mostly apprenticeship positions, ranging from bus drivers (2 courses) to IT assistants (13 courses) and sports referees (1 course). Even common courses like secretary (34 courses) and sales representative (43 courses) represent grouped data from multiple cities and institutions, likely differing in curriculum, class sizes, etc.

Estimating CATE under multiple effective treatments grouped as "treated" or "not treated" can lack external validity, as replicating such treatment mixes in other settings proves challenging. This latent treatment heterogeneity also violates the consistency component of the 'Stable Unit Treatment Value Assumption' (SUTVA), which requires that each unit's potential outcome be uniquely determined by the treatment it receives, thereby endangering the identification process.

Overall, when estimating CATE, it remains unclear whether the observed effect heterogeneity under binary treatment reflects genuine differences in treatment effects or variations in the effective treatments themselves.

Heiler et al. (2021) distinguish two scenarios that define treatment heterogeneity:

- *Scenario 1*: Multiple or continuous treatments are aggregated into a binary indicator post hoc (e.g., different training programs become "training yes/no"). This simplification, driven by data availability or simplicity, can lead to unintended consequences. Discovered effect heterogeneity may be spurious, falsely attributing differences to units' background characteristics. Additionally, actual effect heterogeneity may be obscured by this aggregation.

- *Scenario 2* (multiple treatment versions): A binary treatment offers various versions post-assignment, such as access to a training program with multiple specializations. Heterogeneous effects may arise from varying program effectiveness across groups, curriculum tailoring, or both.

In the 'Youth in Action' program, Scenario 1 is definitely present: participants who underwent different effective treatments (participated in various training courses) were later grouped into a binary 'treatment/no treatment' indicator.

[18] The same result for spline regression can be seen in Figure A2.

[19] Effective treatments refer to treatments that induce variation in potential outcomes. Throughout this discussion, 'treatment' denotes effective treatment unless stated otherwise.

Regarding scenario 2, an illustrative example can be found in Heiler et al. (2021), who examine the Job Corps (JC) program.[20] They highlight gender differences in program effectiveness, noting that women tend to benefit less than men. The potential explanation is that men and women typically receive different vocational training within JC; men often receive training for higher-paying craft jobs, whereas women more frequently focus on service sector training.

Consider the job training program in Columbia once more. Attanasio et al. (2011) continuously emphasize the gap in program effectiveness between men and women. It is notable, however, that 77 courses (18%) are exclusively dominated by either males or females (Figure 8).

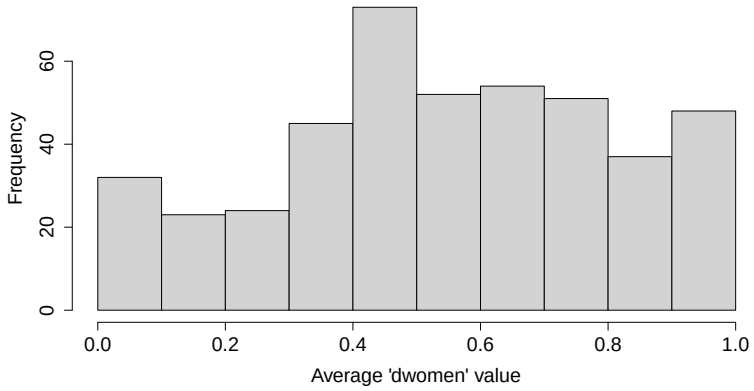


Figure 8: Share of women across the courses

It is not far-fetched that men and women receive inherently different vocational training within the program, potentially contaminating the estimated CATE. Unfortunately, the aggregated data on course participation limits further conclusions.

Heiler et al. (2021) reverse-engineer the identified estimand of $\tau(x)$ in terms of potential outcomes of the effective treatment:

$$
\begin{aligned}
\tau(x) &= \sum_{t \neq 0} E[D_{t,i} Y_i(t) | D_i = 1, X_i = x] - E[Y_i(0) | D_i = 0, X_i = x] \\
&= \sum_{t \neq 0} \underbrace{E[Y_i(t) - Y_i(0) | X_i = x]}_{t-\text{specific CATE}} \frac{e_t(x)}{\sum_{t \neq 0} e_t(x)} \\
&\quad + \sum_{t \neq 0} \underbrace{\{E[Y_i(t) | D_{t,i} = 1, X_i = x] - E[Y_i(t) | X_i = x]\}}_{\text{selection effects } D_i = 1} \frac{e_t(x)}{\sum_{t \neq 0} e_t(x)} \\
&\quad - \underbrace{\{E[Y_i(0) | D_i = 0, X_i = x] - E[Y_i(0) | X_i = x]\}}_{\text{selection effect } D_i = 0} \qquad (5)
\end{aligned}
$$

where $D_{t,i} = 1(T_i = t)$ indicates that unit i is observed in treatment t.

Equation (5) shows that the estimand consists of three components:

1. A weighted average of CATEs of the effective treatments, with weights depending on the *conditional probability* of the respective effective treatment.

[20] Job Corps is the largest training initiative for disadvantaged youth aged 16-24 in the US.

2. A weighted average of effective treatment-specific selection effects.

3. A selection effect into the control group.

The second and third terms are relevant if there's selection into effective treatments even after controlling for observed confounders. In the 'Youth in Action' program, the group indicator C_i and pre-treatment characteristics X_i might not suffice for CIA to hold if not all confounders are available (e.g., health status, family responsibilities).

Additionally, there may be a feedback loop where participation in a course influences future choices, in turn affecting outcomes. For example, if successfully completing a course motivates an individual to take further training.

As noted in Heiler et al., (2021), the decomposition in (5) highlights that the interpretation of the underlying estimand becomes more nuanced in the presence of heterogeneous treatments. Without further assumptions, heterogeneous effects attributed to the binary indicator can be driven by different CATEs, different compositions of the effective treatments, different selection effects of the effective treatments, or combinations thereof.

7 Conclusion

This paper applies double machine learning (DML) methods to replicate and possibly extend a standard program evaluation under the assumption of unconfoundedness. DML allows for the estimation of multiple parameters of interest using the doubly robust score and a combination of standard statistical software.

An alternative estimation method, coined the demeaned augmented inverse propensity weighting (AIPW) estimator, is also applied. In a controlled setting, it effectively addresses unobserved group-invariant effects and successfully identifies treatment effect heterogeneity.

Applying these methods to a job training program in Colombia demonstrates that DML-based methods produce plausible results at a high level of aggregation. However, the transformed data makes interpretation challenging and requires additional adjustments. Overall, several conceptual and implementation issues remain open for investigation and refinement.

When considering cases where a binary treatment represents diverse underlying effective treatments that influence the outcome of interest, it is important to take treatment heterogeneity more seriously and discuss it explicitly in applications, particularly when interpreting heterogeneous effects. Specifically, data collection efforts should document the actual treatments received by participants, even if these treatments are ultimately categorized as binary.

References

Athey, S., and G. W. Imbens (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2), 3–32.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *Annals of Statistics*, 47(2), 1148–78.

Attanasio, O., A. Kugler, and C. Meghir (2011). Subsidizing vocational training for disadvantaged youth in Colombia: Evidence from a randomized trial. *American Economic Journal: Applied Economics*, 3(3), 188–220.

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.

Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., & Syrgkanis, V. (2024). Applied Causal Inference Powered by ML and AI. CausalML-book.org; arXiv:2403.02467.

Gibbons, C. E., Serrato, J. C. S., & Urbancic, M. B. (2019). Broken or Fixed Effects? *Journal of Econometric Methods*, 8(1), 1–12.

Hayfield, T. and J. S. Racine (2008). Nonparametric econometrics: The `np` package. *Journal of Statistical Software* 27(5), 1–32.

Heiler, P. and M. C. Knaus (2021). Effect or Treatment Heterogeneity? Policy Evaluation with Aggregated and Disaggregated Treatments. *arXiv: Econometrics*, arXiv:2110.01427, revised August 2023.

Holland, P. W. (1986). *Statistics and causal inference. Journal of the American Statistical Association*, 81(396), 945–960.

Imai, K., & Kim, I. S. (2019). When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data? *American Journal of Political Science*, 63(2), 467–490.

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press.

Karlan, Dean S., and Jonathan Zinman (2008). Credit Elasticities in Less-Developed Economies: Implications for Microfinance. *American Economic Review*, 98(3), 1040–68.

Knaus, M. C., Lechner, M., & Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *The Econometrics Journal*, 24(1), 134–161. https://doi.org/10.1093/ectj/utaa014

Knaus, M. C. (2022). Double machine learning-based programme evaluation under unconfoundedness. *The Econometrics Journal*, 25(3), 602–627.

Okasa, G. (2022). Meta-Learners for Estimation of Causal Effects: Finite Sample Cross-Fit Performance. *Papers* 2201.12692, arXiv.org.

Racine, J. S. and Z. Nie (2021). *crs: Categorical Regression Splines*. Available at: https://github.com/JeffreyRacine/R-Package-crs/.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688–701.

Sverdrup, E., A. Kanodia, Z. Zhou, S. Athey, and S. Wager (2020). *policytree: Policy learning via doubly robust empirical welfare maximization over trees. Journal of Open Source Software* 5(50), 2232. Available at: https://joss.theoj.org/papers/10.21105/joss.02232.

Wooldridge, Jefrey M. Econometric Analysis of Cross Section and Panel Data. *The MIT Press*, 2010.

Zhou, Z., S. Athey and S. Wager (2018). Offline multi-action policy learning: Generalization and optimization. *arXiv: Machine Learning*, arXiv:1810.04778.

A Appendix

Figure A1: Distribution of Training Courses

Course name	Number of courses	Course name	Number of courses
Inventory and warehouse assistant	18	Agricultural machinery mechanic	1
Taxi/bus driver	2	Cashiers' assistant	2
Electrician	16	Plumbers' assistant	2
Management assistant	3	Seamstress/industrial textile prod.	17
Security guard/building maintenance	8	Library assistant	3
Secretary/administrative assistant	34	Gas station assistant	3
Sales representative	43	Environmental assistant	3
Bakery assistant	5	Organic waste processor	1
Kitchen/cooking assistant	13	Industrial production operator	4
Archival assistant	18	Flower cultivation	5
Pharmacy assistant	6	Metal fabrication	3
Doctor/dentist/nurse assistant	12	Construction operator	1
Carpenter	7	Sports referee	1
IT assistant	13	Senior citizens assistant	6
Clinical lab assistant	2	Marketing assistant	20
Auto/motorcycle mechanic assistant	11	Meat processor	6
Human resources assistant	5	Cleaning services	5
Welding assistant	1	Cattle farming	1
Graphic design assistant	5	Organic farming	1
Refrigeration equipment assistant	1	Waste processor	1
Data entry assistant	14	Packing operator	1
Client relations assistant	16	Shoe repair services	1
Upholster	1	Florist	1
Wooden furniture painter assistant	3	Journeyman	2
Preschool teacher assistant	10	Tourism assistant	1
Accounting assistant	10	Wooden machine operator	1
Foreign trade assistant	2	Molding and foundry worker	2
Beautician	4	Vocational training teacher	2
Mail delivery assistant	10	Journalism assistant	1
Real estate assistant	2	Bank teller	3
Busboy/waiter/waitress	10	Physical rehabilitation	1
Recreation assistant	11	Food processing	4
Call center/telemarketing assistant	5	Quality control assistant	1
Surveyor assistant	9	Worker safety assistant	1
Gas installations	1		

A.1 Identification of the Linear Fixed Effects Model

To identify effects in a linear model, we first impose a functional form assumption:

ASSUMPTION 3: $Y_i(w) = \tau w + C_i + X_i'\beta + U_{Y(w)}, \quad \mathbb{E}[U_{Y(w)} \mid X_i, C_i] = 0, \quad \forall w \in \mathcal{W}$

Potential outcomes are linear functions of confounding variables X_i. The Conditional Expectation Function (CEF) of the potential outcome is then

$$\mathbb{E}[Y_i(w) \mid X_i, C_i] = \tau w + C_i + X_i'\beta.$$

Measured confounding implies that $\mathbb{E}[Y_i(w) \mid X_i, C_i] = \mathbb{E}[Y_i(w) \mid W_i, X_i, C_i]$, i.e., the potential outcome is conditionally mean independent of the treatment. We then rewrite

$$\begin{aligned}\mathbb{E}[Y_i(w) \mid X_i, C_i] &= \tau w + C_i + X_i'\beta + \mathbb{E}[U_{Y(w)} \mid X_i, C_i] \\ &= \tau w + C_i + X_i'\beta + \mathbb{E}[U_{Y(w)} \mid W_i, X_i, C_i].\end{aligned}$$

Measured confounding in combination with linear CEF implies

$$\mathbb{E}[U_{Y(w)} \mid X_i, C_i] = \mathbb{E}[U_{Y(w)} \mid W_i, X_i, C_i] = 0.$$

Under SUTVA and consistency (1), the observed outcome is

$$Y_i = Y_i(W) = \tau W_i + C_i + X_i'\beta + U_{Y(W)},$$

where Y_i is an outcome for person i; C_i are site-by-course fixed effects; and u_i is an error term and X_i are pretreatment characteristics. The CEF of the observed outcome is therefore

$$\begin{aligned}\mathbb{E}[Y_i \mid W_i, X_i, C_i] &= \tau W_i + C_i + X_i'\beta + \mathbb{E}[U_{Y(W)} \mid W_i, X_i, C_i] \\ &= \tau W_i + C_i + X_i'\beta.\end{aligned}$$

The unknown parameter of interest τ can be obtained as the solution of a least squares problem by adding group dummies into the regression or equivalently by using transformed variables. Standard errors in that case can be easily adjusted to account for the loss of degrees of freedom.

A.2 Integrating Program Costs into Policy Recommendations

Accounting for the program cost is not straightforward, since gains are realized over a long period. Theoretically, one could compute them in the following way: compute the future accumulated benefit by adjusting the AIPW score accordingly.

Considering 40 more years of work, the discounted amount of earned extra money would be calculated as follows: the estimated monthly salary gain is \$26,732, which over a year amounts to \$26,732 $\times$ 12 = \$320,784. Dividing this by the exchange rate 1970 results in \$162.83 per year. Over 40 years, this totals to \$162.83 $\times$ 40 = \$6,513.20 under scenario A (no amortization) and \$3,685.36 under scenario B (5% yearly amortization). Subtracting the program costs of \$750, the score is reduced by 52% and 81% under scenarios A and B, respectively. We then adjust the AIPW score for salary by the same amount and re-estimate the policy tree.

A.3 Fixed Effects as a Biased Estimator of ATE

Start with the equation from Gibbons:

$$\hat{\tau}^{FE} \xrightarrow{p} \tau^{ATE} + \sum_c \pi_c \tau_c \left[\frac{\text{Var}\left(\widetilde{W}_i \mid C = c\right)}{\text{Var}\left(\widetilde{W}_i\right)} - 1 \right]$$

Consider that ATE is defined as $\tau^{ATE} = \sum_c \pi_c \tau_c$, which allows us to factor out $\sum_c \pi_c \tau_c$ and simplify the expression to:

$$\hat{\tau}^{FE} \xrightarrow{p} \sum_c \pi_c \tau_c \left[\frac{\text{Var}\left(\widetilde{W}_i \mid C = c\right)}{\text{Var}\left(\widetilde{W}_i\right)} \right]$$

Since the demeaned treatment indicator $\widetilde{W}$ is a binary variable, its variance can be written as

$$\hat{\tau}^{FE} \xrightarrow{p} \sum_c \pi_c \tau_c \frac{P_c(1 - P_c)}{P(1 - P)}$$

where P_c is the share of treated within the course c.

This equation comes very close to the one reported in Attanasio, if one corrects the weighting fot the population size:

$$\hat{\tau}^{FE} = \frac{\sum_c P_c(1 - P_c)\pi_c \tau_c}{\sum_c P_c(1 - P_c)}$$

To connect the two equations, we decompose the total variance $P(1 - P)$. According to the law of total variance:

$$\text{Var}(W) = \mathbb{E}[\text{Var}(W \mid C)] + \text{Var}(\mathbb{E}[W \mid C])$$

For a binary treatment variable W, we have $\text{Var}(W \mid C = c) = P_c(1 - P_c)$ and $\mathbb{E}[W \mid C = c] = P_c$. Thus, the total variance $P(1 - P)$ can be decomposed as:

$$P(1 - P) = \sum_c \pi_c P_c(1 - P_c) + \sum_c \pi_c (P_c - P)^2$$

Here, $\sum_c \pi_c P_c(1-P_c)$ is the weighted sum of the within-group variances, and $\sum_c \pi_c (P_c - P)^2$ represents the between-group variance. Starting with the first equation and including the total variance decomposition:

$$\hat{\tau}^{FE} \xrightarrow{p} \sum_c \pi_c \tau_c \frac{P_c(1 - P_c)}{\sum_c \pi_c P_c(1 - P_c) + \sum_c \pi_c (P_c - P)^2}$$

If one assumes that the between-group variance component is small and can be ignored (hinted in Attanasio), then

$$\hat{\tau}^{FE} \xrightarrow{p} \sum_c \pi_c \tau_c \frac{P_c(1 - P_c)}{\sum_c \pi_c P_c(1 - P_c)}$$

Notice that $\sum_c \pi_c = 1$ by definition, because π_c are population frequencies. Then, we further rearrange:

$$\hat{\tau}^{FE} \xrightarrow{p} \frac{\sum_c P_c(1 - P_c)\pi_c \tau_c}{\sum_c P_c(1 - P_c)}$$

Table A1: Fixed effects as a biased estimator of ATE, simulation

DGP 1	
Covariates	$p = 6$ independent covariates $X_1, \ldots, X_6$ drawn from a normal distribution: $X_k \sim N(k, 1), k \in \{1, 2, \ldots, 6\}$
Treatment	$W \sim$ Bernoulli(0.8): 80% treated
Treatment Effect	$\tau = 1.5$
Outcome (Control)	$Y(0) = \alpha_i X_4^2 - 0.3X_4 + 0.2X_5 + 0.1X_6 + \varepsilon$, with $\varepsilon \sim N(0, 5)$
Outcome (Treated)	$Y(1) = \tau W + Y(0)$
Unit Fixed Effect	$\alpha_i \sim \mathcal{U}(-5, 5)$
DGP 2	
Covariates	$p = 6$ independent covariates $X_1, \ldots, X_6$ drawn from a normal distribution: $X_k \sim N(7 - k, 1), k \in \{1, 2, \ldots, 6\}$
Treatment	$W \sim$ Bernoulli(0.5): 50% treated
Treatment Effect	$\tau = -1.5$
Outcome (Control)	$Y(0) = \alpha_i X_4^2 - 0.3X_1 + 0.2X_2 + 0.1X_3 + \alpha_i + \varepsilon$, with $\varepsilon \sim N(0, 5)$
Outcome (Treated)	$Y(1) = \tau W + Y(0)$
Unit Fixed Effect	$\alpha_i \sim \mathcal{U}(-5, 5)$

A.4 MSE Decomposition

When the true parameter θ varies across simulations (i.e., it is different for each simulation), the mean squared error (MSE) of an estimator $\hat{\theta}$ can be decomposed into three components: bias squared, variance, and the covariance between the estimator and the true parameter.

Let's denote the true parameter vector as θ, and the estimator of θ as $\hat{\theta}$. The MSE of the estimator $\hat{\theta}$ can be decomposed as follows:

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^T(\hat{\theta} - \theta)]$$

Expanding this expression gives:

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] \\
&= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta)^2\right] \\
&= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\right] + (\mathbb{E}[\hat{\theta}] - \theta)^2 + 2\mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}] - \theta)\right]
\end{aligned}$$

The expression can be broken down to:

1. **Bias squared**: $(\mathbb{E}[\hat{\theta}] - \theta)^2$
2. **Variance**: $\mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$
3. **Covariance term**: $2\mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}] - \theta)]$

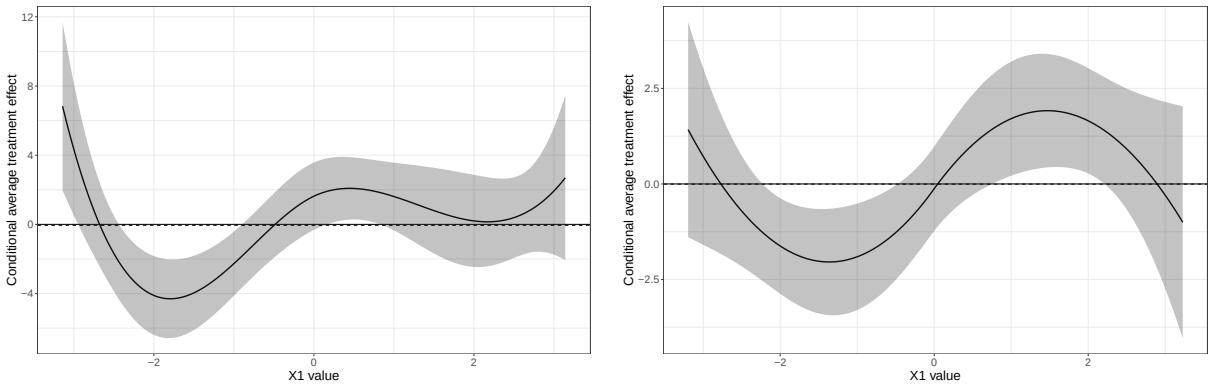
The last term is special for the case where the true parameter θ varies across simulations. It captures the covariance between the estimator $\hat{\theta}$ and the true parameter θ and reflects how the bias of the estimator correlates with the true parameter across different simulations.

Importantly, we report the absolute bias, standard deviation, and MSE, noting that MSE is not equal to the squared bias plus the estimator's variance.

A.5 Simulation Results for Heterogeneity Variables

Case 1	1,000 observations			4,000 observations		
	MSE	|Bias|	SD	MSE	|Bias|	SD
$\mathbf{X_1}$						
POLS	2.77	1.22	1.66	0.71	0.61	0.84
FE	0.81	0.69	0.90	0.20	0.34	0.44
AIPW	0.70	0.65	0.84	0.08	0.22	0.28
$\mathbf{X_2}$						
POLS	2.67	1.25	1.64	0.67	0.62	0.82
FE	0.85	0.71	0.92	0.20	0.33	0.44
AIPW	0.73	0.66	0.85	0.07	0.21	0.27
$\mathbf{X_3}$						
POLS	4.77	1.63	2.19	1.39	0.87	1.18
FE	1.84	1.04	1.36	0.43	0.50	0.66
AIPW	2.23	1.13	1.49	0.23	0.37	0.48
$\mathbf{X_4}$						
POLS	14.02	2.70	3.73	4.33	1.49	2.08
FE	3.45	1.42	1.86	0.88	0.73	0.94
AIPW	2.61	1.25	1.62	0.28	0.41	0.53

Table A2: Detected Effect Heterogeneity w.r.t. X (Case 1)



(a) AIPW estimator on non-demeaned data (b) AIPW estimator on group-demeaned data

Figure A2: Spline regressions for raw and transformed data, $Y = \widetilde{Y}_{ATE}$

Case 2	1,000 observations			4,000 observations		
	MSE	|Bias|	SD	MSE	|Bias|	SD
$\mathbf{X_1}$						
POLS	3.94	1.59	1.98	0.92	0.77	0.96
FE	0.95	0.76	0.97	0.21	0.36	0.46
AIPW	1.13	0.84	1.06	0.25	0.40	0.50
$\mathbf{X_2}$						
POLS	4.01	1.56	2.00	0.92	0.75	0.96
FE	0.89	0.74	0.94	0.23	0.38	0.48
AIPW	1.07	0.82	1.04	0.26	0.40	0.52
$\mathbf{X_3}$						
POLS	7.02	2.11	2.65	1.75	1.04	1.32
FE	1.97	1.14	1.40	0.46	0.54	0.68
AIPW	3.32	1.46	1.82	0.76	0.69	0.87
$\mathbf{X_4}$						
POLS	24.32	3.82	4.93	5.29	1.84	2.30
FE	4.96	1.77	2.23	1.03	0.81	1.02
AIPW	4.78	1.75	2.19	0.96	0.79	0.98

Table A3: Detected Effect Heterogeneity w.r.t. X (Case 2)

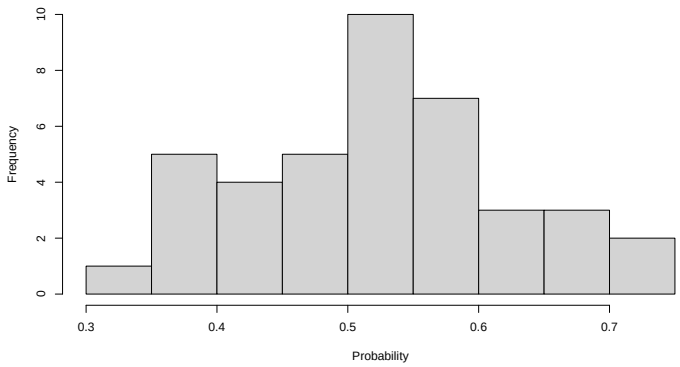


Figure A3: Distribution of Propensity Scores across Courses (Case 2)

DGP: Case 2	
Covariates	$p = 4$ independent covariates $X_1, \ldots, X_4$, where $X_k \sim N(k, 1) + R_i, k \in \{1, 2, 3\}$ and $X_4 \sim$ Bernoulli(0.5)
Treatment	$W \sim$ Bernoulli(0.5)
CATE function	$\tau(X) = -0.5 - 0.8X_1 + 0.8X_2 + 0.3X_3 - 0.6X_4$
Outcome (Treated)	$Y(1) = \tau(X)W + Y(0)$
Outcome (Control)	$Y(0) = u_i + \boldsymbol{\beta X} + \varepsilon$, with $\varepsilon \sim N(0, 5)$
Unit Fixed Effect	$u_i \sim$ Uniform$(-4, 4)$; $R_i \sim$ Uniform$(-2.5, 2.5)$
Confounding	$\boldsymbol{\beta X} = 0.1X_1^2 - 0.1X_2 + 0.5\sin(X_3) - 0.5\exp(X_4)$

Table A4: Data Generating Process (Case 2)

DGP: Case 3	
Covariates	$p = 10$ independent covariates $X_1, \ldots, X_{10}$ drawn from a uniform distribution: $X_k \sim$ Uniform$(-\pi, \pi)$
Propensity Score	$W \sim$ Bernoulli$(\Phi(\sin(X_1)))$, where $\Phi(\cdot)$ is the standard normal cdf
Outcome (Control)	$Y(0) = \cos(X_1 + \frac{1}{2}\pi) + u_i \times 0.5X^2 + \varepsilon$, with $\varepsilon \sim N(0, 5)$
Outcome (Treated)	$Y(1) = \sin(X_1) + u_i \times 0.5X^2 + \varepsilon$, with $\varepsilon \sim N(0, 5)$
CATE Function	$\tau(X) = \sin(X_1) - \cos(X_1 + \frac{1}{2}\pi)$
Unit Effects	$u_i \sim N(0, 5)$

Table A5: Data Generating Process (Case 3)