Effect or Treatment Heterogeneity? Policy Evaluation with Aggregated and Disaggregated Treatments

Phillip Heiler[†] Michael C. Knaus[‡]

September 28, 2021

Abstract

Binary treatments in empirical practice are often (i) ex-post aggregates of multiple treatments or (ii) can be disaggregated into multiple treatment versions after assignment. In such cases it is unclear whether estimated heterogeneous effects are driven by effect heterogeneity or by treatment heterogeneity. This paper provides estimands to decompose canonical effect heterogeneity into the effect heterogeneity driven by different responses to underlying multiple treatments and potentially different compositions of these underlying effective treatments. This allows to avoid spurious discovery of heterogeneous effects, to detect potentially masked heterogeneity, and to evaluate the underlying assignment mechanism of treatment versions. A nonparametric method for estimation and statistical inference of the decomposition parameters is proposed. The framework allows for the use of machine learning techniques to adjust for high-dimensional confounding of the effective treatments. It can be used to conduct simple joint hypothesis tests for effect heterogeneity that consider all effective treatments simultaneously and circumvent multiple testing procedures. It requires weaker overlap assumptions compared to conventional multi-valued treatment effect analysis. The method is applied to a reevaluation of heterogeneous effects of smoking on birth weight. We find that parts of the differences between ethnic and age groups can be explained by different smoking intensities. We further reassess the gender gap in the effectiveness of the Job Corps training program and find that it is largely explained by gender differences in the type of vocational training received.

Keywords: causal inference, debiased/orthogonal estimation, heterogeneous treatment effects, causal machine learning, double machine learning, treatment versions

JEL classification: C14, C21

*Michael Knaus gratefully acknowledges financial support from the Swiss National Science Foundation (SNSF) (grant number SNSF 407740_187301). The paper was circulated and presented previously under different titles and we would like to thank Martin Huber, Michael Lechner, Jana Mareckova, Julian Schüssler, Anthony Strittmatter, and the participants of sessions at IAAE2021, ESEM2021, COMPIE2021, and the FFHKT Econometrics Seminar for valuable comments and discussions. All remaining errors are ours.

[†]Aarhus University, Department of Economics and Business Economics, CREATES, TrygFonden's Centre for Child Research, Fuglesangs Allé 4, 8210 Aarhus V, Denmark, pheiler@econ.au.dk.

[‡]University of St. Gallen, Varnbüelstrasse 14, 9000 St. Gallen, Switzerland. Michael C. Knaus is also affiliated with IZA, Bonn, michael.knaus@unisg.ch.

1 Introduction

The analysis of causal effects is at the heart of empirical research in economics, political science, the biomedical sciences, and beyond. To evaluate and design policies, interventions, or programs for observational units with different background characteristics, it is necessary to develop a thorough understanding of the heterogeneities present in causal relationships. There is now a large literature that develops and applies identification and estimation strategies for causal or treatment parameters that explicitly take into account these heterogeneities (see for recent overviews Athey & Imbens, 2017; Abadie & Cattaneo, 2018).

Most attention is put on the analysis of *effect heterogeneity* of binary treatments, while less is given to *treatment heterogeneity*. In particular, many analyzed binary treatments in applications can be conceived as actually being non-homogeneous in the sense that they summarize underlying multi-valued treatments that directly impact the outcome of interest. In such cases it is not clear whether effect heterogeneity defined in the canonical binary treatment setting reflects heterogeneous effects or heterogeneity in the effective treatments. This paper proposes new estimands and estimators to disentangle these two sources of heterogeneity in a variety of settings where the analyzed binary indicator does not coincide with the effective treatment. The distinction between these sources of heterogeneity is crucial for evaluating and improving assignment mechanisms. As leading illustrations, consider the following two common scenarios where the analyzed treatment indicator is binary but the effective treatment is multi-valued:[1]

Scenario 1 (binarized treatments): Multiple or continuous treatments are *ex post* subsumed into a binary indicator (e.g. different smoking intensities are subsumed into "smoking yes/no"). The motivations for such aggregations are manifold and include simplicity, convenience or data availability. This can have unintentional consequences: First, discovered effect heterogeneity could be a spurious byproduct of aggregation and thus falsely attributed to unit background characteristics. Second, actual effect heterogeneity could be masked as a consequence of the aggregation.

Scenario 2 (multiple treatment versions): A binary treatment takes different versions after (potentially random) assignment (e.g. access to a job training program where different

[1]See also Appendix A for a motivating toy example.

specializations are possible). Here effect heterogeneity could result from a better version targeting across groups and not from different effectiveness of the versions themselves. Understanding this difference is crucial for policy makers to assess quality and potentially fairness of the version assignment mechanism in place.

In this paper we propose a novel method for decomposing effect heterogeneity in a more general scenario with observed confounders. We decompose canonical effect heterogeneity into new estimands that are representative of (i) heterogeneous effects and (ii) heterogeneity stemming from different underlying treatment compositions. These decomposition parameters can help to diagnose and understand the impact of (dis)aggregating treatment variables. They can also serve as summary measures to evaluate the impact of (dis)aggregating treatments on the causal analysis both on a global or on conditional levels based on unit level characteristics such as age, gender, or income. Furthermore they can be used by policy makers to evaluate the quality of treatment version assignment and its potential heterogeneity across different groups.

We propose a simple but flexible nonparametric method for estimation of the decomposition parameters and show how to conduct valid statistical inference for their heterogeneity along a low-dimensional set of background characteristics. Our framework allows for the use of machine learning techniques such as random forests, deep neural networks, high-dimensional sparse regression or likelihood models in the estimation of the required nuisance parameters. It can be used to conduct simple joint hypothesis tests for global or conditional decomposition parameters that consider the potentially many effective treatments simultaneously. This allows us to test necessary conditions for different types of heterogeneity without the need for multiple testing procedures.

The large sample theory extends beyond the decomposition parameters considered in this paper. In particular it demonstrates that, for parameters of a specific structure such as our decomposition parameters, flexible machine learning estimators for causal parameters can be combined with weighting schemes that are themselves estimated and still obtain asymptotically normally distributed estimators. This asymptotic property can then be exploited to construct simple analytical confidence intervals. Our Monte Carlo simulations suggest that the proposed intervals have coverage rates close to the nominal

level in finite samples.

The decomposition estimators also have superior large sample properties compared to the ones employed in conventional multi-valued treatment effect analysis when there are many treatments. The commitment to the decomposition parameters instead of considering all possible treatment comparisons allows us to relax the overlap conditions compared to conventional multi-valued treatment effect estimators and to consider potentially many effective treatments.

We provide two applications of our decomposition method, one for each of the leading scenarios. For the first scenario, we show that parts of the common finding that the detrimental effect of smoking on birth weight is largest for white mothers can be explained by white mothers smoking more heavily conditional on being smokers. Similarly, different effects for different age groups are partly due to the fact that teenage mothers smoke less intensive than older mothers. For the second scenario, we investigate the lower effectiveness of access to the Job Corps training program for women compared to men. We find evidence that this gender difference is largely explained by the vocational training curriculum, which focuses more on lower paying service jobs for women and more on higher paying craft jobs for men. Imposing the same mix of vocational training as part of our decomposition method removes around three quarters of the total gender differences in the effect on earnings.

The paper is structured as follows: Section 2 discusses the related literature. Section 3 outlines the decomposition of the causal effect parameters and discusses their identification. Section 4 contains the estimation and inference method. Section 5 introduces the technical assumptions and discusses the large sample properties of the proposed methodology. Section 6 contains some discussion regarding the comparison to multi-valued treatment effect analysis. Section 7 provides the Monte Carlo study. Section 8 contains the application. Section 9 provides some concluding remarks.

2 Related Literature

The proposed decomposition estimands for heterogeneous effects complement the literature that consider (dis)aggregated binary treatments. For cases like scenario 1, Lechner (2002) discusses how to meaningfully aggregate average effects of multiple treatments into composite treatment effects. Hotz, Imbens, and Mortimer (2005) and Hotz, Imbens, and Klerman (2006) consider the issue of collecting different training components into a binary indicator. They emphasize the potential lack of external validity due to the latent treatment heterogeneities. Similarly, a more recent stream of papers formalizes the underlying causal structure and clarifies the interpretation of such compound treatments (Cole & Frangakis, 2009; VanderWeele, 2009; Hernán & VanderWeele, 2011; Petersen, 2011). VanderWeele and Hernan (2013) note that non-homogeneous treatments violate the "Stable Unit Treatment Value Assumption" (SUTVA) of Rubin (1980) that consists of two parts: (i) no interference, which rules out that potential outcomes of a unit depend on the treatment status of other units, and (ii) no-multiple-versions-of-treatment, which requires a homogeneous treatment or at least the treatment variation irrelevance assumption of VanderWeele (2009). While the first part of SUTVA is discussed and addressed in numerous studies, the second part is often ignored. Thus, VanderWeele and Hernan (2013) formalize a setting where this assumption is violated and provide several new identification results and estimands. The consequences of aggregating heterogeneous causal effects have also been discussed in the context of instrumental variables estimation (e.g. Angrist & Imbens, 1995; Marshall, 2016; Andresen & Huber, 2021). While the above papers discuss the consequences of (dis)aggregated treatments for average or unconditional causal effects, we focus on effect heterogeneity and the additional complications arising there.

The focus on effect heterogeneity is motivated by and related to the surging literature that develops (e.g. Imai & Ratkovic, 2013; Tian, Alizadeh, Gentles, & Tibshirani, 2014; Athey & Imbens, 2016; Wager & Athey, 2018; Athey, Tibshirani, & Wager, 2019; Künzel, Sekhon, Bickel, & Yu, 2019; Knaus, Lechner, & Strittmatter, 2021; Nie & Wager, 2021) and applies (e.g. Davis & Heller, 2020; Knaus, Lechner, & Strittmatter, 2020) flexible machine learning inspired methods to the estimation of heterogeneous causal effects. We

build on the de-biased/double machine learning framework by Chernozhukov et al. (2018). They use doubly robust score functions and sample splitting in conjunction with machine learning methods to develop a theory for estimation and statistical inference on average effect parameters. There is now a series of papers which uses such score functions as pseudo-outcomes in auxiliary regression models to estimate conditional effect parameters (Lee, Okui, & Whang, 2017; Zimmert & Lechner, 2019; Fan, Hsu, Lieli, & Zhang, 2020; Semenova & Chernozhukov, 2021; Kennedy, 2020; Knaus, 2020; Curth & van der Schaar, 2021). These methods obtain functional parameters of interest by localizing the robust score function in (nonparametric) regression or machine learning frameworks. Our theoretical contribution is most directly linked to Semenova and Chernozhukov (2021) who analyze the properties of linear predictor approximations for a variety of structural functions. Like them we exploit approximation theory in the literature on series regression such as Newey (1997) and Belloni, Chernozhukov, Chetverikov, and Kato (2015). We extend some of the inferential results by Semenova and Chernozhukov (2021) to settings where pseudo-outcomes are constructed as a weighted average of Neyman-orthogonal scores with estimated weights, which might be of interest beyond the application to the decomposition parameters of this paper.

Many studies document the detrimental effect of smoking during pregnancy on birth weight (e.g. Almond, Chay, & Lee, 2005; Abrevaya, 2006; Cattaneo, 2010; Almond & Currie, 2011). Our methodology allows us to understand how much of the heterogeneous effects of this binarized treatment is spuriously attributed to subgroup characteristics instead of to different intensities of smoking across these subgroups.

Previous studies evaluate the US training program Job Corps from different angles based on a large scale experiment (e.g. Schochet, Burghardt, & Glazerman, 2001; Schochet, Burghardt, & McConnell, 2008; Flores, Flores-Lagunes, Gonzalez, & Neumann, 2012; Eren & Ozbelik, 2014; Strittmatter, 2019). While most of them document and discuss heterogeneous effects, our new decomposition allows to disentangle how much of the effects and their heterogeneity is driven by selection into different curricula. This provides a complementary angle to the existing evaluations and illustrates how policy makers can learn more about the quality of the existing assignment mechanism of treatment versions.

3 Decomposition and Identification

3.1 The Setting

Assume we observe independent data (Y_i, D_i, T_i, X_i) for $i = 1, \ldots, n$. Y_i denotes the outcome of interest, $D_i \in \{0, 1\}$ is the analyzed binary indicator, $T_i \in \mathcal{T} = \{0, 1, \ldots, J\}$ indicates the effective treatment, and X_i contains confounding variables. We consider settings that are characterized by two features: (i) Not D_i, but the effective treatment T_i has a direct influence on the outcome creating potential outcomes $Y_i(t)$ for each $t \in \mathcal{T}$. Thus, we assume SUTVA with respect to the effective treatment such that $Y_i = \sum_t \mathbb{1}[T_i = t] Y_i(t)$. (ii) Conditional on T_i, the binary indicator D_i is deterministic, i.e. it perfectly separates the support $\mathcal{T}$. We use directed acyclic graphs (DAGs) (see e.g. Pearl, 1995) to represent and formalize our main scenarios in this setting.

Figure 1: Analyzed indicator is ex-post aggregate of confounded multiple treatment:

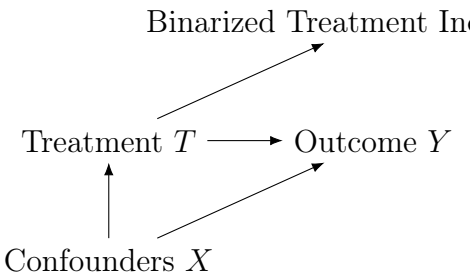
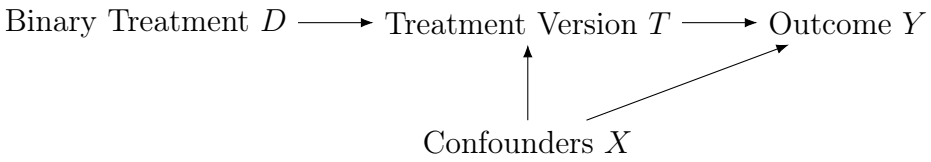


Figure 1 outlines the causal structure of *Scenario 1* where the binary indicator variable D_i is the result of an ex-post aggregation and not structurally related to the outcome. In practice, this aggregation is often conducted after the outcome realizes, which makes it unlikely for D_i to affect Y_i directly. This is indicated by a missing arrow from D_i to Y_i. However, as T_i is ancestor of both binarized indicator D_i and outcome Y_i, they are not statistically independent of each other even conditional on X_i. For example, a statistical relationship between birth weight (Y_i) and smoking (D_i) as an aggregate of the consumed dose of cigarettes (T_i) can be derived from observational data. Conditional on the number of cigarettes, however, smoking is deterministic and no association remains.

The DAG in Figure 2 depicts the causal structure of *Scenario 2* where a randomized binary treatment D_i precedes the confounded allocation of treatment versions T_i. Here, D_i is not an ex-post variable with regards to Y_i. Y_i and D_i are associated as the latter

Figure 2: Randomized binary treatment precedes confounded treatment versions:



determines which treatment versions are available, but has no direct effect beyond that. Its effect is completely mediated through the treatment versions T_i. For example, this implies that any association between D_i being access to a training program (yes/no) on earnings Y_i would disappear if we would condition on all training types including no training access (T_i).

It is important to note that, while conceptually different in terms of the causal interpretation, the DAGs in Figures 1 and 2 imply the same conditional independence relationships regarding D_i, T_i, and potential outcomes $Y_i(t)$. In standard Neyman-Rubin notation for multi-valued treatments (Rubin, 1974; Imbens, 2000; Lechner, 2001; Cattaneo, 2010), we have that

$$Y_i(0), Y_i(1) \ldots, Y_i(J) \perp\!\!\!\perp D_i \mid T_i \tag{1}$$

$$Y_i(0), Y_i(1) \ldots, Y_i(J) \perp\!\!\!\perp T_i \mid X_i \tag{2}$$

where (1) is a consequence of our setting that D_i is a constant given T_i and (2) follows from the causal structure encoded in the DAG (see Appendix B.1). Thus, from a statistical identification point of view, we treat both scenarios as being equivalent in the following. Note that conditions (1) and (2) and everything that follows can be extended to apply to causal graphs where additional observed confounders can affect D_i, T_i, and Y_i simultaneously (see Appendix B.2).

We use $D_{t,i} = \mathbb{1}[T_i = t]$ to indicate that unit i is observed in treatment t and let $e_t(x) = P(D_{t,i} = 1 | X_i = x)$ denote the corresponding generalized propensity scores. Without loss of generality, we assume throughout that $t = 0$ denotes a homogeneous control condition. Thus the binary indicator is defined as $D_i = \sum_{t \neq 0} D_{t,i}$ and $D_{0,i} = 1 - D_i$ in what follows.

3.2 Heterogeneous effects if treatment heterogeneity is ignored

We are interested in the case where the causal structure is accurately described by DAGs like in the previous section, but the analyst imposes the canonical setting considering only the binary indicator D_i, which is deterministic in T_i. Usually the analyst is interested in the conditional average treatment effect ($CATE$) $\tau(x)$ or aggregations thereof like the average treatment effect ($ATE = E[\tau(X_i)]$). However, the potential outcome under the binary indicator being one is not uniquely defined in our setting unless $J = 1$ and everything collapses to the standard case. The question is then, what does the quantity $\tau(x) = E[Y_i|D_i = 1, X_i = x] - E[Y_i|D_i = 0, X_i = x]$ identify, which is commonly deployed based on the standard identification assuming strong ignorability for D_i (see e.g. Imbens & Rubin, 2015). Given the setting outlined in Section 3.1 we can backwards engineer the actually identified estimand in terms of potential outcomes from the effective treatment:[2]

$$
\begin{aligned}
\tau(x) &= E[Y_i|D_i = 1, X_i = x] - E[Y_i|D_i = 0, X_i = x] \\
&= \sum_{t\neq 0} E\left[D_{t,i}Y_i(t)|D_i = 1, X_i = x\right] - E[Y_i(0)|X_i = x] \\
&= \sum_{t\neq 0} E[Y_i(t)|D_{t,i} = 1, D_i = 1, X_i = x]P(D_{t,i} = 1|X_i, D_i = 1) - E[Y_i(0)|X_i = x] \\
&= \sum_{t\neq 0} E[Y_i(t)|D_{t,i} = 1, X_i = x]\frac{e_t(x)}{\sum_{t\neq 0} e_t(x)} - E[Y_i(0)|X_i] \\
&= \sum_{t\neq 0} \underbrace{E[Y_i(t) - Y_i(0)|X_i = x]}_{t-\text{specific CATE}} \frac{e_t(x)}{\sum_{t\neq 0} e_t(x)} \\
&\quad + \sum_{t\neq 0} \underbrace{\{E[Y_i(t)|D_{t,i} = 1, X_i = x] - E[Y_i(t)|X_i = x]\}}_{\text{selection effect}} \frac{e_t(x)}{\sum_{t\neq 0} e_t(x)}
\end{aligned} \tag{3}
$$

Equation (3) shows that the actually identified estimand consists of two components. The first one is a weighted average of $CATEs$ of the effective treatments, $\tau_t(x) = E[Y_i(t) - Y_i(0)|X_i = x]$, where the weights depend on the conditional probability of being treated in the respective effective treatment. The second one is a weighted

[2] We use that $Y_i = \sum_t D_{t,i}Y_i(t)$ and that for $t \neq 0$ $P(D_i = 1|D_{t,i} = 1, X_i) = 1$ holds by definition and thus by Bayes' Law:

$$
P(D_{t,i} = 1|D_i = 1, X_i) = \frac{P(D_{t,i} = 1|X_i)}{\sum_{t\neq 0} P(D_{t,i} = 1|X_i)} = \frac{e_t(X_i)}{\sum_{t\neq 0} e_t(X_i)}
$$

average of effective treatment specific selection effects. The selection effects are positive if those with characteristics x who are actually observed in treatment t show higher potential outcomes than the general population described by x, or negative if vice versa. This term becomes relevant if the considered confounding variables matter beyond the selection into the binary indicator. This can e.g. occur in the case of a randomized binary treatment in Scenario 2 where the selected X_i might not include all confounders for the treatment versions.

The decomposition in (3) highlights that the interpretation of the underlying estimand becomes more nuanced in the presence of heterogeneous treatments. What is supposed to be an easily interpretable $CATE$ depends now on the potentially unknown distribution of effective treatments and selection into those treatments. Thus, 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.

To be able to meaningfully decompose estimand (3) below, we impose a strong ignorability assumption at the effective treatment level:

Assumption 1 *(strong ignorability of effective treatment)*

(a) Unconfoundedness: $Y_i(t) \perp\!\!\!\perp D_{t,i} | X_i = x, \forall\ t \in \mathcal{T}$ *and* $x \in \mathcal{X}$.

(b) Common support: $0 < P[D_{t,i} = 1 | X_i = x] \equiv e_t(x), \forall\ t \in \mathcal{T}$ *and* $x \in \mathcal{X}$.

Assumption 1 is a standard assumption in the multiple treatments setting (Imbens, 2000; Lechner, 2001). It imposes that a) the set of conditioning variables is rich enough such that after conditioning all residual variation in potential outcomes is independent of the allocated effective treatment and b) there are comparable units across effective treatments in terms of their confounders. Under this assumption, the selection effects in (3) disappear and the underlying estimand becomes

$$\tau(x) = \sum_{t \neq 0} \frac{e_t(x)}{\sum_{t \neq 0} e_t(x)} \tau_t(x) \equiv nATE(x). \tag{4}$$

We call this estimand the natural conditional average treatment effect ($nATE(x)$) because it is the result of the actual or "natural" effective treatment composition. However,

even under Assumption 1, the differences between units characterized by x and x' can result from different treatment shares, different treatment $CATEs$, or both. We thus could detect seemingly heterogeneous effects, even if the treatment $CATEs$ are constant within treatments but not homogeneous between treatments, i.e. $\tau_t(x) = \tau_t \; \forall \; t \in \mathcal{T}, x \in \mathcal{X}$ but $\tau_t \neq const. \; \forall \; t \in \mathcal{T}$, as long as the probabilities to be observed in the different effective treatments are heterogeneous. In this case any difference is driven by treatment heterogeneity:

$$nATE(x) - nATE(x') = \sum_{t \neq 0} \left[\frac{e_t(x)}{\sum_{t \neq 0} e_t(x)} - \frac{e_t(x')}{\sum_{t \neq 0} e_t(x')} \right] \tau_t \tag{5}$$

This should be kept in mind when interpreting heterogeneous effects even if the underlying effective treatments are not observable. If they are observable, however, we can further decompose heterogeneous effects of the binary indicator in what follows.

3.3 The Decomposition

In this section we demonstrate how to disentangle actual effect heterogeneity and heterogeneity driven by heterogeneous selection into effective treatments. We propose to decompose the $nATE(x)$ in two parts:

$$\underbrace{\sum_{t \neq 0} \frac{e_t(x)}{\sum_{t \neq 0} e_t(x)} \tau_t(x)}_{nATE(x)} = \underbrace{\sum_{t \neq 0} \frac{\pi_t}{\sum_{t \neq 0} \pi_t} \tau_t(x)}_{rATE(x)} + \underbrace{\sum_{t \neq 0} \left(\frac{e_t(x)}{\sum_{t \neq 0} e_t(x)} - \frac{\pi_t}{\sum_{t \neq 0} \pi_t} \right) \tau_t(x)}_{\Delta(x)} \tag{6}$$

where $\pi_t = E[D_{t,i}]$ are the unconditional treatment probabilities. The first component on the right hand side fixes the composition of the effective treatments at the population value. It resembles a situation where effective treatments are randomly allocated using the population level selection probabilities. Thus, we refer to it as the random conditional average treatment effect ($rATE(x)$). All heterogeneities in $rATE(x)$ are thus driven by "real" effect heterogeneity within treatments, $\tau_t(x) \neq \tau_t(x')$ for some $x, x' \in \mathcal{X}$, as the underlying treatment composition is held fixed. In other words differences in $rATE(x)$ describe effect heterogeneity *compositionis paribus*. Thus, we can consider heterogeneity

in $rATE(x)$ as possibility to test necessary conditions for classic (or "within") effect heterogeneity.

The second component of the decomposition $\Delta(x)$ is the part of $nATE(x)$ resulting from the interaction of non-constant effective treatment probabilities and different effective treatments having different effects ("between" treatment effect heterogeneity). Thus, the decomposition is redundant, i.e. $\Delta(x) = 0 \ \forall \ x \in \mathcal{X}$, under (i) effective treatment composition homogeneity $\frac{e_t(x)}{\sum_{t\neq 0} e_t(x)} - \frac{\pi_t}{\sum_{t\neq 0} \pi_t} = 0 \ \forall \ t \in \mathcal{T}$ and $x \in \mathcal{X}$, (ii) treatment variation irrelevance $E[Y_i(t)|X_i = x] = E[Y_i(t')|X_i = x] \ \forall \ x \in \mathcal{X}, \ t, t' \in \mathcal{T}$ of VanderWeele (2009)[3], or if (iii) positive and negative components of the sum net out to zero. Hence, $\Delta(x) \neq 0$ is a necessary condition for unequal treatment probabilities and between treatment effect heterogeneity. Furthermore, heterogeneity in $\Delta(x)$ is a necessary condition for heterogeneous assignment probabilities, within treatment effect heterogeneity, or both. Thus, the decomposition can be used to address a variety of relevant policy questions, see also Remark 2 below. Moreover, the focus on such necessary conditions offers statistical advantages over testing similar conditions in the standard multi-valued treatment effect setup when there are many effective treatments. We return to this point in Section 6.

Under Assumption 1, the conditional average potential outcome of treatment t is identified as $\mu_t(X_i) \equiv E[Y_i(t)|X_i] = E[Y_i(t)|D_{t,i} = 1, X_i] = E[Y_i|D_{t,i} = 1, X_i]$ and accordingly the decomposition terms are identified as:

$$
\begin{aligned}
nATE(x) &= \sum_{t\neq 0} \frac{e_t(x)}{\sum_{t\neq 0} e_t(x)} (\mu_t(x) - \mu_0(x)) \\
rATE(x) &= \sum_{t\neq 0} \frac{\pi_t}{\sum_{t\neq 0} \pi_t} (\mu_t(x) - \mu_0(x)) \\
\Delta(x) &= \sum_{t\neq 0} \left(\frac{e_t(x)}{\sum_{t\neq 0} e_t(x)} - \frac{\pi_t}{\sum_{t\neq 0} \pi_t} \right) (\mu_t(x) - \mu_0(x))
\end{aligned} \tag{7}
$$

Aggregations or projections of the three estimands are thus also identified. In particular,

[3] In this case $\tau_t(x) = \tau(x)$ and consequently $\Delta(x) = \tau(x) \underbrace{\sum_{t\neq 0} \left(\frac{e_t(x)}{\sum_{t\neq 0} e_t(x)} - \frac{\pi_t}{\sum_{t\neq 0} \pi_t} \right)}_{=0} = 0, \ \forall \ x \in \mathcal{X}$.

let Z_i denote a (low dimensional) subset of confounders supported on $\mathcal{Z} \subset \mathcal{X}$ and define

$$\begin{aligned} nATE(z) &= E[nATE(X_i)|Z_i = z] \\ rATE(z) &= E[rATE(X_i)|Z_i = z] \\ \Delta(z) &= E[\Delta(X_i)|Z_i = z]. \end{aligned} \tag{8}$$

These parameters provide more concise, predictive summaries of heterogeneity or allocation differences for specific subgroups defined by $Z_i = z$. The unconditional decomposition terms $nATE = E[nATE(X_i)]$, $rATE = E[rATE(X_i)]$, and $\Delta = E[\Delta(X_i)]$ are special cases thereof. We propose an estimation and inference method for these parameters in Section 4.

Remark 1: In principle, an analogous decomposition could also be constructed with alternative weights, e.g. $1/J$. However, using the unconditional effective treatment probabilities ensures that $nATE(x) = rATE(x)$ in the case of completely randomized effective treatments.

Remark 2: The interpretation of $\Delta(x)$ depends on the scenario:

- *Scenario 1:* $\Delta(x)$ and its aggregates have a descriptive interpretation. It describes how much of $nATE(x)$ is driven by an underlying effective treatment mix that deviates from the population mix. Thus, it helps to understand the heterogeneous effects resulting from the binarized treatment. A non-constant $\Delta(x)$ indicates that the choice of the binarization has consequences for the detected heterogeneous effects.

- *Scenario 2:* $\Delta(x)$ and its aggregates provide interesting information for policy makers. Positive values indicate that the assignment of treatment versions is better than random. Negative values indicate worse than random version assignment assuming that individuals act equivalently under the hypothetical random assignment compared to the observational assignment (Heckman, 2020). A non-constant $\Delta(x)$ indicates that the selection quality of versions varies across different groups. Thus, the estimand provides an evaluation of the actual assignment mechanism.

Remark 3: The comparison of $nATE(x)$ and $rATE(x)$ shows some resemblance to the

relationship between the ATE and the average treatment effect on the treated (ATET) in the canonical setting with a homogeneous binary treatment. The $ATET$ gives the average effect of those treated under the actual treatment assignment, while the ATE gives the average effect under the hypothetical random assignment of treatment.

Remark 4: The unconditional $rATE$ is a special case of a composite treatment effect described in Lechner (2002). The unconditional Δ is similar in spirit to the Population Average Prescriptive Effect defined by Imai and Li (2021) in the context of policy learning.

4 Estimation and Inference

In this section we outline a flexible estimation approach for the (conditional) decomposition terms and propose a method for conducting valid statistical inference. The method accommodates the use of modern machine learning and other non- or semiparametric methods in the estimation of the required nuisance parameters.

We propose to approximate the conditional expectations of the decomposition terms $g(z)$ by a linear combination of transformations $b(z)$ of heterogeneity variables z, i.e.

$$g(z) = b(z)'\beta_0 + r_g(z) \tag{9}$$

where β_0 is the parameter vector of the best linear predictor given as the solution to the equation $E[b(Z_i)(g(Z_i) - b(Z_i)'\beta_0)] = 0$. $r_g(z)$ is the approximation error and $b(z)$ can be basis transformations of the regressors of interest such as polynomials, splines, wavelets, or other functions. The number of components in $b(\cdot)$ is allowed to grow with the sample size which allows us to be agnostic about the shape of the true g-function.[4]

Let in the following $\eta = \eta(x) = (\mu_0(x), \ldots, \mu_J(x), e_0(x), \ldots, e_J(x))'$ denote the vector of nuisance quantities and write $\eta = \eta_i = \eta(X_i)$ with argument and subscript suppressed out of convenience. Also define $\pi = (\pi_0, \ldots, \pi_J)$.

[4]We choose the series approach here due to its analytical tractability. Compared to kernel regression, it extends easier to increasingly smooth function classes without the need for e.g. higher order kernels as the dimensionality of the heterogeneity variables increases. We believe that, with modified assumptions, it is possible to develop a corresponding methodology with localization using kernel or other semiparametric/machine learning estimators along the lines of Fan et al. (2020), Zimmert and Lechner (2019), or Kennedy (2020) as well.

Table 1: Score Functions of the Decomposition Parameters

Parameter	Score function $\psi_i(\eta, \pi) = \psi_i^{[Parameter]}(\eta, \pi)$
$nATE$	$\Psi_i(\eta) - \psi_i^{[0]}(\eta)$
$rATE$	$\frac{\sum_{t\neq 0} \pi_t \psi_i^{[t]}(\eta)}{\sum_{t\neq 0} \pi_t} - \psi_i^{[0]}(\eta)$
Δ	$\Psi_i(\eta) - \frac{\sum_{t\neq 0} \pi_t \psi_i^{[t]}(\eta)}{\sum_{t\neq 0} \pi_t}$

The scores $\psi_i^{[t]}(\eta)$ and $\Psi_i(\eta)$ are defined in equations (10) and (11), respectively.

We follow the general idea of Semenova and Chernozhukov (2021) to construct robust or "Neyman-orthogonal" scores $\psi_i(\eta, \pi)$ such that $g(z) = E[\psi_i(\eta, \pi)|Z_i = z]$. These scores are defined by having an (approximate) zero Gateaux derivative with respect to the underlying nuisance parameters at the true parameter vector (Chernozhukov et al., 2018). The robust scores for the three decomposition parameters considered here are weighted combinations of the well-known doubly robust scores for average potential outcomes (Robins & Rotnitzky, 1995), also known as Augmented Inverse Probability Weighting scores:

$$\psi_i^{[t]}(\eta) = \mu_t(X_i) + \frac{D_{t,i}(Y_i - \mu_t(X_i))}{e_t(X_i)} \tag{10}$$

$$\begin{aligned}\Psi_i(\eta) &= E[Y_i|D_i = 1, X_i] + \frac{D_i(Y_i - E[Y_i|D_i = 1, X_i])}{P(D_i = 1|X_i)} \\ &= \frac{\sum_{t\neq 0} \mu_t(X_i)e_t(X_i)}{\sum_{t\neq 0} e_t(X_i)} + \frac{D_i\left[Y_i - \frac{\sum_{t\neq 0} \mu_t(X_i)e_t(X_i)}{\sum_{t\neq 0} e_t(X_i)}\right]}{\sum_{t\neq 0} e_t(X_i)}\end{aligned} \tag{11}$$

where $\psi_i^{[t]}(\eta)$ is the score of the treatment t specific average potential outcome and $\Psi_i(\eta)$ is the score for the group described by the binary indicator. Table 1 shows how we combine these scores to form unbiased signals of our decomposition parameters. These combinations retain Neyman-orthogonality with respect to η, see Appendix E.1, but inference has to be adjusted for uncertainty in the estimation of π, see Section 5.

Consider now the regression of the score functions onto the space spanned by the

k-dimensional transformation of Z_i, $b(Z_i)$. This yields the estimator

$$\hat{\beta} = \left(\sum_{i=1}^{n} b(Z_i) b(Z_i)' \right)^{-1} \sum_{i=1}^{n} b(Z_i) \psi_i(\hat{\eta}, \hat{\pi}) \tag{12}$$

where the score of a decomposition term with estimated nuisance quantities $\psi_i(\hat{\eta}, \hat{\pi})$ serves as pseudo-outcome in the corresponding least squares regression on $b(Z_i)$. For $\hat{\pi}$ we use simple sample averages, i.e. $\hat{\pi}_t = n^{-1} \sum_{i=1}^{n} D_{t,i}$. Estimation of $\hat{\eta}$ can be done via modern machine learning such as random forests, deep neural networks, high-dimensional sparse likelihood and regression models or other non- and semiparametric estimation methods with good approximation qualities for the functions at hand. For details regarding the technical assumptions, consider Section 5. We require that all components in $\hat{\eta}$ are obtained via K-fold cross-fitting:

Definition 4.1 ***K-fold cross-fitting*** *(see Definition 3.1 in Chernozhukov et al. (2018))* *Take a K-fold random partition $(I_f)_{f=1}^{K}$ of observation indices $[K] = \{1, \ldots, n\}$ with each fold size $n_f = n/K$. For each $f \in [K] = \{1, \ldots, K\}$, define $I_f^c := \{1, \ldots, n\} \backslash I_f$. Then for each $f \in [K]$, the machine learning estimator of the nuisance function are given by*

$$\hat{\eta}_f = \hat{\eta}((Y_i, X_i, T_i)_{i \in I_f^c}).$$

Thus for any observation $i \in I_f$ the estimated score only uses the model for η learned from the complementary folds $\psi_i(\hat{\eta}, \hat{\pi}) = \psi_i(\hat{\eta}_f, \hat{\pi})$.

The use of cross-fitting allows to control the potential bias arising from overfitting using flexible machine learning methods without the need to evaluate entropy conditions for the function class that contains true and estimated nuisance quantities. If finite parametric models such as linear or logistic are assumed and estimated for the nuisance quantities, the proposed methodology can be applied without the need for cross-fitting.

Under suitable assumptions, the predictions using the estimator in (12) are consistent for $g(z)$. Moreover, it is possible to conduct asymptotically valid inference around the best linear predictor, i.e. for any $z_0 = z_{0,n}$ we can construct $(1 - \alpha)\%$ confidence intervals for

the true decomposition function as

$$CI_{1-\alpha}(g(z_0)) = \left[b(z_0)'\hat{\beta} \pm q_{1-\alpha/2} \sqrt{\frac{b(z_0)'\hat{\Omega} b(z_0)}{n}} \right] \tag{13}$$

where $q_{1-\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the standard normal distribution and $\hat{\Omega}$ is a consistent sample estimator of the asymptotic variance Ω_0 (see Appendices C and D.3.2). The estimator explicitly takes into account the additional uncertainty from estimating the unconditional treatment probabilities in the decomposition terms. The interval in (13) is also valid for the best linear predictor $b(z_0)'\beta_0$ under moderate misspecification if the variance of the approximation error is not too large. It provides asymptotically accurate confidence intervals around the true g-function if the approximation error vanishes at a suitable rate as the number of basis functions or transformations increases. For the technical details consider Section 5.

5 Large Sample Properties

In this section we provide the assumptions for the asymptotic validity of the confidence intervals proposed in (13) and some more technical discussion. Recall that $E[\psi_i(\eta, \pi)|Z_i = z] = g(z)$ where $g \in \mathcal{G}$. $g(z) = b(z)'\beta_0 + r_g(z)$ where β_0 is the parameter of the best linear predictor defined as the root of equation $E[b(Z_i)(g(Z_i) - b(Z_i)'\beta_0)] = 0$. Let $m > 2$ and define $\varepsilon_i = \psi_i(\eta, \pi) - E[\psi_i(\eta, \pi)|Z_i]$. We assume the following conditions:

A.1) *(Identification)* $Q = E[b(Z_i)b(Z_i)']$ *has eigenvalues bounded above and away from zero uniformly over* n.

A.2) *(Basis)* $\xi_k = \sup_{z \in \mathcal{Z}} ||b(z)||$ *grows sufficiently slow, i.e.* $\xi_k = O((n/\log(k))^{\frac{m-1}{2m}})$.

A.3) *(Approximation error) For each* n *and* k, *and approximation error* r_g *with* $g \in \mathcal{G}$, *there exists finite constants* c_k *such that* $||r_g||_{P,2} = O(c_k)$ *and* $||r_g||_{P,\infty} = O(c_k + c_k\xi_k)$ *with*

$$\sup_{g \in \mathcal{G}} \inf_{\beta} ||g - b'\beta||_{P,\infty} \leq c_k = O\left((\log(k)/n)^{\frac{m-1}{2m}} \right)$$

A.4) (Disturbances) There exist finite constants $\underline{\sigma}^2, \bar{\sigma}^m > 0$ *such that* $\underline{\sigma}^2 < \inf_{z \in \mathcal{Z}} E[\varepsilon_i^2 | Z_i = z]$ *and* $\sup_{z \in \mathcal{Z}} E[|\varepsilon_i|^m | Z_i = z] < \bar{\sigma}^m$.

A.5) (Machine learning bias) Let $u_n = o(1)$ *such that with probability of at least* $1 - u_n$, *for all* $f \in [K]$, $\hat{\eta}_f$ *obtained via cross-fitting belongs to a shrinking neighborhood* $\mathcal{H}_n$ *around* η. *Assume that uniformly over* $\mathcal{H}_n$ *we have that*

$$
\begin{aligned}
B_n &= \sqrt{n} \sup_{\eta \in \mathcal{H}_n} ||E[b(Z_i)(\psi(\eta, \pi) - \psi(\eta, \pi))]|| &= o(1) \\
\Lambda_n &= \sup_{\eta \in \mathcal{H}_n} E[||b(Z_i)(\psi(\eta, \pi) - \psi(\eta, \pi))||^2]^{1/2} &= o(1) \\
\kappa_n^1 &= \sup_{\eta \in \mathcal{H}_n} E[\max_{1 \leq i \leq n} |\psi(\eta, \pi) - \psi(\eta, \pi)|] &= o(n^{-\frac{1}{m}}) \\
\kappa_n &= \sup_{\eta \in \mathcal{H}_n} E[\max_{1 \leq i \leq n} (\psi(\eta, \pi) - \psi(\eta, \pi))^2]^{1/2} &= o(1)
\end{aligned}
$$

A.6) (Bounded effects) The treatment specific conditional average treatment effects are uniformly bounded, i.e. $\sup_{z \in \mathcal{Z}, t \in \mathcal{T}} |\tau_t(z)| = O(1)$.

A.7) (Lipschitz constant). Let $\alpha(z) = b(z)/||b(z)||$. *Assume that* $\xi_k^L = \sup_{z, z' \in \mathcal{Z}, z \neq z'} ||\alpha(z) - \alpha(z')||/||z - z'|| = O(\log(k))$.

Assumption A.1 limits possible collinearity of the technical regressors and A.2 controls their growth. A.3 bounds the size and variation of the approximation error universally for any approximating class chosen for the technical transformations. A.4 imposes some regularity on the tails of the conditional mean errors of the decomposition model.[5] A.5 assumes that the cross-fitted machine learning estimates have good approximation qualities around the true parameter with high probability. A.6 imposes bounds for the effective treatment specific effects conditional on the heterogeneity variables. A.7 imposes some smoothness on the technical transformations. Assumptions A.1 – A.4 combined with A.7 assure consistency and asymptotic normality with estimated variance matrix using the true score functions as left hand side variable in the series estimation.[6] These assumptions are

[5] Note that Assumption A.4 implicitly imposes some regularity on the doubly robust scores, i.e. it puts restrictions on the moments of the potential outcomes and the distribution of the generalized propensity scores. In particular, it implies regularly identified decomposition parameters (Khan & Tamer, 2010; Heiler & Kazak, 2021).

[6] For the sake of brevity we just provide the asymptotic normality result here and the corresponding ingredients that differ from Belloni et al. (2015) and Semenova and Chernozhukov (2021) in Appendix D.

stronger than required just for consistency or asymptotic normality due to the additional estimation of Ω_0. Instead of providing sharp minimal conditions for each step, we think it is more useful to combine them with respect to the applied goal of choosing the technical transformations and constructing feasible confidence intervals. With true score functions, the assumptions are implied by the conditions for equivalent results in Belloni et al. (2015). A.5 accommodates the use of estimated nuisance quantities and is equivalent to the conditions in Semenova and Chernozhukov (2021) required for similar results for the conditional treatment effects.[7] The stronger the tail condition in A.4, the weaker the rate requirements in A.2, A.3, and A.5 become. A.6 is a mild heterogeneity restriction. We obtain the following result:

Theorem 5.1 *Let $\Phi(\cdot)$ denote the Gaussian cumulative distribution function. Suppose Assumptions A.1 - A.7 hold and $\hat{\beta}$ and $\hat{\Omega}$ are estimated according to (12) and (21) respectively. Then, for any $z_0 = z_{0,n}$,*

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} \left| P\left(\sqrt{n} \frac{b(z_0)'(\hat{\beta} - \beta_0)}{\sqrt{b(z_0)'\hat{\Omega} b(z_0)}} \leq t \right) - \Phi(t) \right| = 0.$$

Moreover if the approximation error is small, i.e. $\sqrt{n} r_g(z_0) / \sqrt{b(z_0)'\Omega_0 b(z_0)} \to 0$, then

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} \left| P\left(\sqrt{n} \frac{b(z_0)'\hat{\beta} - g(z_0)}{\sqrt{b(z_0)'\hat{\Omega} b(z_0)}} \leq t \right) - \Phi(t) \right| = 0.$$

Theorem 5.1 demonstrates the asymptotic validity of the confidence intervals proposed in (13). The result accommodates the case of mild misspecification often present in applied econometric research. It is most useful under the additional slight undersmoothing condition that makes any misspecification bias vanish sufficiently fast.[8]

Note that Theorem 5.1 also applies to alternative combinations of Neyman-orthogonal scores. In particular it extends to any weighted combination of conditional average

For the remaining approximations please consider the aforementioned papers.

[7] They first only assume $\kappa_n^1 = o(1)$ for the asymptotic normality with true variance which is a weaker condition. However, for consistent variance estimation, they also require that $n^{1/m}\kappa_n^1 = o(1)$, see their additional assumption (ii) in Theorem 3.3. Thus there is no qualitative difference to our assumption.

[8] In particular, when $\mathcal{G}$ is in a s-dimensional ball on $\mathcal{Z}$ of finite diameter, then the condition simplifies to $n^{1/2} k^{-(\frac{1}{2}+\frac{s}{d})} \log(k) \to 0$. See also Belloni et al. (2015), Comment 4.3 for additional details.

treatment effects as long as the weights are either fixed or can be estimated at the parametric rate similarly to $\hat{\pi}_t - \pi_t = O_p(n^{-1/2})$. This can be useful when comparing the heterogeneity in a given selection mechanism to alternative, hypothetical (estimated or true) policies different from random selection as considered in this paper.

6 Decomposition versus Multi-valued Treatment Analysis under Many Treatments and Weak Overlap

In a classic multi-valued treatment effect setup, we can test for effect heterogeneity by comparing different (conditional) potential outcomes pairwise across treatments. For that, the estimation of many effects is required. On one hand, this requires correction for multiple testing. On the other, estimation and inference becomes increasingly difficult with many treatments as already in the simple unconditional effect analysis one has to test over $J(J-1)/2$ hypotheses. The decomposition, however, always consists of only three (functional) parameters independently of the underlying dimensions of the effective multi-valued treatment. Thus, estimation is also feasible even if there are many treatments in large samples, i.e. if $J \to \infty$. With many treatments, the strong overlap assumption usually required for regular identification of standard multi-valued effects will be violated by construction and estimation and conventional statistical inference will perform poorly (Heiler & Kazak, 2021). Strong overlap means, that all generalized propensity scores are uniformly bounded away from zero by a positive constant. In finite samples, a very small bound for the propensities is hard to distinguish from a zero lower bound, see Rothe (2017). When estimating generalized propensity scores for many treatments, practitioners routinely encounter many extreme generalized propensities close to zero (e.g. Uysal, 2015). This problem is even exacerbated when focusing on conditional effect heterogeneity reducing effective sample sizes. Much of this can be avoided by resorting to the decomposition parameters instead. Relaxing the overlap requirements comes at the cost of testing weaker conditions regarding effect heterogeneity. In the multi-valued treatment effect setting we can test *sufficient* conditions for effect heterogeneity between and within all treatments, while our decomposition tests *necessary* conditions for within treatment effect heterogeneity

and between treatment heterogeneity.

For illustration consider the case of the following sequence of probability measures $\{F_n\}_{n=1}^{\infty}$ such that $\mathcal{T} \equiv \mathcal{T}_n = \{0, 1, \ldots, J_n\}$ with $J_n \to \infty$. We still assume strong overlap for control propensities, i.e $\inf_{x \in \mathcal{X}} e_0(x) > \underline{e} > 0$ almost surely for some $\underline{e} > 0$. Now additionally assume that almost surely for $t \neq 0$

$$e_t(X_i) \equiv e_{t,n}(X_i) = \frac{\underline{e}_{t,n}(X_i)}{J_n} \tag{14}$$

such that

$$\inf_{n \geq 1} \inf_{x \in \mathcal{X}} \underline{e}_{t,n}(x) > \underline{e} > 0 \tag{15}$$

$$\sup_{n \geq 1} \sup_{x \in \mathcal{X}} \underline{e}_{t,n}(x) < C \tag{16}$$

for $t \neq 0$. (15) is a uniform strong overlap assumption for the *rescaled* generalized propensities $\underline{e}_{t,n}(x) = J_n e_{t,n}(x)$. This allows the generalized propensity scores to converge to zero, i.e. for many treatments with small selection probabilities, but controls their rate. This assumption says that the increase in the number of treatments will qualitatively be similar over the support of confounders, i.e. units have a somewhat comparable decrease in their relative propensities along the sequences. Intuitively, the distributions of the generalized propensity scores should not concentrate too quickly around zero when multiplying with the total number of treatments along such sequences. (16) is only for normalization. Now consider the variance of the leading terms for estimating the unconditional $rATE$ (or equivalently any other decomposition parameter). Along the sequences we have that

$$V_{F_n}[\sqrt{n}(\hat{\theta}_{rATE} - \theta_{rATE})] = O(1). \tag{17}$$

Thus, the limiting variance is bounded independently of the number of treatments J_n. For estimating a multivalued average treatment effect $\theta_{t,t'}$ comparing levels $t, t' \in \mathcal{T}$, however,

we obtain that

$$V_{F_n}[\sqrt{n}(\hat{\theta}_{t,t'} - \theta_{t,t'})] = O(J_n). \tag{18}$$

For the derivations consider Appendix F. Thus, estimating and testing heterogeneity will become increasingly difficult under many treatments $J_n \to \infty$ when constructing multi-valued effect parameters. The decomposition parameters circumvent this problem as for all $nATE$, $rATE$, and Δ, the more problematic contributions of treatments with little probability mass are also weighted down correspondingly. Note that, without knowledge of the nuisance parameters, the approximation rates for the semiparametric double machine learning framework in Section 5 are also likely to require stronger sparsity/complexity type assumptions or cross-treatment restrictions in the many treatment setup. We leave an extension along this line for future work.

7 Monte Carlo Study

In this section we analyze the finite sample performance of the analytical confidence bounds proposed in Section 4. In particular we analyze the empirical coverage rates of the corresponding confidence intervals in a setup with heterogeneous effective treatment probabilities for all the decomposition parameters. We consider the case of three effective treatment levels and a univariate linear model for the heterogeneity analysis using different sample sizes and total number of confounding variables. In particular, in the final step, we regress the estimated pseudo outcomes on a single confounder and evaluate the coverage rates for the parameters of this linear predictor. We consider two ways to estimate the nuisance parameters (i) correctly specified parametric models and (ii) double machine learning estimators. For the latter we apply 2-fold cross-fitting using ℓ_1-regularized linear regression for the outcome models as well as ℓ_1-regularized multinomial logistic regression for the generalized propensity scores. Tuning parameter selection is done via 5-fold cross-validation. The true models satisfy the necessary sparsity assumptions required for high-quality approximation of the machine learning methods (Belloni & Chernozhukov, 2013; Farrell, 2015; Belloni, Chernozhukov, & Wei, 2016). For more details on the designs

please consider Appendix G.

Table 2 contains the coverage rates of the confidence intervals based on (13) using double machine learning at a significance level of 5%. For $rATE(x)$ and $nATE(x)$ all results are very close to the nominal coverage rate. For $\Delta(x)$, there is some undercoverage for the intercept α by 1.7 to 9.6 percentage points which increases in the number of parameters and decreases with the sample size. The slope parameter β is accurate for any sample or regressor set size.

Table 2: Monte Carlo Simulation: Results (Double Machine Learning)

(a) n = 1000

		$rATE$	$nATE$	Δ
k = 10	α	0.9472	0.9494	0.8964
	β	0.9456	0.9452	0.9430
k = 100	α	0.9438	0.9420	0.8538
	β	0.9486	0.9476	0.9542

(b) n = 5000

		$rATE$	$nATE$	Δ
k = 10	α	0.9516	0.9500	0.9326
	β	0.9498	0.9506	0.9478
k = 100	α	0.9512	0.9508	0.9204
	β	0.9534	0.9520	0.9534

The table entries contain the coverage rates under the null hypothesis for the parameters (α, β) of the linear predictor for different number of regressors (k), sample sizes (n) and decomposition parameters $rATE$, $nATE$ and Δ. The nominal coverage rate is 95%. Results are based on 5000 simulations.

Table 3 contains the coverage rates of the confidence intervals based on (13) using correctly specified parametric models at a significance level of 5%. All results are very close to their nominal coverage rate. For the smallest $n = 1000$ and $k = 100$, there is undercoverage for β of 5 percentage points for the $rATE$ which largely vanishes for $n = 5000$. For the other parameters, there are no relevant size distortions.[9] Overall the inference based on the asymptotic approximation in (13) seems to be reliable in finite samples.

8 Applications

8.1 Smoking and Birth Weight (Scenario 1)

The detrimental effect of smoking on birth weight and its economic costs are well documented (see e.g. Almond et al., 2005; Abrevaya, 2006; Almond & Currie, 2011, and

[9]Note that the results are insensitive with respect to the design. We have also experimented with other heterogeneous effects and treatment probabilities which do not seem to affect the coverage rates substantially.

Table 3: Monte Carlo Simulation: Results (Parametric Model)

(a) n = 1000

		$rATE$	$nATE$	Δ
k = 10	α	0.9460	0.9456	0.9524
	β	0.9444	0.9488	0.9470
k = 100	α	0.9468	0.9488	0.9518
	β	0.8974	0.9434	0.9474

(b) n = 5000

		$rATE$	$nATE$	Δ
k = 10	α	0.9486	0.9480	0.9482
	β	0.9448	0.9528	0.9570
k = 100	α	0.9506	0.9508	0.9508
	β	0.9380	0.9464	0.9458

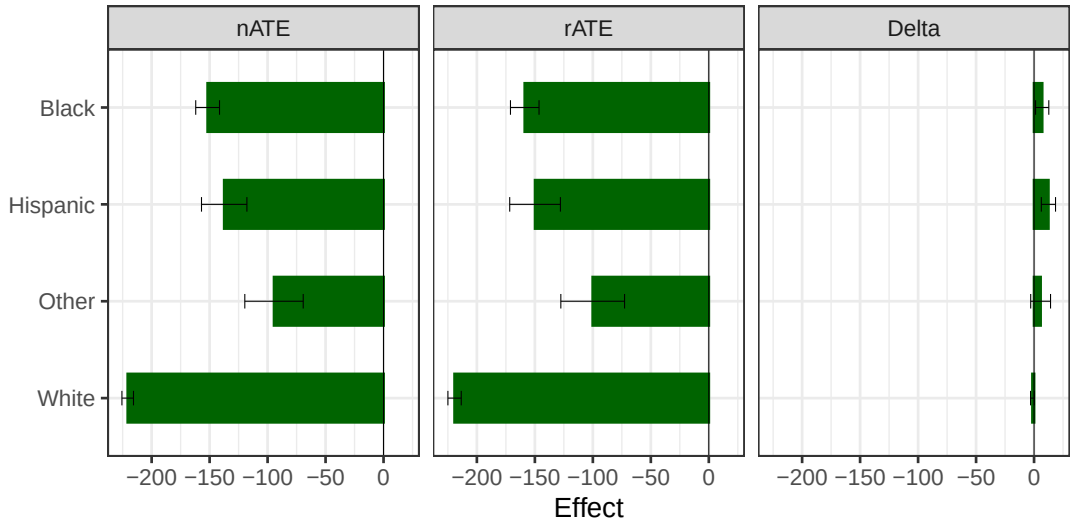
The table entries contain the coverage rates under the null hypothesis for the parameters (α, β) of the linear predictor for different number of regressors (k), sample sizes (n) and decomposition parameters $rATE$, $nATE$ and Δ. The nominal coverage rate is 95%. All results are based on 5000 simulations.

references therein). Beyond the standard average effects it is important to understand the heterogeneous effects to e.g. identify for which subgroups interventions to reduce smoking during pregnancy would be most beneficial. Abrevaya (2006) documents that the negative effect of smoking is less pronounced for black compared to white mothers in a standard subgroup analysis. A variety of papers analyze heterogeneous effects of smoking as a function of mother's age (Abrevaya, Hsu, & Lieli, 2015; Lee et al., 2017; Zimmert & Lechner, 2019; Fan et al., 2020). They all document increasingly negative effects with higher age. The aforementioned studies consider "smoking yes/no" as the binary treatment. Cattaneo (2010) notes that smoking is not a homogeneous treatment, but that the negative effects become more extreme for higher intensities of smoking. Thus, the binary indicator "smoking" represents only an aggregation of smoking intensities which directly affect birth weight. This corresponds to Scenario 1 of Figure 1. We investigate whether the heterogeneous effects documented in the literature can be at least partly explained by different smoking intensities of different groups.

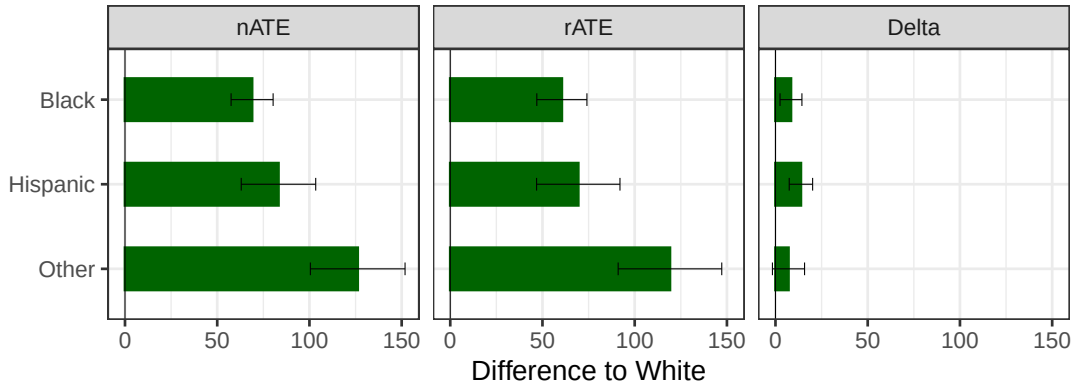
We analyze the dataset of Almond et al. (2005) used by Cattaneo (2010) with five intensities of smoked cigarettes per day as the effective treatment $T_i \in \mathcal{T} = \{0, 1-5, 6-10, 11-15, 16-20, >20\}$, the binary indicator defined as $D_i = \mathbb{1}[T_i > 0]$, the outcome Y_i being birth weight in gram, and the confounders X_i including age, education, ethnicity, and marital status of mother and father as well as health indicators and pregnancy history of the mother.[10] The dataset comprises 511,940 observations after removing the 0.1% of the observations with missing values in relevant variables and 52 confounders. The

[10]We thank Matias Cattaneo for sharing the full data. A random subsample is available on his GitHub repository.

Figure 3: Heterogeneous effects and decomposition by ethnicity



(a) Subgroup effects for ethnicity



(b) Effect heterogeneity with white as benchmark

Note: Point estimates of the decomposition parameters with 95%-confidence interval.

nuisance parameters are estimated with 2-fold cross-fitting using an ensemble learner of the unconditional mean, Random Forests, Lasso and Ridge regression with 2-fold cross-validated weights. For the propensity scores, we use logistic Lasso and Ridge.

Smoking behavior differs along the heterogeneity variables ethnicity and age showing that white and older smoking mothers smoke more heavily.[11] Combined with the result of Cattaneo (2010) that different smoking intensities have different effects, this suggests that at least part of the heterogeneity could be explained by different smoking intensities.

Figure 3 shows the result of the decomposition for the heterogeneity variable "ethnicity". The upper panel shows the decomposition for each subgroup. It is obtained by running an

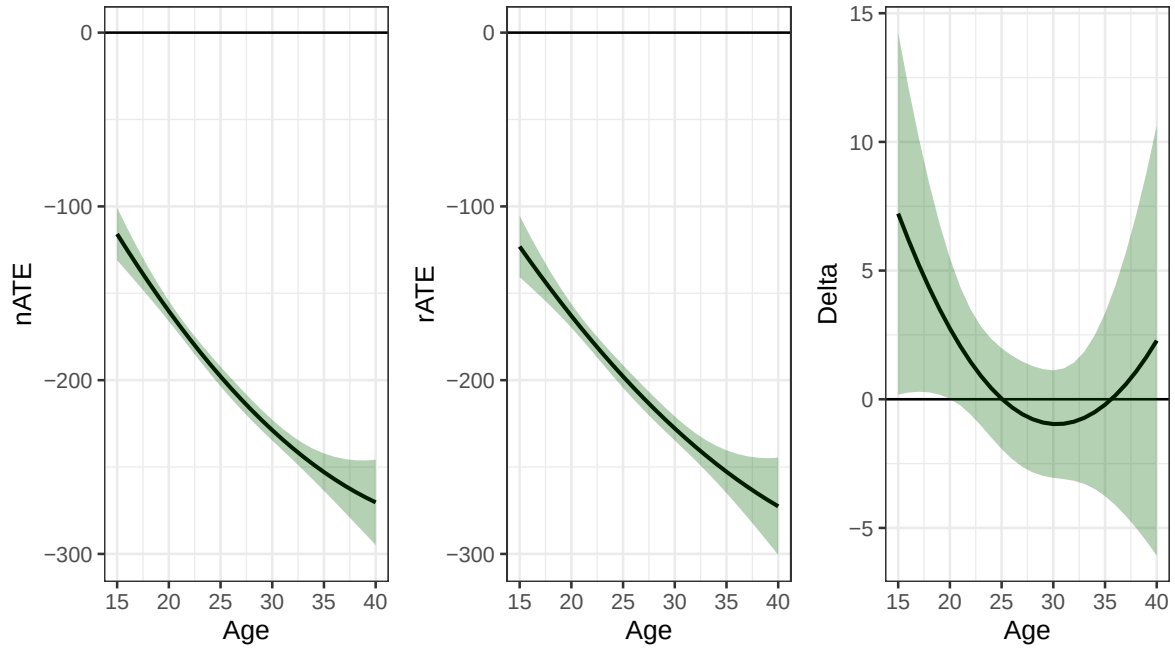
[11] Appendix H and in particular Figure H.1 provides the smoking distributions by heterogeneity variables.

OLS regression of the estimand specific pseudo-outcome on a set of four dummy variables indicating ethnicity of the mother without a constant. The standard errors are then adjusted as described in Section 4. The $nATE$ in the left part can be considered as the result of a standard subgroup analysis. Like previous studies we find that smoking reduces the birth weight of newborns more for white women than for Blacks, Hispanics and others. Given that smoking is a binarized treatment, it is not clear how much is really effect heterogeneity and how much is driven by the fact that subgroups differ in their smoking intensity. The decomposition term $rATE$ fixes the intensity of smoking for all subgroups at the population level. It provides the subgroup specific effect of smoking if all groups had the same smoking intensity. Under this harmonized smoking intensity the negative average effect of smoking is smaller for white women and larger for the others. Δ in the right graph quantifies the difference between $nATE$ and $rATE$. It shows relatively small differences suggesting that different smoking intensities are not the main driver of the differences between white mothers and the other groups. However, they are also not negligible as the lower panel of Figure 3 shows. It quantifies the heterogeneous effects by subtracting the effects for white mothers from the other three groups. We observe that a significant portion of the difference between black/hispanic mothers and white mothers is driven by different smoking intensities. For black vs. white mothers the difference in the $nATE$ is 69 gram of which 12% are due to different smoking intensities ($\Delta = 8$). For hispanic vs. white mothers it explains around 17% ($\Delta = 14$).

Figure 4 depicts the heterogeneity analysis along age. We use B-splines as basis functions of age. We select the nodes and order via leave-one-out cross-validation for each parameter and apply the most flexible/low-bias model for all parameters to ensure that the $rATE$ and Δ curves add up to the $nATE$ curve. The left panel of Figure 4 replicates the well-established findings of previous papers that the $nATE$ is much smaller for younger mothers than for older mothers.

In the extreme case where different smoking intensities would fully explain the heterogeneous $nATE$, we would see a flat $rATE$ curve in the middle graph. However, we only observe that the effect of teenage mothers would be more negative if we harmonize smoking intensity over all age groups.

Figure 4: Effect Heterogeneity by Age



Notes: B-spline estimated decomposition parameters with 95%-confidence interval.

Overall, only a relatively small part of the heterogeneous effects of the binarized smoking indicator can be attributed to different smoking intensities and the larger part seems to be driven by different age groups actually being affected differently. This is expected as younger mothers had less time to inflict severe long-run damage to their own physiology also prior to pregnancy and are on average more healthy.

8.2 Job Corps (Scenario 2)

We illustrate Scenario 2 of Figure 2 with an evaluation of the Job Corps (JC) program. JC operates since 1964 and is the largest training program for disadvantaged youth aged 16-24 in the US (see Schochet et al., 2001, 2008, for a detailed description). The roughly 50,000 participants per year receive an intensive treatment as a combination of different components like academic education, vocational training, and job placement assistance. Participants plan their educational and vocational curricula together with counselors. This means that although the variable "access to JC" is a binary indicator, different versions of JC participation are conceivable. Heterogeneous effects might thus be driven by different effectiveness of JC for different groups, by different tailoring of the curriculum, or a

combination thereof.

We investigate this based on data from an experiment in 1994-1996 (Schochet, Burghardt, & McConnell, 2019).[12] This experiment is basis of a variety of studies looking at different aspects of JC. Many of them report gender differences in the effectiveness of the programs with women benefiting less than men from access to JC (e.g. Schochet et al., 2001, 2008; Flores et al., 2012; Eren & Ozbelik, 2014; Strittmatter, 2019). One potential explanation for this finding is that men and women focus on average on different vocational training within JC. In particular men receive more often training for higher paying craft jobs, while women focus more often on training for the service sector (Quadagno & Fobes, 1995; Inanc, Needels, & Berk, 2017).[13] We apply our decomposition method to investigate this potential explanation of the gender gap in program effectiveness.

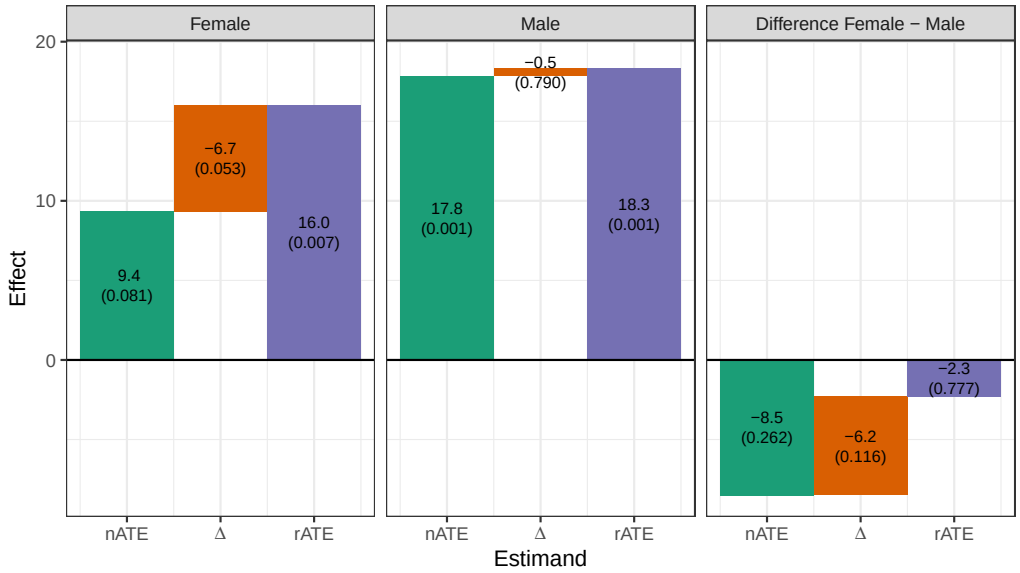
We analyze the intention to treat effect (ITT) of the binary variable indicating random access to JC (D_i) on weekly earnings four years after random assignment (Y_i). We consider the following eleven versions of treatment (T_i): (i) *No JC* if eligible individuals did not participate (non-compliers), (ii) *JC without vocational training* if eligible individuals entered JC but did not receive vocational training, (iii-ix) training for jobs in the clerical, health, auto mechanics, welding, electrical/electronics, construction, or food sector, (x) other vocational training, (xi) training for multiple sectors.

We estimate the nuisance parameters with the same ensemble as in the previous section with 5-fold cross-fitting. We control for 55 covariates that include pre-treatment information about labor market history, socio-economic characteristics, education, health, crime, and JC related variables. These control variables overlap mostly with those of Flores et al. (2012) who also employ an unconfoundedness strategy. Considering second-order interactions and polynomials results in a total of 1428 variables after screening nearly empty cells (less than 1% observations) and nearly perfectly correlated variables (correlation higher than 0.99). In total we work with a sample of 9,708 observations.

The unconditional $nATE$, corresponding to the ITT of eligibility for JC on monthly earnings, is estimated at \$14.2 (S.E. 3.8), which is an increase of 7% in line with previous studies. The unconditional $rATE$ is larger (\$17.4, S.E. 4.1) indicating that randomly

[12]The data is available as public use file via https://doi.org/10.3886/E113269V1.
[13]Appendix I and in particular Figure I.1 provides the distribution of trainings by gender.

Figure 5: Effect heterogeneity and decomposition by gender



Notes: The numbers in the bar show the point estimate and the p-value in parentheses.

allocating the curricula would have been better than the actual assignment. However, the unconditional difference Δ is insignificant ($\$ - 3.1$, S.E. 1.8). This suggests that, on average, the selection of versions is not statistically distinguishable from random allocation.

Figure 5 depicts the decomposition of the gender specific effects. We observe that the effect for women with the actual composition of vocational training ($nATE$) is not significant at $\alpha = 0.05$, but under the hypothetical treatment composition of the population would show a clear positive effect. The gender gap in effectiveness basically disappears when both groups receive the same hypothetical mix of vocational training. The right part of Figure 5 suggests that 73% of the gender gap in the effectiveness of JC is due to different training curricula. This means that the worse than average performance of the assignment mechanism seen in the unconditional parameters is mostly driven by women. While the assignment to vocational training for men is as well targeted as random assignment, for women it is even worse. This indicates that there is room for improvement to target vocational training in general and for women in particular. Our results suggest that removing the worse than random targeting of vocational training for women could decrease the gender gap in the effectiveness of access to JC.

9 Concluding Remarks

The method proposed in this paper provides a practical way of decomposing effect heterogeneity obtained from analyzing a binary treatment indicator that does not coincide with the effective multi-valued treatment. We believe that our approach can be extended to other causal parameters and identification strategies such as continuous effective treatments, selection on unobservables/instrumental variables, or mediation analysis. It would also be interesting to see whether our ideas could be further developed to find the most relevant dimensions of effective treatments for cases with multiple treatment versions instead of requiring the researcher to manually specify them.

The conceptual and empirical results highlight that potential treatment heterogeneity underlying the analyzed binary indicator should be taken more seriously and explicitly discussed in applications, especially when interpreting heterogeneous effects. The decomposition provides one principled way to do this. However, it requires to observe the effective treatment. Data collection can anticipate the goal of better understanding treatment heterogeneity by recording effective treatment information beyond a binary indicator. Furthermore, the decomposition shows that boiling down the analysis to such binary indicators, while facilitating the analysis, can come at the cost of a more intricate interpretation of empirical results.

References

Abadie, A., & Cattaneo, M. D. (2018). Econometric methods for program evaluation. *Annual Review of Economics*, *10*, 465–503.

Abrevaya, J. (2006). Estimating the effect of smoking on birth outcomes using a matched panel data approach. *Journal of Applied Econometrics*, *21*(4), 489–519.

Abrevaya, J., Hsu, Y.-C., & Lieli, R. P. (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, *33*(4), 485–505.

Almond, D., Chay, K. Y., & Lee, D. S. (2005). The costs of lower birth weight. *The Quarterly Journal of Economics*, *120*(3), 1031–1083.

Almond, D., & Currie, J. (2011). Human capital development before age five. In *Handbook of labor economics* (Vol. 4, pp. 1315–1486). Elsevier.

Andresen, M. E., & Huber, M. (2021). Instrument-based estimation with binarised treatments: issues and tests for the exclusion restriction. *The Econometrics Journal*, *utab002*, published ahead of print 29 August 2020.

Angrist, J. D., & Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, *90*(430), 431–442.

Athey, S., & Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

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

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *Annals of Statistics*, *47*(2), 1148 - 1178.

Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, *19*(2), 521–547.

Belloni, A., Chernozhukov, V., Chetverikov, D., & Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, *186*(2), 345–366.

Belloni, A., Chernozhukov, V., & Wei, Y. (2016). Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, *34*(4), 606–619.

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, *155*(2), 138–154.

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

Cole, S. R., & Frangakis, C. E. (2009). The consistency statement in causal inference. *Epidemiology*, *20*(1), 3–5.

Curth, A., & van der Schaar, M. (2021). Nonparametric estimation of heterogeneous

treatment effects: From theory to learning algorithms. In *Proceedings of the 24th international conference on artificial intelligence and statistics* (Vol. 130, pp. 1810–1818). PMLR.

Davis, J. M. V., & Heller, S. B. (2020). Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs. *The Review of Economics and Statistics*, *102*(4), 664–677.

Eren, O., & Ozbelik, S. (2014). Who benefits from Job Corps? A distributional ananlysis of an active labor market program. *Journal of Applied Econometrics*, *29*(4), 586–611.

Fan, Q., Hsu, Y.-C., Lieli, R. P., & Zhang, Y. (2020). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics*, published ahead of print 14 September 2020.

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

Flores, C. A., Flores-Lagunes, A., Gonzalez, A., & Neumann, T. C. (2012). Estimating the effects of length of exposure to instruction in a training program: The case of job corps. *Review of Economics and Statistics*, *94*(1), 153–171.

Heckman, J. J. (2020). Epilogue: Randomization and social policy evaluation revisited. In F. Bédécarrats, I. Guérin, & F. Roubaud (Eds.), *Randomized control trials in the field of development: A critical perspective* (pp. 304–330). Oxford University Press. doi: 10.1093/oso/9780198865360.003.0014

Heiler, P., & Kazak, E. (2021). Valid inference for treatment effect parameters under irregular identification and many extreme propensity scores. *Journal of Econometrics*, *222*(2), 1083–1108.

Hernán, M. A., & VanderWeele, T. J. (2011). Compound treatments and transportability of causal inference. *Epidemiology*, *22*(3), 368–377.

Hotz, V. J., Imbens, G. W., & Klerman, J. A. (2006). Evaluating the differential dffects of alternative welfare-to-work training aomponents: A reanalysis of the California GAIN program. *Journal of Labor Economics*, *24*(3), 521–566.

Hotz, V. J., Imbens, G. W., & Mortimer, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*,

125(1-2), 241–270.

Imai, K., & Li, M. L. (2021). Experimental evaluation of individualized treatment rules. *Journal of the American Statistical Association*. doi: 10.1080/01621459.2021.1923511

Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *Annals of Applied Statistics*, *7*(1), 443–470.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*(3), 706–710.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Inanc, H., Needels, K., & Berk, J. (2017). *Gender segregation in training programs and the wage gap* (Tech. Rep. Nos. Cambridge, NJ: Mathematica Policy Research).

Kennedy, E. H. (2020). Optimal doubly robust estimation of heterogeneous causal effects. *arXiv:2004.14497*. Retrieved from http://arxiv.org/abs/2004.14497

Khan, S., & Tamer, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, *78*(6), 2021–2042.

Knaus, M. C. (2020). Double machine learning based program evaluation under unconfoundedness. Retrieved from http://arxiv.org/abs/2003.03191

Knaus, M. C., Lechner, M., & Strittmatter, A. (2020). Heterogeneous employment effects of job search programmes: A machine learning approach. *Journal of Human Resources*, *0718-9615R*, published ahead of print 26 March 2020.

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.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In M. Lechner & E. Pfeiffer (Eds.), *Econometric evaluation of labour market policies* (pp. 43–58). Heidelberg: Physica.

Lechner, M. (2002). Program heterogeneity and propensity score matching: an application to the evaluation of active labor market policies. *The Review of Economics and Statictics*, *84*, 205–220.

Lee, S., Okui, R., & Whang, Y.-J. (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics*, *32*(7), 1207–1225.

Marshall, J. (2016). Coarsening bias: How coarse treatment measurement upwardly biases instrumental Variable Estimates. *Political Analysis*, *24*(2), 157–171.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, *79*(1), 147–168.

Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, *108*(2), 299–319.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688.

Petersen, M. L. (2011). Compound treatments, transportability, and the structural causal model: The power and simplicity of causal graphs. *Epidemiology*, *22*(3), 378–381.

Quadagno, J., & Fobes, C. (1995). The welfare state and the cultural reproduction of gender: Making good girls and boys in the Job Corps. *Social Problems*, *42*(2), 171–190.

Richardson, T., & Robins, J. M. (2013). Single World Intervention Graphs (SWIGs): Unifying the Counterfactual and Graphical Approaches to Causality – Presentation. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, *128*.

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, *90*(429), 122–129.

Rothe, C. (2017). Robust Confidence Intervals for Average Treatment Effects Under Limited Overlap. *Econometrica*, *85*(2), 645–660.

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

Rubin, D. B. (1980). Randomization Analysis of Experimental Data: The Fisher Random-

ization Test Comment. *Journal of the American Statistical Association*, *75*(371), 591.

Rudelson, M. (1999). Random vectors in the isotropic position. *Journal of Functional Analysis*, *164*(1), 60–72.

Schochet, P. Z., Burghardt, J., & Glazerman, S. (2001). *National job corps study: The impacts of job jorps on participants' employment and related outcomes* (Tech. Rep.). Princeton, NJ: Mathematica Policy Research Inc.

Schochet, P. Z., Burghardt, J., & McConnell, S. (2008). Does job corps work? Impact findings from the national job corps study. *American Economic Review*, *98*(5), 1864–1886.

Schochet, P. Z., Burghardt, J., & McConnell, S. (2019). Replication data for: Does job corps work? Impact findings from the national job corps study. *Inter-university Consortium for Political and Social Research (ICPSR)*.

Semenova, V., & Chernozhukov, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, *24*(2), 264–289.

Strittmatter, A. (2019). Heterogeneous earnings effects of the job corps by gender: A translated quantile approach. *Labour Economics*, *10760*.

Tian, L., Alizadeh, A. A., Gentles, A. J., & Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, *109*(508), 1517–1532.

Uysal, D. (2015). Doubly robust estimation of causal effects with multivalued treatments: An application to the returns to schooling. *Journal of Applied Econometrics*, *30*(5), 763–786.

VanderWeele, T. J. (2009). Concerning the consistency assumption in causal inference. *Epidemiology*, *20*(6), 880–883.

VanderWeele, T. J., & Hernan, M. A. (2013). Causal inference under multiple versions of treatment. *Journal of Causal Inference*, *1*(1), 1–20.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523),

1228–1242.

Zimmert, M., & Lechner, M. (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *arXiv:1908.08779*. Retrieved from http://arxiv.org/abs/1908.08779

Appendices

Not meant for publication but to be provided as supplementary material in the online repositories of the Journal and the homepages of the authors.

A Toy example

Consider a setting with a binary heterogeneity variable $X_i \in \{0, 1\}$ and three effective treatments $T_i \in \{0, 1, 2\}$. We impose deterministic potential outcomes that are homogeneous within treatment status, but heterogeneous between treatments:

	$Y_i(0)$	$Y_i(1)$	$Y_i(2)$
$X_i = 0$	0	-1	1
$X_i = 1$	0	-1	1

Both groups defined by X_i have the same potential outcomes under the different treatments. This means there can be no real effect heterogeneity. However, consider now that the probability to receive the effective treatments varies with X_i:

	$P(T_i = 0\|X_i)$	$P(T_i = 1\|X_i)$	$P(T_i = 2\|X_i)$
$X_i = 0$	0.5	$1/8$	$3/8$
$X_i = 1$	0.5	$3/8$	$1/8$

Collapsing treatments one and two into a binary treatment $D_i = \mathbb{1}[T_i > 0]$ and running a subgroup analysis for the "treatment" D_i results in the following conditional average treatment effects ($CATE$):

$$CATE(X_i) = 1 - 4 \cdot P(T_i = 1|X_i) = \begin{cases} 0.5 & \text{if } X_i = 0 \\ -0.5 & \text{if } X_i = 1. \end{cases}$$

Thus, the aggregation into the binary indicator leads us to "find" a positive effect for one group and a negative effect for another group although the effective treatments actually do not create heterogeneous effects. Everything is just driven by them receiving a different mix of effective treatments.

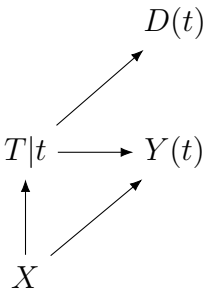
B Identification

B.1 Conditional independencies

We can read off the conditional independencies with respect to potential, not observed, outcomes encoded in DAGs (1) and (2) from single-world intervention graphs (SWIG) of Richardson and Robins (2013). We intervene on T_i to read off the independencies we require for identification of our decompositions.

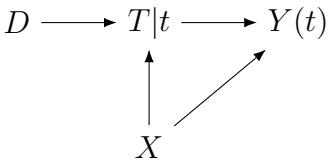
Scenario 1:

Figure B.1: SWIG with intervention on T_i



Scenario 2:

Figure B.2: SWIG with intervention on T_i

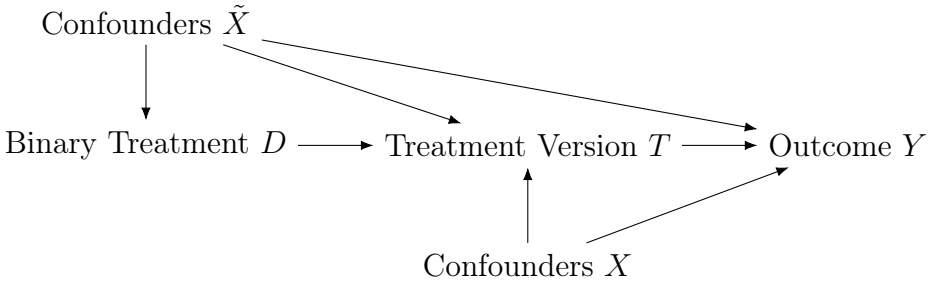


Both SWIGs (B.1) and (B.2) imply the conditional independence shown in Equation (2). This links the observed effective treatment to the unobserved potential outcomes and justifies our assumption 1a required for identifying our decomposition terms.

B.2 Identification in Scenario 2 with confounded binary treatment

Figure B.3 considers the case of a binary treatment that is potentially confounded with both treatment version selection and outcome due to the backdoor through $\tilde{X}$. This could occur e.g. in the case of the evaluation of Job Corps access on earnings when access was

Figure B.3: DAG: Confounded binary treatment precedes confounded treatment version:



not allocated randomly but is based on observables. In this case, the derived conditional independence assumptions change to

$$Y_i(0), Y_i(1) \ldots, Y_i(J) \perp\!\!\!\perp D_i \mid T_i, \tilde{X}_i \tag{19}$$

$$Y_i(0), Y_i(1) \ldots, Y_i(J) \perp\!\!\!\perp T_i \mid X_i, \tilde{X}_i \tag{20}$$

Thus, the adjustment set required for identification and entering estimation of the nuisance parameters of the decomposition terms needs to incorporate the additional set of confounders $\tilde{X}_i$, but all results hold equivalently.

C Estimation of Asymptotic Variance

Let $E_n[X_i] = \frac{1}{n}\sum_{i=1}^{n} X_i$. Define

$$\begin{aligned} \hat{Q} &= E_n[b(Z_i)b(Z_i)'] \\ \hat{\Omega} &= \hat{Q}^{-1}\hat{\Sigma}\hat{Q}^{-1} \end{aligned} \tag{21}$$

For the $rATE$ we use

$$\hat{\Sigma} = E_n\Big[(b(Z_i)e_i + \hat{a}_i - \bar{\hat{a}_i})(b(Z_i)e_i + \hat{a}_i - \bar{\hat{a}_i})'\Big]$$

with

$$
\begin{aligned}
e_i &= \psi_i^{[rATE]}(\hat{\eta}, \hat{\pi}) - b(Z_i)'\hat{\beta} \\
\hat{a}_i &= \sum_{t \neq 0} \frac{E_n[b(Z_i)(\psi_i^{[t]}(\hat{\eta}) - \psi_i^{[0]}(\hat{\eta}))](D_{t,i}(1-\hat{\pi}_0) + D_{0,i}\hat{\pi}_t)}{(1-\hat{\pi}_0)^2} \\
\bar{\hat{a}}_i &= E_n[\hat{a}_i] \\
\hat{\pi}_t &= E_n[D_{t,i}]
\end{aligned}
$$

For Δ, the $\psi_i^{[rATE]}(\hat{\eta}, \hat{\pi})$ has to be replaced by the corresponding score function and $\hat{\Sigma}$ changes to

$$
\hat{\Sigma} = E_n\left[(b(Z_i)e_i - \hat{a}_i + \bar{\hat{a}}_i)(b(Z_i)e_i - \hat{a}_i + \bar{\hat{a}}_i)'\right].
$$

For the $nATE$ we use only $\hat{\Sigma} = E_n[b(Z_i)b(Z_i)'e_i^2]$ as there are no estimated unconditional weights.

D Supplementary Material for Section 5

D.1 Notation and Outline

$$
\begin{aligned}
E_n[X_i] &:= \frac{1}{n}\sum_{i=1}^n X_i \\
G_n[X_i] &:= \frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - E[X_i]) \\
||r_g||_{P,2} &:= \sqrt{\int_{z\in\mathcal{Z}} r_g^2(z) dP(z)} \\
||r_g||_{P,\infty} &:= \sup_{z\in\mathcal{Z}} |r_g(z)| \\
\tau_t(z) &= E[\tau_t(X_i)|Z_i = z]
\end{aligned}
$$

Also let $C > 0$ denote generic constants that do not depend on n, $a \lesssim_P b$ means that $a/b = O_p(1)$. Let $b_i = b(Z_i)$ and $\psi_i^{[t,0]}(\eta) = \psi_i^{[t]}(\eta) - \psi_i^{[0]}(\eta)$ for any t and η. First we

provide some auxiliary results. Then we derive the asymptotically linear representation of the best linear predictor and show its asymptotic normality using the true variance covariance matrix. In the last step we provide the necessary derivations to replace the true variance with the estimated sample counterpart proposed in Section 4. For reference, we refer with BCCK to Belloni et al. (2015) and with SC to Semenova and Chernozhukov (2021).

D.2 Auxiliary results

H.1: $||\hat{\gamma}_t - \gamma_t||$ **bound**

The triangle inequality and using Assumption A.5 yields

$$\begin{aligned}
||\hat{\gamma}_t - \gamma_t|| &= ||E_n[b_i \psi_i^{[t,0]}(\hat{\eta}) - E[b_i \psi_i^{[t,0]}(\eta)]|| \\
&\leq ||E_n[b_i(\psi_i^{[t,0]}(\hat{\eta}) - \psi_i^{[t,0]}(\eta_0))|| + ||E_n[b_i(\psi_i^{[t,0]}(\eta_0))] - E[b_i \psi_i^{[t,0]}(\eta)]|| \\
&\leq O_p((B_n + \Lambda_n) n^{-1/2}) + O_p(n^{-1/2}) \\
&= o_p(1)
\end{aligned}$$

H.2: $||\gamma_t||$ **rate**

$$||\gamma_t|| = ||E[b_i \psi_i^{[t,0]}(\eta)]|| = ||E[b_i \tau_t(Z_i)]|| \leq \sup_{z \in \mathcal{Z}, t \in \mathcal{T}} |\tau_t(z)|| \sup_z ||b(z)|| = O(\xi_k)$$

H.3: Bound $||\gamma\gamma'||$ **and** $||E_n[\gamma b_i'(a_i + \varepsilon_i)]||$ First consider

$$\begin{aligned}
||\gamma\gamma'|| &\leq \sup_{z \in \mathcal{Z}, t \in \mathcal{T}} |\tau_t(z)|^2 ||E[b_i] E[b_i]|| \\
&\leq \sup_{z \in \mathcal{Z}, t \in \mathcal{T}} |\tau_t(z)|^2 ||E[b_i b_i']|| \\
&= O_p(1)
\end{aligned}$$

by Assumption A.1 and A.6. For the second term note that using H.2 yields

$$\begin{aligned}||E_n[\gamma b_i']|| &\leq \sup_{z\in\mathcal{Z},t\in\mathcal{T}} |\tau_t(z)|(||E_n[b_i b_i']|| + o_p(1)\ ) \\ &= O_p(1)\end{aligned}$$

as $E_n[b_i b_i'] > E_n[b_i]E[b_i]$ w.p.a. 1 and Assumption A.1. Thus

$$\begin{aligned}||E_n[\gamma b_i'(a_i+\varepsilon_i)]|| &\leq \max_{1\leq i\leq n}|a_i+\varepsilon_i|||E_n[\gamma b_i']|| \\ &= O_p(n^{1/m})O_p(1) \\ &= O_p(n^{1/m})\end{aligned}$$

where the max rate follows from Assumption A.4 and the boundedness of a_i.

D.3 Asymptotic Normality

D.3.1 Asymptotically linear representation

First we provide some additional auxiliary results:

(i) Linearization of the unconditional weights:

$$\begin{aligned}\frac{\hat{\pi}_t}{\sum_{t\neq 0}\hat{\pi}_t} - \frac{\pi_t}{\sum_{t\neq 0}\pi_t} &= \frac{\hat{\pi}_t}{1-\hat{\pi}_0} - \frac{\pi_t}{1-\pi_0} \\ &= \frac{\hat{\pi}_t(1-\pi_0) - (1-\pi_0)\pi_t + (1-\pi_0)\pi_t - \pi_t(1-\hat{\pi}_0)}{(1-\pi_0)(1-\hat{\pi}_0)} \\ &= \frac{1-\pi_0}{1-\hat{\pi}_0}\frac{1}{(1-\pi_0)^2}E_n[([D_{t,i}-\pi_t](1-\pi_0) + [D_{0,i}-\pi_0]\pi_t)]\end{aligned}$$

(ii) Simplification of estimated weights due to multiplicative structure:

$$\begin{aligned}
&\sqrt{n}E_n[b_i(\psi_i(\hat{\eta},\hat{\pi})-\psi_i(\hat{\eta},\pi))] \\
&\qquad = \sum_{t\neq 0} E_n[b_i\psi_i^{[t,0]}(\hat{\eta})]\sqrt{n}\left(\frac{\hat{\pi}_t}{\sum_{t\neq 0}\hat{\pi}_t}-\frac{\pi_t}{\sum_{t\neq 0}\pi_t}\right) \\
&\qquad = \sum_{t\neq 0}\frac{E[b_i\psi_i^{[t,0]}(\eta)]}{(1-\pi_0)^2}G_n[(D_{t,i}(1-\pi_0)+D_{0,i}\pi_t)]+O_p(n^{-1/2}) \\
&\qquad = G_n\left[\sum_{t\neq 0}\frac{E[b_i\tau_t(Z_i)](D_{t,i}(1-\pi_0)+D_{0,i}\pi_t)}{(1-\pi_0)^2}\right]+O_p(n^{-1/2}) \\
&\qquad \equiv G_n[\sum_{t\neq 0}E[b_i\tau_t(Z_i)]a_i^{[t]}]+O_p(n^{-1/2}) \\
&\qquad = G_n[\sum_{t\neq 0}\gamma_t a_i^{[t]}]+o_p(1).
\end{aligned}$$

where the second equality and its remainder follows from H.1.

(iii) Simplification for the estimated nuisance part:

$$\sqrt{n}E_n[b_i(\psi_i(\hat{\eta},\pi)-\psi_i(\eta,\pi))]=\sum_{t\neq 0}\frac{\pi_t}{\sum_{t\neq 0}\pi_t}\sqrt{n}E_n[b_i(\psi_i^{[t,0]}(\hat{\eta})-\psi_i^{[t,0]}(\eta))]$$

implies that

$$\sqrt{n}||E_n[b_i(\psi_i(\hat{\eta},\pi)-\psi_i(\eta,\pi))]||=O_p(B_n+\Lambda_n)=o(1)$$

by the triangle inequality together with SC, Lemma A.3.

(iv) LLN for Q matrix (Rudelson, 1999): Let $Q=\frac{1}{n}\sum_{i=1}^n E[b_ib_i']$ and assume that Q has bounded eigenvalues from above and below uniformly over n.

$$\begin{aligned}
E||\hat{Q}-Q|| &\lesssim_P \frac{\xi_k^2\log k}{n}+\sqrt{\frac{\xi_k^2||Q||\log k}{n}} \\
&\lesssim_P \sqrt{\frac{\xi_k^2\log k}{n}}
\end{aligned}$$

due to the bounded eigenvalues.

(v) Bound approximation error with G_n: Let α be a k dimensional vector of unit norm, i.e. $||\alpha||=1$ (or any other constant).

misspecification part

$$||\alpha'(\hat{Q}^{-1} - Q^{-1})G_n[b_i r_i]|| \lesssim_P \sqrt{\frac{\xi_k^2 \log k}{n}} \xi_k c_k \sqrt{k} = o(1)$$

where the last equality comes from A.2. See BCCK, proof of Theorem 4.1 for the derivations.

regular part We consider the simplified case with approximation term for the treatment probabilities, i.e. $\gamma = \gamma_t$ and $a_i = a_i^{[t]}$ as J is finite and all elements are of the same rate and therefore this will not affect any bounds in the following. Now note that a_i are iid and bounded. Denote x_i as the set of all observables for observation i. We have that

$$V[G_n[b_i\varepsilon_i + \gamma a_i]|x_1, \ldots, x_n] \leq C(\hat{Q} + \gamma\gamma') \leq C(\hat{Q} + Q)$$

in a positive semi-definite sense. The second bound follows from H.3. Now note that

$$E[\alpha'(\hat{Q}^{-1} - Q^{-1})G_n[b_i\varepsilon_i + \gamma a_i]|x_1, \ldots, x_n] = 0.$$

Thus by A.1, we have that

$$\begin{aligned}
&\alpha'(\hat{Q}^{-1} - Q^{-1})V[G_n[b_i\varepsilon_i + \gamma a_i]|x_1, \ldots, x_n](\hat{Q}^{-1} - Q^{-1})\alpha \\
&\qquad \lesssim C(||\hat{Q}|| + ||Q||)||\hat{Q}^{-1}||^2||||Q^{-1}||^2||\hat{Q} - Q||^2 \\
&\qquad \lesssim_P \frac{\xi_k^2 \log k}{n}
\end{aligned}$$

by (iv). Chebyshev's inequality then implies that

$$\alpha'(\hat{Q}^{-1} - Q^{-1})G_n[b_i\varepsilon_i + \gamma a_i] \lesssim_P \sqrt{\frac{\xi_k^2 \log k}{n}}$$

(vi) Asymptotically linear representation: Now we expand the original estimator

around the BLP:

$$\begin{aligned}
\sqrt{n}\alpha'(\hat{\beta}-\beta_0) &= \alpha'\hat{Q}^{-1}\sqrt{n}E_n[b_i(\psi(\hat{\eta},\hat{\pi})-b_i'\beta_0)] \\
&= \alpha'\hat{Q}^{-1}\Big(\sqrt{n}E_n[b_i(\psi(\hat{\eta},\pi)-\psi(\eta_0,\pi))]+\sqrt{n}E_n[b_i(\psi(\eta_0,\pi)-b_i'\beta_0)] \\
&\quad +\sqrt{n}E_n[b_i(\psi(\hat{\eta},\hat{\pi})-\psi(\hat{\eta},\pi))]\Big) \\
&= \alpha'(Q^{-1}+(\hat{Q}^{-1}-Q^{-1}))\Big(\sqrt{n}E_n[b_i(\psi(\hat{\eta},\pi)-\psi(\eta_0,\pi))]+G_n[b_i(r_i+\varepsilon_i)+\gamma a_i]\Big) \\
&= \alpha'Q^{-1}G_n[b_i\varepsilon_i+\gamma a_i]+R_n(\alpha)
\end{aligned}$$

where with probability approaching one

$$\sup_{\eta\in T_n}||R_n(\alpha)|| = O_p(B_n\vee\Lambda_n\vee\sqrt{\frac{\xi_k^2\log k}{n}}(1+c_k\xi_k\sqrt{k}\wedge\xi_k^2)) = o(1)$$

where we use (i) - (v) from above and assumptions A.1 - A.6. The result differs in the sense that the decomposition term has a different first order term but the remainder order is the same as in SC under the given assumptions.

(vii) second decomposition term $\Delta = nATE - rATE$: Note that due to the multiplicative structure of the decomposition we have that for any η

$$\begin{aligned}
\psi_i^{[nATE]}(\eta,\hat{\pi})-\psi_i^{[rATE]}(\eta,\hat{\pi})-(\psi_i^{[nATE]}(\eta,\pi)-\psi_i^{[rATE]}(\eta,\pi)) \\
&= \sum_{t\neq 0}\psi_i^{[t,0]}(\eta)\left[\frac{\pi_t}{\sum_{t\neq 0}\pi_t}-\frac{\hat{\pi}_t}{\sum_{t\neq 0}\pi_t}\right] \\
&= -(\psi_i^{[rATE]}(\eta,\hat{\pi})-\psi_i^{[rATE]}(\eta,\pi))
\end{aligned}$$

Thus we obtain an analogous asymptotically linear representation with a simple sign flip which does not affect the rates of any bounds used above:

$$\sqrt{n}\alpha'(\hat{\beta}-\beta_0) = \alpha'Q^{-1}G_n[b_i\varepsilon_i-\gamma a_i]+R_n(\alpha)$$

Note that the interpretation of ε_i and r_i are different as they are now estimation and approximation error for the conditional mean of the Δ decomposition term.

D.3.2 Asymptotic normality with known variance

Here we show that

$$\frac{\sqrt{n}\alpha'(\hat{\beta}-\beta_0)}{\alpha'\Omega_0\alpha} \xrightarrow{d} \mathcal{N}(0,1)$$

with

$$\Omega_0 = Q^{-1}V[b_i\varepsilon_i + \gamma a_i]Q^{-1} = Q^{-1}V_0Q^{-1}.$$

Let

$$\begin{aligned}
\Omega_0 &= Q^{-1}V[b_i\varepsilon_i + \gamma a_i]Q^{-1} = Q^{-1}V_0Q^{-1} \\
\Omega_1 &= Q^{-1}E[b_ib_i'\varepsilon_i^2]Q^{-1} = Q^{-1}V_1Q^{-1} \\
\Omega_2 &= Q^{-1}V[\gamma a_i]Q^{-1} = Q^{-1}V_2Q^{-1}
\end{aligned}$$

These matrices have bounded eigenvalues uniformly over n due to Assumptions A.1 and A.6 (see also H.3 for the auxiliary step in Ω_2).

Now we use the following auxiliary steps:

(viii) Ratio bounds The eigenvalue assumptions imply that the ratio of quadratic forms are bounded, i.e. for j=1,2

$$0 < \underline{\lambda} < \frac{\lambda_{min}(\Omega_j)}{\lambda_{max}(\Omega_0)} \leq \frac{\alpha'\Omega_j\alpha}{\alpha'\Omega_0\alpha} \leq \frac{\lambda_{max}(\Omega_j)}{\lambda_{min}(\Omega_0)} < \bar{\lambda} < \infty$$

(ix) Square bounds for sum of two random variables. Let A, B denote two continuous random variables. Note that for any $\delta > 0$ we have that

$$\begin{aligned}
E[A^2\mathbb{1}(|A+B|>\delta)] &\leq E[A^2\mathbb{1}(|A|+|B|>\delta)(\mathbb{1}(A>B)+\mathbb{1}(A\leq B))] \\
&\leq E[A^2\mathbb{1}(2|A|>\delta)\mathbb{1}(A>B)] + E[B^2\mathbb{1}(2|B|>\delta)\mathbb{1}(A\leq B)] \\
&\leq E[A^2\mathbb{1}(A>\delta/2)] + E[B^2\mathbb{1}(B>\delta/2)]
\end{aligned}$$

Thus we obtain that

$$\begin{aligned} E[(A+B)^2\mathbb{1}(|A+B|>\delta)] &\leq 2E[(A^2+B^2)\mathbb{1}(|A+B|>\delta)] \\ &\leq 4(E[A^2\mathbb{1}(A>\delta/2)]+E[B^2\mathbb{1}(B>\delta/2)]) \end{aligned}$$

Now we use (viii), (ix), and H.2 and H.3 to examine the Lindeberg condition for the empirical process. Without loss of generality, assume that $R_n(\alpha)\to 0$. The part of the approximation error could also be kept analogously to BCCK, proof of Theorem 4.2 under the given assumptions. We obtain that

$$\frac{\sqrt{n}\alpha'(\hat{\beta}-\beta_0)}{\sqrt{\alpha'\Omega_0\alpha}} = \frac{\alpha'}{\sqrt{\alpha'\Omega_0\alpha}}Q^{-1}G_n[b_i\varepsilon_i+\gamma a_i]+o_p(1)$$

Note that the $b_i\varepsilon_i+\gamma a_i$ are iid. Thus, by construction, the variance of the leading term is one. Now note that for each $\delta>0$

$$\begin{aligned} &\sum_{i=1}^n E\left[\left|\frac{\alpha'}{\sqrt{\alpha'\Omega_0\alpha}}Q^{-1}(b_i\varepsilon_i+\gamma a_i)\right|^2\left|\mathbb{1}\left(\left|\frac{\alpha'}{\sqrt{\alpha'\Omega_0\alpha}}Q^{-1}(b_i\varepsilon_i+\gamma a_i)\right|>\delta\right)\right] \\ &\quad\leq 4\sum_{i=1}^n\left(\left(\frac{\alpha'\Omega_1\alpha}{\alpha'\Omega_0\alpha}\right)\left(E\left[\left|\frac{\alpha'}{\sqrt{\alpha'\Omega_1\alpha}}Q^{-1}(b_i\varepsilon_i)\right|^2\left|\mathbb{1}\left(\left|\frac{\alpha'}{\sqrt{\alpha'\Omega_1\alpha}}Q^{-1}b_i\varepsilon_i\right|>\delta/2\sqrt{\frac{\alpha'\Omega_0\alpha}{\alpha'\Omega_1\alpha}}\right)\right]\right)\right. \\ &\qquad\left.+\left(\frac{\alpha'\Omega_2\alpha}{\alpha'\Omega_0\alpha}\right)\left(E\left[\left|\frac{\alpha'}{\sqrt{\alpha'\Omega_2\alpha}}Q^{-1}\gamma a_i\right|^2\left|\mathbb{1}\left(\left|\frac{\alpha'}{\sqrt{\alpha'\Omega_2\alpha}}Q^{-1}\gamma a_i\right|>\delta/2\sqrt{\frac{\alpha'\Omega_0\alpha}{\alpha'\Omega_2\alpha}}\right)\right]\right)\right) \\ &\quad\leq 4\sum_{i=1}^n\left(\bar{\lambda}\left(E\left[\left|\frac{\alpha'}{\sqrt{\alpha'\Omega_1\alpha}}Q^{-1}(b_i\varepsilon_i)\right|^2\left|\mathbb{1}\left(\left|\frac{\alpha'}{\sqrt{\alpha'\Omega_1\alpha}}Q^{-1}b_i\varepsilon_i\right|>\delta/(2\sqrt{\underline{\lambda}})\right)\right]\right)\right. \\ &\qquad\left.+\bar{\lambda}\left(E\left[\left|\frac{\alpha'}{\sqrt{\alpha'\Omega_2\alpha}}Q^{-1}\gamma a_i\right|^2\left|\mathbb{1}\left(\left|\frac{\alpha'}{\sqrt{\alpha'\Omega_2\alpha}}Q^{-1}\gamma a_i\right|>\delta/(2\sqrt{\underline{\lambda}})\right)\right]\right)\right) \\ &\quad= o(1) \end{aligned}$$

where the last line follows from A.2 and H.3 together with the same arguments as in BCCK, Proof of Theorem 4.2. Note that we only need a uniform integrability condition for ε_i as in BCCK, not for a_i as these are (uniformly) bounded random variables. The CLT in then follows from the sufficiency of Lindeberg's condition.

D.4 Asymptotic Variance Estimation

D.4.1 Variance Decomposition

Define

$$\Sigma = E[(b_i\varepsilon_i - \gamma a_i)(b_i\varepsilon_i - \gamma a_i)']$$
$$\Sigma_n = E_n[(b_i\varepsilon_i - \gamma a_i)(b_i\varepsilon_i - \gamma a_i)']$$
$$\hat{\Sigma} = E_n[(b_i e_i - \hat{\gamma}\hat{a}_i)(b_i e_i - \hat{\gamma}\hat{a}_i)']$$

with $e_i = \psi_i(\hat{\eta}_0, \hat{\pi}) - b_i'\hat{\beta}$ and $\hat{a}_i = a_i(\hat{\pi}_0 - \pi_0)$ obtained by replacing the true probabilities π in a_i with the sample estimates $\hat{\pi}$. Using $\hat{\pi}_0$ here is without loss of generality as all $\hat{\pi}$ have the same convergence rate. In the following we proof that $||\hat{\Sigma} - \Sigma|| = o_p(1)$. Consider the decomposition:

$$\begin{aligned}
&||\hat{\Sigma} - \Sigma_n|| \\
&= ||E_n[(b_i e_i - b_i\varepsilon_i + \gamma a_i - \hat{\gamma}a_i)(b_i e_i - b_i\varepsilon_i + 2b_i\varepsilon_i - [\hat{\gamma}\hat{a}_i - \gamma a_i + 2\gamma a_i])]|| \\
&\leq 2(||E_n[(\hat{\gamma}\hat{a}_i - \gamma a_i)(\gamma a_i + b_i\varepsilon_i)]|| + ||E_n[(b_i e_i - b_i\varepsilon_i)(\gamma a_i + b_i\varepsilon_i)]|| \\
&\quad + ||E_n[(b_i e_i - b_i\varepsilon_i)(b_i e_i - b_i\varepsilon_i)']|| + ||E_n[(\hat{\gamma}\hat{a}_i - \gamma a_i)(\hat{\gamma}\hat{a}_i - \gamma a_i)']||) \\
&= 2(M.1 + M.2 + M.3 + M.4)
\end{aligned}$$

where the first equality comes from the definition and the second inequality from the binomial formula $2E[aa'] + E[bb'] \geq E[(a+b)(a+b)']$. Decomposing the individual term $M.1$ yields:

$$\begin{aligned}
M.1 &\leq ||E_n[(\hat{\gamma} - \gamma)(\hat{a}_i - a_i)(b_i\varepsilon_i + \gamma a_i)']|| + ||E[\gamma(\hat{a}_i - a_i)(b_i\varepsilon_i + \gamma a_i)']|| \\
&\quad + ||E[(\hat{\gamma} - \gamma)a_i(b_i\varepsilon_i + \gamma a_i)']|| \\
&= M.1a + M.1b + M.1c.
\end{aligned}$$

Further bounding yields

$$\begin{aligned} M.1a &\leq ||\hat{\gamma} - \gamma|||\hat{\pi}_0 - \pi_0|(\sup_z ||b(z)||E_n[|a_i\varepsilon_i|] + ||\gamma||E_n[a_i^2])|| \\ &= O_p(n^{-1/2})O_p(n^{-1/2})[\xi_k O_p(1) + \xi_k O_p(1)] \\ &= O_p(x_i/n) \\ &= o_p(1) \end{aligned}$$

where the rates come from H.1 and H.2 together with Assumption A.2. Moreover

$$\begin{aligned} M1.b &\leq |\hat{\pi}_0 - \pi|(||E_n[\gamma b_i' a_i \varepsilon_i]|| + ||\gamma\gamma'||E_n[a_i^2]) \\ &= O_p(n^{-1/2})(O_p(n^{1/m}) + O_p(1)) \\ &= o_p(1) \end{aligned}$$

by using H.3. For the last term we have that

$$\begin{aligned} M1.c &\leq ||\hat{\gamma} - \gamma||(\sup_z ||b(z)||E_n[|a_i\varepsilon_i|] + ||\gamma||E_n[a_i^2])|| \\ &= O_p(n^{-1/2})\xi_k O_p(1) \\ &= o_p(1) \end{aligned}$$

similarly to $M1.a$ using H.1 and H.2 together with Assumption A.2. For $M.2$ note that Assumption A.5 implies that $\max_{1\leq i\leq n} |e_i - \varepsilon_i| \leq \kappa_n^1$ with probability going to one and a_i is bounded a.s. and $\max_{1\leq i\leq n} |\varepsilon_i| = O_p(n^{1/m})$ due to Assumption A.4. Thus

$$\begin{aligned} M.2 &= ||E[(e_i - \varepsilon_i)(b_i\gamma' a_i + b_i b_i' \varepsilon_i)]|| \\ &\leq \max_{1\leq i\leq n} |e_i - \varepsilon_i|(\max_{1\leq i\leq n} |a_i|||E[b_i\gamma']|| + \max_{1\leq i\leq n} |\varepsilon_i|||E_n[b_i b_i']|| \\ &\leq \kappa_n^1 O_p(1) + \kappa_n^1 n^{1/m} O_p(1) \\ &= O_p(\kappa_n^1 n^{1/m}) \\ &= o_p(1) \end{aligned}$$

$M.3$ can be found in SC in their proof of Theorem 3.3. For that note that in their proof they rely on the conditions of BCCK on uniform error bounds for Lemma 4.2 and Theorem 4.6. It can be easily verified that these are implied by assumption A.2 and A.3. and A.7 together. They obtain that

$$M.3 = O_p(\kappa_n^2) = o_p(1)$$

For the last term, using the triangle inequality and binomial formula twice yields

$$\begin{aligned} M.4 &\leq 4(||E[(\hat{\gamma}-\gamma)(\hat{\gamma}-\gamma)'(\hat{a}_i - a_i)^2]|| + ||E[\gamma\gamma'(\hat{a}_i - a_i)^2]|| + ||E[(\hat{\gamma}-\gamma)(\hat{\gamma}-\gamma)'\hat{a}_i^2|| \\ &= ||\hat{\gamma}-\gamma||^2|\hat{\pi}_0 - \pi_0|^2 O_p(1) + ||\gamma\gamma'||\hat{\pi}_0 - \pi_0|^2 O_p(1) + ||\hat{\gamma}-\gamma||^2 O_p(1) \\ &= O_p(n^{-1}) \\ &= o_p(1) \end{aligned}$$

using H.1. and H.3. Now finally note that

$$||\Sigma_n - \Sigma|| = o_p(1)$$

by (iv). The final result then follows straightforwardly from Theorem 4.6 in BCCK.

D.4.2 Adaptions to Lemma 4.2. in Belloni et al. (2015)

Here we accommodate our assumptions to obtain a result similar to Lemma 4.2 in BCCK to use the bounds from their Theorem 4.6. Note that A.1 - A.5. imply all the assumptions for Lemma 4.2, however the asymptotic linearization and the order of its remainder has to be adjusted to the different empirical process at hand that also depends on the estimated unconditional weights. In particular, we need to derive an equivalent or stronger result as in (A.52) in BCCK. For the approximation error part, the proof works analogously.

Let $\rho = (\rho_1, \ldots, \rho_n) \in P := \{\rho \in \mathbb{R}^n : \rho_i = \alpha(z)'(\hat{Q}^{-1} - Q^{-1})\gamma a_i, z \in \mathcal{Z}\}$. Define norm $||\rho||_{n,2}^2 = E_n[\rho_i^2]$. Let ω_i be independent Rademacher random variables with $P(\omega_i = 1) = P(\omega_i = -1) = 1/2$ and let $\omega = (\omega_1, \ldots, \omega_n)$ and E_ω denote the expectation w.r.t. ω. First

note that

$$\begin{aligned}
&\sup_z |\alpha(z)'(\hat{Q}^{-1} - Q^{-1})G_n[(p_i\varepsilon_i + \gamma a_i)]| \\
&\quad \leq \sup_{z\in\mathcal{Z}} |\alpha(z)'(\hat{Q}^{-1} - Q^{-1})G_n[p_i\varepsilon_i]| + \sup_{z\in\mathcal{Z}} |\alpha(z)'(\hat{Q}^{-1} - Q^{-1})G_n[\gamma a_i]| \\
&\quad = M.5 + M.6
\end{aligned}$$

$M.5$ is controlled for in BCCK, proof of Lemma 4.2. For $M.6$ note that by Dudley's inequality

$$\begin{aligned}
E_\omega[\sup_{z\in\mathcal{Z}} |\alpha(z)'(\hat{Q}^{-1} - Q^{-1})G_n[\omega_i\gamma a_i]|] &\leq \int_0^{2\sup_{\rho\in P}||\rho||_{n,2}} \sqrt{\log N(P, ||\cdots||_{n,2}, \varepsilon)}d\varepsilon \\
&\leq \int_0^{2C||\hat{Q}^{-1}-\hat{Q}^{-1}||||\gamma\gamma'||^{1/2}} \sqrt{\log N(P, ||\cdots||_{n,2}, \varepsilon)}d\varepsilon
\end{aligned}$$

as a_i are bounded random variables. Now note that for any $z, z' \in \mathcal{Z}$

$$\begin{aligned}
&E[(\alpha(z) - \alpha(z'))(\hat{Q}^{-1} - Q^{-1})\gamma a_i)^2]^{1/2} \\
&\qquad \leq \max_{1\leq i\leq n} |a_i|||(\alpha(z) - \alpha(z')||||(\hat{Q}^{-1} - Q^{-1})||||\gamma\gamma'||^{1/2} \\
&\qquad \leq \xi_k^L C||(\hat{Q}^{-1} - Q^{-1})||||\gamma\gamma'||^{1/2}||z - z'||
\end{aligned}$$

and thus

$$N(P, ||\cdots||_{n,2}, \varepsilon) \leq \left(\xi_k^L C_2||(\hat{Q}^{-1} - Q^{-1})||||\gamma\gamma'||^{1/2}/\varepsilon\right)^d$$

for some $C_2 > 0$. Plugging this back into the integral yields

$$\begin{aligned}
&\int_0^{2C||Q^{-1}-Q^{-1}||||\gamma\gamma'||^{1/2}} \sqrt{\log N(P, ||\cdots||_{n,2}, \varepsilon)} d\varepsilon \\
&\qquad \leq \xi_k^L C||(\hat{Q}^{-1} - Q^{-1})||||\gamma\gamma'||^{1/2} \int_0^2 \sqrt{d \log(\xi_k^L C_2/\varepsilon) d\varepsilon} \\
&\qquad \leq ||\hat{Q} - Q||||Q^{-1}||||\hat{Q}^{-1}||||\gamma\gamma'||^{1/2} \log(\xi_k^L) O(1) \\
&\qquad \leq \sqrt{\zeta_k^2 \log(k)/n} \log(\xi_k^L) O_p(1) \\
&\qquad = O_p\left(\sqrt{\zeta_k^2 \log^2(k)/n}\right)
\end{aligned}$$

where the second last line follows from (iv), H.3, and Assumption A.1. The last line is due to Assumption A.7. This rate is clearly smaller than the rate for M.5 in BCCK $n^{1/m}\sqrt{\zeta_k^2 \log^2(k)/n}$ and therefore does not affect the relevant approximation error order. Overall we obtain:

$$\begin{aligned}
&E\left[\sup_z |\alpha(z)'(\hat{Q}^{-1} - Q^{-1})G_n[(p_i\varepsilon_i + \gamma a_i)]|\middle| z_1, \ldots, z_n\right] \\
&\qquad \leq CE\left[E_\omega\left[\sup_z |\alpha(z)'(\hat{Q}^{-1} - Q^{-1})G_n[\omega_i(p_i\varepsilon_i + \gamma a_i)]|\right]\middle| z_1, \ldots, z_n\right] \\
&\qquad = O_p\left(n^{1/m}\sqrt{\frac{\xi_k^2 \log^2(k)}{n}}\right) \\
&\qquad = o_p(1)
\end{aligned}$$

by A.2. Using (iv) then implies that $||\hat{\Omega} - \Omega_0|| = o_p(1)$. This in conjunction with the asymptotic normality above proves Theorem 5.1.

E Supplementary Material for Section 4

E.1 Neyman-orthogonality

The key insight required here is that the $nATE$ and $rATE$ scores are Neyman-orthogonal with known unconditional probabilities π_t, $t = 1, \ldots, J$. We show how the additional estimation error can be incorporated in Appendix D. Here we are concerned with the Gateaux derivative of the $nATE$ and $rATE$ scores with respect to the vector of infinite-

dimensional nuisance parameters $\eta = (\mu(x), p(x)) = (\mu_0(x), \dots, \mu_J(x), e_0(x), \dots, e_J(x))'$. As π is assumed to be known, we suppress dependence $\psi(\eta, \pi) = \psi(\eta)$ out of convenience for now. Suppressing also the dependencies of the nuisance parameters on x, we write the path-wise derivative of the conditional expectation of a score with respect to the vector of nuisance parameters as

$$\partial_\eta E[\psi_i(\eta)|X_i = x] = \partial_r E[\psi_i(\dots, \mu_t + r(\tilde{\mu}_t - \mu_t), \dots, e_t + r(\tilde{e}_t - e_t), \dots)|X_i = x]|_{r=0}$$

First, we revisit Neyman-orthogonality of the doubly robust score:

$$\begin{aligned}
&\partial_r E[\psi_i^{[t]}(\eta + r(\tilde{\eta} - \eta))|X_i = x]|_{r=0} \\
&\quad = \partial_r E\left[(\mu_t + r(\tilde{\mu}_t - \mu_t)) + \frac{D_{t,i}Y_i}{e_t + r(\tilde{e}_t - e_t)} - \frac{D_{t,i}(\mu_t + r(\tilde{\mu}_t - \mu_t))}{e_t + r(\tilde{e}_t - e_t)}\middle| X_i = x\right]\Bigg|_{r=0} \\
&\quad = (\tilde{\mu}_t - \mu_t) - \frac{e_t\mu_t(\tilde{e}_t - e_t)}{e_t^2} - \frac{e_t^2(\tilde{\mu}_t - \mu_t) - e_t\mu_t(\tilde{e}_t - e_t)}{e_t^2} \\
&\quad = 0
\end{aligned}$$

where we use that $E[D_{t,i}Y_i|X_i = x] = E[D_{t,i}\sum_t D_{t,i}Y_i(t)|X_i = x] = E[D_{t,i}Y_i(t)|X_i = x] = e_t\mu_t$ by the observational rule and Assumption 1.

E.1.1 rATE

As the $rATE$ score is a linear combination of doubly robust scores, it inherits the Neyman-orthogonality of its components:

$$\begin{aligned}
&\partial_r E[\psi_i^{[rATE]}(\eta + r(\tilde{\eta} - \eta))|X_i = x]|_{r=0} \\
&\quad = \sum_{t\neq 0} \frac{\pi_t}{1 - \pi_0} \partial_r E[\psi_i^{[t]}(\eta + r(\tilde{\eta} - \eta))|X_i = x]|_{r=0} \\
&\qquad - \partial_r E[\psi_i^{[0]}(\eta + r(\tilde{\eta} - \eta))|X_i = x]|_{r=0} \\
&\quad = 0
\end{aligned}$$

E.1.2 nATE

The $nATE$ score differs from the standard doubly robust scores but can still be shown to be Neyman-orthogonal:

$$
\begin{aligned}
&\partial_r E[\psi_i^{[nATE]}(\eta + r(\tilde{\eta} - \eta))|X_i = x]|_{r=0} \\
&\quad = \partial_r E\left[\frac{\sum_{t\neq 0}[(\mu_t + r(\tilde{\mu}_t - \mu_t))(e_t + r(\tilde{e}_t - e_t))]}{\sum_{t\neq 0}(e_t + r(\tilde{e}_t - e_t))} + \frac{D_i Y_i}{\sum_{t\neq 0}(e_t + r(\tilde{e}_t - e_t))}\right. \\
&\qquad \left. - \frac{D_i \sum_{t\neq 0}[(\mu_t + r(\tilde{\mu}_t - \mu_t))(e_t + r(\tilde{e}_t - e_t))]}{[\sum_{t\neq 0}(e_t + r(\tilde{e}_t - e_t))]^2}\middle| X_i = x\right]\Bigg|_{r=0} \\
&\quad - \partial_r E[\psi_i^{[0]}(\eta + r(\tilde{\eta} - \eta))|X_i = x]|_{r=0} \\
&\quad = \frac{\sum_{t\neq 0}[\mu_t(\tilde{e}_t - e_t) + e_t(\tilde{\mu}_t - \mu_t)]\sum_{t\neq 0} e_t}{[\sum_{t\neq 0} e_t]^2} - \frac{\sum_{t\neq 0}\mu_t e_t \sum_{t\neq 0}(\tilde{e}_t - e_t)}{[\sum_{t\neq 0} e_t]^2} \\
&\quad - \frac{\sum_{t\neq 0} e_t\mu_t \sum_{t\neq 0}(\tilde{e}_t - e_t)}{[\sum_{t\neq 0} e_t]^2} - \frac{\sum_{t\neq 0} e_t \sum_{t\neq 0}[\mu_t(\tilde{e}_t - e_t) + e_t(\tilde{\mu}_t - \mu_t)]}{[\sum_{t\neq 0} e_t]^2} \\
&\quad + 2\frac{\sum_{t\neq 0}\mu_t e_t \sum_{t\neq 0}(\tilde{e}_t - e_t)}{[\sum_{t\neq 0} e_t]^2} - \partial_r E[\psi_i^{[0]}(\eta + r(\tilde{\eta} - \eta))|X_i = x]|_{r=0} \\
&\quad = 0
\end{aligned}
$$

where we use that $E[D_i Y_i|X_i = x] = E[D_i \sum_t D_{t,i} Y_i(t)|X_i = x] = \sum_{t\neq 0} e_t\mu_t$ by the observational rule and Assumption 1. Consequently, the difference between the $nATE$ and $rATE$ score that forms the Δ score is Neyman-orthogonal as well:

$$\partial_\eta E[\psi_i^{[nATE]}(\eta) - \psi_i^{[rATE]}(\eta)|X_i = x] = 0$$

F Supplementary Material for Section 6

Consider the following sequence of probability measures $\{F_n\}_{n=1}^{\infty}$ such that $\mathcal{T} \equiv \mathcal{T}_n = \{0, 1, \ldots, J_n\}$ with $J_n \to \infty$. We still assume strong overlap for control propensities, i.e $\inf_{x\in\mathcal{X}} e_0(x) > \underline{e} > 0$ almost surely. Now additionally assume that almost surely for $t \neq 0$

$$e_t(X_i) \equiv e_{t,n}(X_i) = \frac{\underline{e}_{t,n}(X_i)}{J_n} \tag{22}$$

such that

$$\inf_{n\geq 1}\inf_{x\in\mathcal{X}} \underline{e}_{t,n}(x) > \underline{e} > 0 \tag{23}$$

$$\sup_{n\geq 1}\sup_{x\in\mathcal{X}} \underline{e}_{t,n}(x) < C \tag{24}$$

for $t \neq 0$. (23) is a uniform strong overlap assumption for the *rescaled* generalized propensities $\underline{e}_{t,n}(x) = J_n e_{t,n}(x)$. This allows the generalized propensity scores to converge to zero, i.e. for many treatment with small selection probabilities but controls their rate. This assumption says that the increase in the number of treatments will qualitatively be similar over the support of confounders, i.e. units have a somewhat comparable decrease in their relative propensities along the sequences. (24) is only for normalization in what follows. Note that this also implies that the unconditional version probabilities $\pi_t \equiv \pi_{t,n}$ follow

$$\begin{aligned}
\pi_{t,n} &= E_{F_n}[e_{t,n}(X_i)] \\
&= E_{F_n}\left[\frac{\underline{e}_{t,n}(X_i)}{J_n}\right] \\
&\equiv \frac{\underline{\pi}_{t,n}}{J_n}
\end{aligned}$$

with $\underline{\pi}_{t,n}$ being uniformly bounded away from zero and from above. Now we consider the variance along sequences of probability measures for the $rATE$. The dominating term for the variance of the $rATE$ estimator is given by

$$\sqrt{n}(\hat{\theta}_{rATE} - \theta_{rATE}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{t\neq 0}\pi_t \frac{D_i(t)\varepsilon_i(t)}{e_t(X_i)}$$

where $\varepsilon_i(t)$ denotes the conditional mean error of potential outcome for version t. Without loss of generality we assume conditional homoskedasticity $E[\varepsilon_i(t)^2|X_i] = \sigma^2$ for all t. The

variance is then given by

$$\begin{aligned} V_{F_n}[\sqrt{n}(\hat{\theta}_{rATE} - \theta_{rATE})] &= E_{F_n}\left[\left(\sum_{t\neq 0} \pi_{t,n} \frac{D_i(t)\varepsilon_i(t)}{e_{t,n}(X_i)}\right)^2\right] \\ &= E_{F_n}\left[\sum_{t\neq 0} \pi_{t,n}^2 \frac{D_i(t)\varepsilon_i(t)^2}{e_{t,n}(X_i)^2}\right] \\ &= \sigma^2 \sum_{t\neq 0} \pi_{t,n}^2 E_{F_n}\left[\frac{1}{e_{t,n}(X_i)}\right] \end{aligned}$$

where we used the iid assumption in the first equation, orthogonality of the version dummies $D_i(t)D_i(t') = 0$ for all $t \neq t'$ in the second, and homoskedasticity as well as treatment conditional independence Assumption 1 in the third. Plugging in the sequences yields that

$$\begin{aligned} V_{F_n}[\sqrt{n}(\hat{\theta}_{rATE} - \theta_{rATE})] &= \sigma^2 \sum_{t\neq 0} \frac{\pi_{t,n}^2}{J_n^2} E_{F_n}\left[\frac{J_n}{\underline{e}_{t,n}(X_i)}\right] \\ &\leq \sigma^2 J_n \times \frac{C^2}{J_n}\underline{e}^{-1} \\ &= O(1) \end{aligned}$$

uniformly over n by (23) and (24). Thus the variance is always bounded along the sequences, no matter the rate at which $J_n \to \infty$. In contrast to that, consider the estimation of a treatment effect in a standard multi-valued framework. Here the leading term for the asymptotic variance for the estimator of a potential outcome t is given by

$$\sqrt{n}(\hat{\mu}_t - \mu_t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{D_i(t)\varepsilon_i(t)}{e_t(X_i)} \tag{25}$$

for every $t \in \mathcal{T}$. Along the sequences this has variance

$$\begin{aligned} V_{F_n}[\sqrt{n}(\hat{\mu}_t - \mu_t)] &= \sigma^2 E_{F_n}\left[\frac{1}{e_{t,n}(X_i)}\right] \\ &= \sigma^2 J_n E_{F_n}\left[\frac{1}{\underline{e}_{t,n}(X_i)}\right] \\ &= O(J_n) \end{aligned}$$

by (24). Thus, it diverges at the same rate as the number of treatments. An equivalent argument applies to the effect parameter $\theta_{t,t'} = \mu_t - \mu_{t'}$.

G Supplementary Material for Section 7

We simulate n observations of (Y_i, X_i, T_i). Let X_i be a k-dimensional vector of uniform random variables $X_{i,j} \sim \mathcal{U}[-1, 1]$ for $j = 1, \ldots, p$ and $\varepsilon_i \sim \mathcal{N}(0, 1)$. We let $Y_i(t) = u_i$ for $t \neq 1$ and $Y_i(1) = \tau + u_i$. Treatment probabilities $P(T_i = t|X_i) = e_t(X_i)$ for $t = 0, 1 \ldots, J$ (with $t = 0$ denoting control) are generated under independence of irrelevant alternatives as

$$
\begin{aligned}
e_0(x) &= \frac{1}{1 + \sum_{j \neq 0} \exp(x_1 \beta_j)} \\
e_t(x) &= \frac{\exp(x_1 \beta_t)}{1 + \sum_{j \neq 0} \exp(x_1 \beta_j)}
\end{aligned}
$$

with $\beta_1 = 1$ and $\beta_t = 0$ for all $t \neq 1$. Thus, conditional treatment effects are given by $\tau_1(x) = \tau = 10$ and $\tau_t(x) = 0$ for all $t \neq 1$. This implies the following conditional decomposition terms (II):

$$
\begin{aligned}
E[rATE(X_i)|X_{i,1} = x_1] &= \tau \left[\frac{\pi_1}{1 - \pi_0} \right] \\
E[nATE(X_i)|X_{i,1} = x_1] &= \tau \left[\frac{e_1(x_1)}{1 - e_0(x_1)} \right] \\
E[\Delta(X_i)|X_{i,1} = x_1] &= \tau \left[\frac{e_1(x_1)}{1 - e_0(x_1)} - \frac{\pi_1}{1 - \pi_0} \right]
\end{aligned}
$$

Note that $E[X_{i,1}] = 0$ and $V[X_{i,1}] = 1/3$. Thus the best linear approximation of $E[\Delta(X_i)|X_{i,1}]$ has population parameters (α, β) with

$$\begin{aligned}
\alpha &= \tau E\left[\frac{e_1(X_{i,1})}{1-e_0(X_{i,1})} - \frac{\pi_1}{1-\pi_0}\right] - \beta E[X_{i,1}] \\
&= \tau E\left[\frac{e_1(X_{i,1})}{1-e_0(X_{i,1})} - \frac{\pi_1}{1-\pi_0}\right] \\
\beta &= \frac{\tau}{V[X_{i,1}]} E\left[\frac{e_1(X_{i,1})}{1-e_0(X_{i,1})} X_{i,1} - \frac{\pi_1}{1-\pi_0} X_{i,1}\right] \\
&= 3\tau E\left[\frac{e_1(X_{i,1})}{1-e_0(X_{i,1})} X_{i,1}\right]
\end{aligned}$$

and equivalently for the $rATE$ and $nATE$. Evaluating the expectation yields the following parameterization:

Table G.1: Monte Carlo Study: Parameterization

	$rATE$	$nATE$	Δ
α	5.127	5.000	-.127
β	0.000	2.383	2.383

H Supplementary Material for Section 8.1

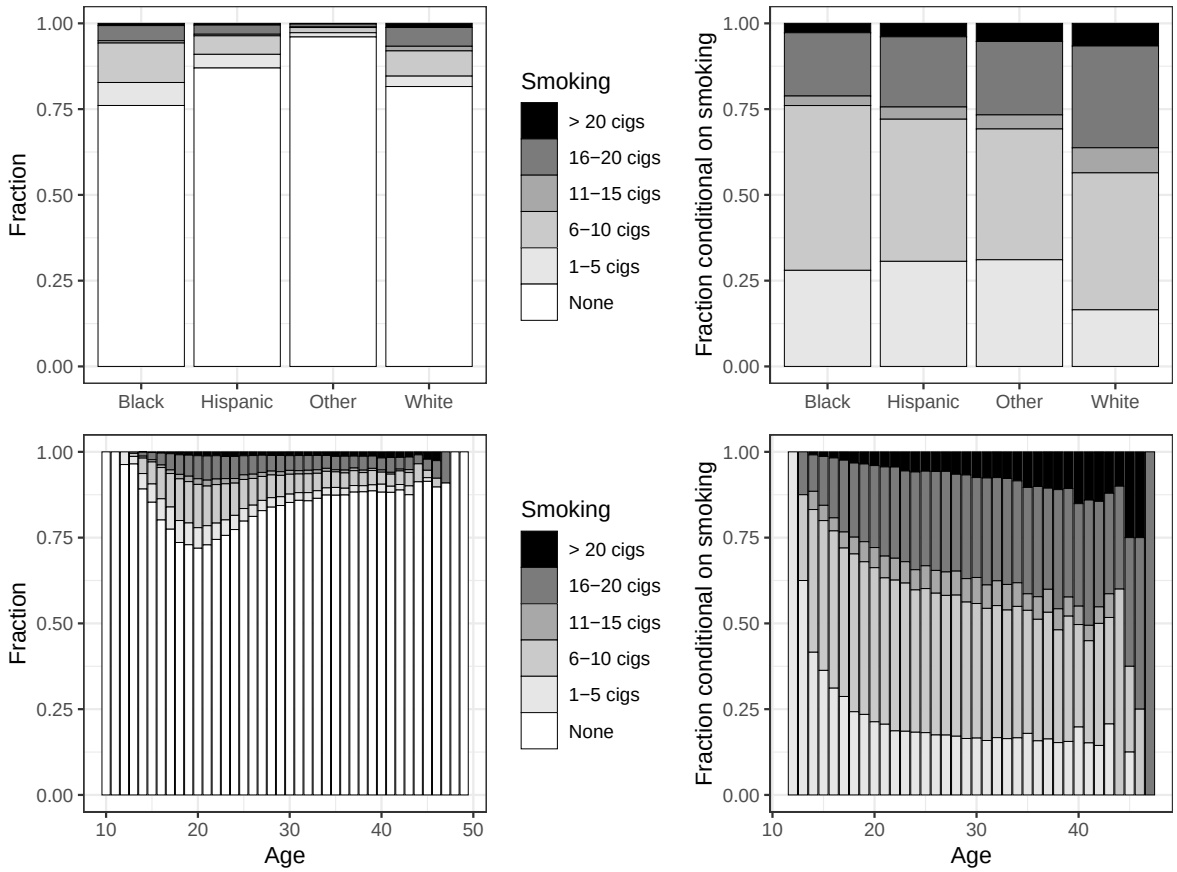
The distribution of smoking intensities is shown in Figure H.1 and Table H.1. The majority of mothers do not smoke during pregnancy ranging from 76% for Black mothers to 96% in the category "Other". However, the right panel of Figure H.1 shows that conditional on smoking white mothers and older mothers smoke more heavily.

Table H.1: Distribution of smoking intensities by ethnicity (in percent)

	Black	Hispanic	Other	White	All
> 20 cigs	0.7	0.5	0.2	1.2	1.1
16-20 cigs	4.4	2.7	0.9	5.5	5.1
11-15 cigs	0.7	0.5	0.2	1.3	1.2
6-10 cigs	11.5	5.4	1.5	7.4	7.8
1-5 cigs	6.7	4.0	1.2	3.0	3.6
None	76.1	87.0	96.0	81.6	81.2

Figure H.2 replicates the solid line of Figure 1 in Cattaneo (2010) with Double Machine Learning as a byproduct. Our results are very similar and show that average potential

Figure H.1: Distribution of smoking intensities along heterogeneity variables



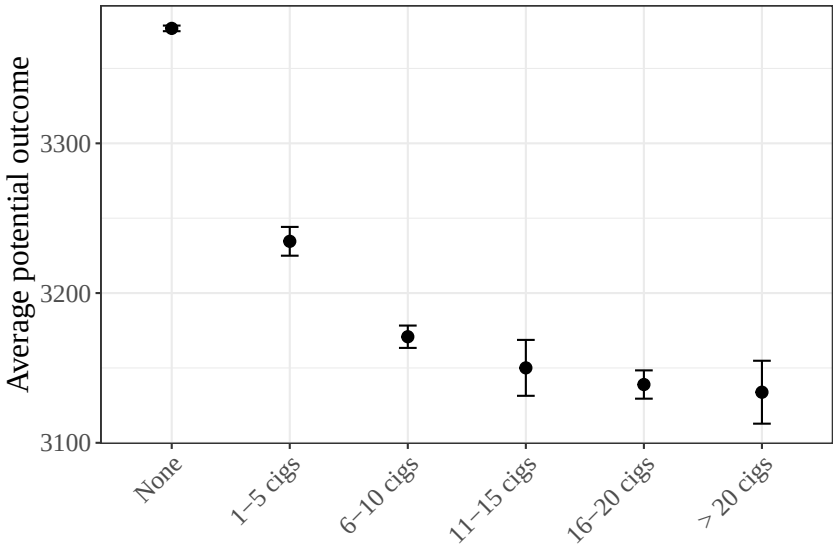
outcomes become smaller the higher the intensity of smoking.

I Supplementary Material for Section 8.2

The distribution of versions is shown in Figure I.1 and Table I.1. We observe that women are overrepresented in clerical, health and food training, while men are more likely to be observed in automechanics, welding, electrical and construction training.

As a byproduct of the decomposition estimation, we create the AIPW scores for every treatment version. This allows us to inspect their often noisily estimated average potential outcomes in Figure I.2. We observe a clear pattern. The point estimates of the predominantly male trainings are all larger than the predominantly female ones.

Figure H.2: Average potential outcomes of smoking intensities



Note: Average potential outcomes estimated with Double Machine Learning using an ensemble of Ridge, Lasso and Random Forest regression. Point estimates and 95%-confidence interval.

Figure I.1: Distribution of treatment versions by gender

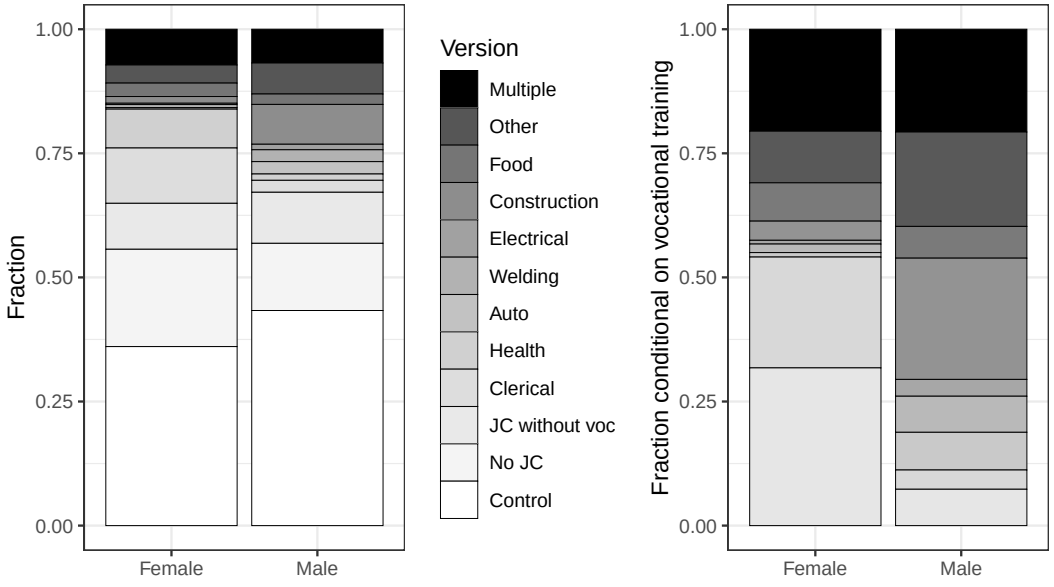
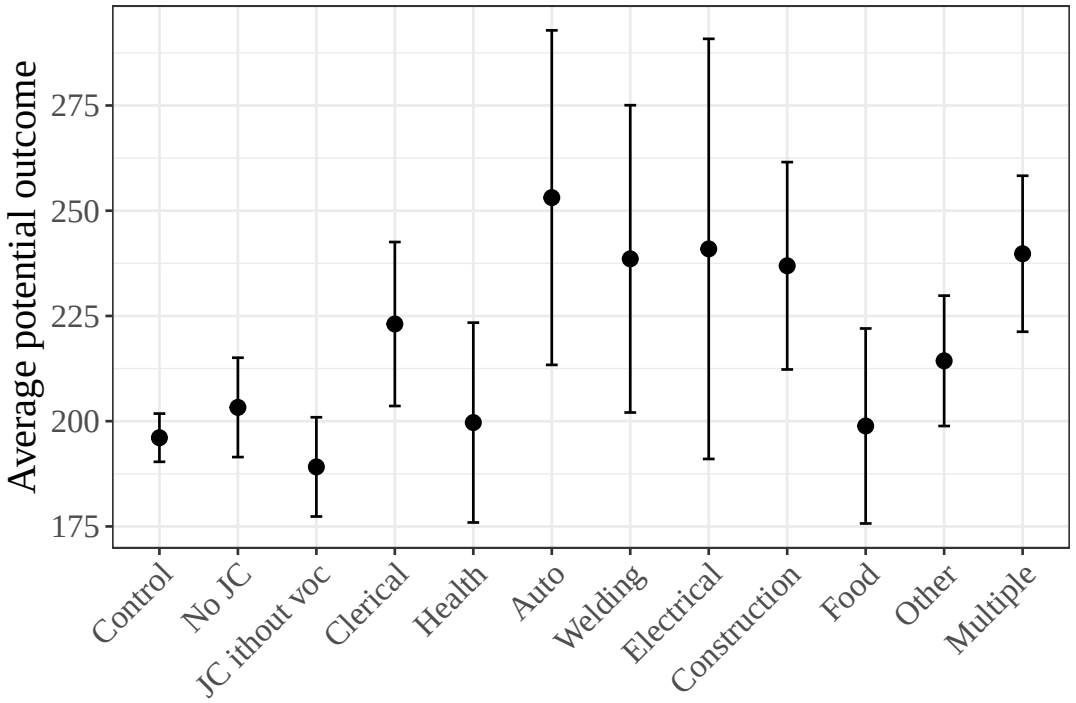


Table I.1: Share of observations in treatment versions (in percent)

	Female	Male	All
Control	36.1	43.3	40.2
No JC	19.6	13.5	16.1
JC without voc	9.3	10.3	9.8
Clerical	11.1	2.4	6.1
Health	7.8	1.3	4.1
Auto	0.3	2.5	1.6
Welding	0.6	2.4	1.6
Electrical	0.3	1.1	0.8
Construction	1.4	8.0	5.2
Food	2.7	2.1	2.4
Other	3.6	6.2	5.1
Multiple	7.2	6.8	7.0

Figure I.2: Average potential outcomes of treatment versions



Note: Average potential outcomes estimated with Double Machine Learning using an ensemble of Ridge, Lasso and Random Forest regression. Point estimates and 95%-confidence interval.