Fixed Effects and the Generalized Mundlak Estimator

DMITRY ARKHANGELSKY

CEMFI; CEPR, Spain

and

GUIDO W. IMBENS

Stanford University; SIEPR; and NBER, USA

We develop a new approach for estimating average treatment effects in observational studies with unobserved group-level heterogeneity. We consider a general model with group-level unconfoundedness and provide conditions under which aggregate balancing statistics—group-level averages of functions of treatments and covariates—are sufficient to eliminate differences between groups. Building on these results, we re-interpret commonly used linear fixed-effect regression estimators by writing them in the Mundlak form as linear regression estimators without fixed effects but including group averages. We use this representation to develop Generalized Mundlak Estimators that capture group differences through group averages of (functions of) the unit-level variables and adjust for these group differences in flexible and robust ways in the spirit of the modern causal literature.

Key words: Fixed effects, Cross-section data, Groups, Causal effects, Treatment effects, Unconfoundedness

JEL codes: C14, C21, C23

1. INTRODUCTION

A common specification for regression functions when estimating causal effects with grouped data is a fixed-effect model:

$$Y_i = \alpha_{g(i)} + W_i \tau + X_i^\top \beta + \varepsilon_i, \tag{1.1}$$

where the index $g(i)$ indicates the group a unit i belongs to, the α_g are the group fixed effects, and ε_i is an error term, independent of the regressors W_i and X_i (see Arellano, 2003; Angrist and Pischke, 2008; Wooldridge, 2010 for textbook discussions). In empirical work, the groups (also referred to as strata, subpopulations, or clusters) may correspond to states, cities, MSAs, classrooms, birth cohorts, families, siblings, or other geographic or demographic groups. The model is typically estimated by the ordinary least squares (OLS). We refer to the corresponding estimator as the fixed-effect estimator. The parameter τ, which we would like to interpret as an average

---

The editor in charge of this paper was Aureo de Paula.

causal effect of the treatment W_i (binary through most of this paper), is the object of interest. The fixed effects α_g are intended to capture unobserved differences between the groups. The motivation for including these fixed effects in the specification is that their presence improves the credibility of a causal interpretation of the fixed-effect estimator for τ (*e.g.* Arellano, 2003; Angrist and Pischke, 2008). Inference for τ is typically based on asymptotic approximations with a growing number of groups and a fixed number of units per group. In that setting one cannot rely on consistent estimation of the fixed effects because of the incidental parameter problem (*e.g.* Neyman and Scott, 1948; Lancaster, 2000).

In this paper, we make two sets of contributions. First, we unpack the assumptions underlying the fixed-effects estimator based on specification (1.1), and second, we re-interpret the fixed effect estimator and use that re-interpretation to propose a new class of estimators.

The popularity of the fixed-effect estimator may be due partly because it is viewed as accounting for differences between groups in a flexible way. However, the specification in equation (1.1) embodies multiple strong assumptions. The current paper clarifies these assumptions by fseparating them into three parts. First, the specification in equation (1.1) assumes a constant treatment effect. In practice, it is likely the effects of the treatment are heterogeneous. We show that in general the average effect of the treatment is *not* point-identified. We then describe the estimand corresponding to the fixed-effect estimator under general treatment-effect heterogeneity and demonstrate that it estimates, under some conditions, a weighted average treatment effect, with the weights depending in an unusual way on the sampling scheme. For example, we show that in the presence of treatment-effect heterogeneity, the interpretation of the fixed-effect estimator changes if we double the number of units we sample per cluster. Second, a key assumption motivating equation (1.1), what we call group unconfoundedness (Assumption 3.2 below), validates all within-group comparisons of treated and control units with the same covariates. However, the additional assumptions underlying equation (1.1) also implicitly validate some, but not all, between-group comparisons of treated and control units with the same covariates. We describe which between-group comparisons are validated by the fixed-effect specification. Third, the fixed-effect specification assumes linearity and additivity in the fixed effects, covariates, and treatments.

In the second set of contributions, we propose a new class of estimators. To motivate these new estimators we start with the re-interpretation of the fixed-effect estimator due to Mundlak (1978). Instead of thinking of the fixed-effect estimator for τ as corresponding to the specification in equation (1.1), we can also think of it as the least-squares estimator for τ corresponding to the specification

$$Y_i = W_i \tau + \overline{W}_{g(i)} \gamma + X_i^\top \beta + \overline{X}_{g(i)}^\top \delta + \eta_i, \qquad (1.2)$$

where, throughout the paper, we use $\overline{W}_g$ and $\overline{X}_g$ to denote the average value of W_i and X_i for units in group g, that is, for units with $g(i) = g$. Compared to equation (1.1), this specification does not have the fixed effects α_g. Instead, it has two additional regressors, the group averages of W_i and X_i. The least-squares estimators for τ based on equations (1.1) and (1.2) are identical, as first noted by Mundlak (1978) (see also Wooldridge, 2021). We can therefore interpret the fixed-effect estimator as controlling for group differences by simply including group averages of the covariates as additional regressors in a linear regression. This interpretation motivates our exploration of new estimators, which we refer to as Generalized Mundlak Estimators (GMEs). These estimators, for suitable defined average treatment effects as made precise in Section 3, maintain group unconfoundedness (Assumption 3.2) which justifies within-group comparisons of treated and control units. We then formulate additional assumptions that allow for some, but not all, cross-group comparisons of treated and control units. The estimators can be thought of as starting with a parsimonious specification as in equation (1.2), and, like equation (1.2), are

based on adjusting for group characteristics to justify cross-group comparisons. They differ from equation (1.2) in that they free up the specification or change the estimation procedure in three distinct ways, inspired by the modern literature on estimating average treatment effects under unconfoundedness (*e.g.* Imbens and Rubin, 2015; Abadie and Cattaneo, 2018). First, we can adjust for differences in the group averages in a more flexible, non-linear way, for example, by including non-linear functions of the averages $\overline{W}_g$ and $\overline{X}_g$ in the regression function in equation (1.2). Second, we can make the estimator more robust by estimating the propensity score and using inverse propensity score weighting in combination with regression to create doubly robust estimators (Robins and Rotnitzky, 1995; Zubizarreta, 2015; Chernozhukov *et al.*, 2017; Athey *et al.*, 2018). Third, we can account for more general differences between groups than accounted for by the fixed-effect estimator by adjusting for differences in averages of other functions of the covariates X_i and the treatment indicators W_i, for example, the average of the product, $\overline{WX}_g$. We show how the choice of group characteristics to adjust for can be motivated by structural choice models in Section 3.3.

We can capture the three modifications in the proposed GME relative to the fixed-effect estimators by an unconfoundedness assumption that strengthens group unconfoundedness. Instead of conditioning on group indicators, it assumes it is sufficient to condition on a set of group characteristics:

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i, \overline{H}_{g(i)},$$

where $Y_i(0)$ and $Y_i(1)$ are potential outcomes for unit i (Neyman, 1923/1990; Rubin, 1974; Imbens and Rubin, 2015), and $\overline{H}_{g(i)}$ is the group average of some pre-specified function $H(W_i, X_i)$ of the treatment indicators and the covariates. We refer to $\overline{H}_{g(i)}$ as a balancing score, following the terminology in Rosenbaum and Rubin (1983). By the unconfoundedness condition, treated and control units with the same values for covariates and balancing statistics can be compared even if they belong to different groups. If $H(W_i, X_i) = (W_i, X_i)$ and the adjustment is only linear, this gets us back to the fixed-effect estimator, but in general the proposed GME differs in two ways. First, it adjusts more flexibly for the components of $\overline{H}_{g(i)}$, and second, $H(\cdot)$ can contain more components than just (W_i, X_i).

Our formal asymptotic results focus on the case with a fixed number of units per group. In particular, we discuss settings where the group size is not large enough to carry out a two-stage procedure where we first estimate the effects entirely within groups by flexibly adjusting for covariates, followed by averaging over the groups. In other words, because of the incidental parameter problem (Neyman and Scott, 1948; Lancaster, 2000), we need to rely on comparisons of treated and control units in different groups. In addition, there is concern that accounting for the group differences solely through additive fixed effects as in specification (1.1) is not sufficient to adjust for all relevant differences (*e.g.* Altonji and Matzkin, 2005; Imai and Kim, 2019). Finally, in this setting we cannot estimate the propensity score, that is, the conditional probability of treatment assignment, as a function of group membership.

2. EXAMPLES

In this section, we discuss the main insights of the paper in the context of two examples. We start with a case with no covariates and present the interpretation of equation (1.1) under general heterogeneity in treatment effects. We then introduce binary covariates and connect (1.1) to a particular unconfoundedness condition. We finish the section with a set of practical recommendations.

2.1. *No covariates*

A key assumption underlying most group analyses is that assignment is unconfounded within the groups. Let each unit i in a large population be characterized by a pair of potential outcomes $(Y_i(0), Y_i(1))$ and group label $L_{g(i)}$. These labels can correspond to names, for example, geographic locations, legal entities, and markets. We use the labels at this stage to make clear that there is not necessarily an ordering to the groups, nor a distance measure that allows us to say that some groups are closer to each other. Suppose each unit is assigned to a binary treatment W_i. The sampling process has two stages. First, we randomly sample M groups from a large population of groups. Second, we sample N_g units from group g, for each of the sampled groups. In the absence of individual-level covariates, we can express group unconfoundedness in the following way (Rosenbaum and Rubin, 1983):

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid L_{g(i)}. \tag{2.1}$$

This assumption validates comparisons between treated and control units within groups.

As we show formally in the next section, this condition implies a different independence restriction:

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid \overline{W}_{g(i)}, \tag{2.2}$$

where $\overline{W}_g$ is the share of treated units in group g. This unconfoundedness condition allows us to combine groups with the same value for $\overline{W}_g$. This does not immediately have any empirical content. To see this, consider the special case with two units per group. This implies there are three values for $\overline{W}_g$, namely 0, 1/2, and 1. If $\overline{W}_g = 0$ or $\overline{W}_g = 1$, the combined set of groups has only control units or only treated units, so there is no basis to estimate treatment effects. If we look at a set of groups with $\overline{W}_g = 1/2$, comparing treated and control units gives us an average of within-group comparisons of treated and control units based on equation (2.2). However, within-group comparisons were validated already by equation (2.1).

Continuing with this two-units-per-group example, let M_s denote the number of groups in the sample with $\overline{W}_g = s$. Without covariates the fixed-effect estimator for τ in equation (1.1), which we denote by $\hat{\tau}^{\text{fe}}$, reduces to averaging the difference between the treated unit and the control unit over the $M_{1/2}$ groups with exactly one treated and one control unit:

$$\hat{\tau}^{\text{fe}} = \frac{1}{M_{1/2}} \sum_{g : \overline{W}_g = 1/2} \hat{\tau}_g, \quad \text{where } \hat{\tau}_g = \sum_{i : g(i) = g} W_i Y_i - \sum_{i : g(i) = g} (1 - W_i) Y_i.$$

This fixed-effect estimator $\hat{\tau}^{\text{fe}}$ does not depend on outcomes for groups with either only treated units $\overline{W}_g = 1$, or groups with only control units, $\overline{W}_g = 0$. For groups with $\overline{W}_g = 1/2$, $\hat{\tau}_g$ is the difference between a treated and a control outcome from the same group, so it is a natural, and in fact the only natural, estimator for the average effect within that group given that there are exactly two units from that group in the sample.

We can now characterize what $\hat{\tau}^{\text{fe}}$ is estimating in settings with heterogeneous treatment effects under group unconfoundedness. Let $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$ denote the population average treatment effect and define for $s \in \{0, 1/2, 1\}$ the weighted average effect

$$\tau(s) = \mathbb{E}[Y_i(1) - Y_i(0) | \overline{W}_{g(i)} = s]. \tag{2.3}$$

This is a critical, though somewhat unusual, object. $\tau(s)$ is a conditional average treatment effect, but the conditioning depends on the sampling scheme and the assignment mechanism. To illustrate the unusual nature of this object, suppose that for the same population we sampled three

units per group instead of two. That would change the values of s for which $\tau(s)$ is defined, and it would potentially change the value of $\tau(0)$ and $\tau(1)$ which are defined under both sampling schemes.

Now consider the overall average effect of the treatment. This can be expressed in terms of the $\tau(s)$ as

$$\tau = \text{pr}(\overline{W}_{g(i)} = 0)\tau(0) + \text{pr}(\overline{W}_{g(i)} = 1/2)\tau(1/2) + \text{pr}(\overline{W}_{g(i)} = 1)\tau(1).$$

Under group unconfoundedness (2.2), $\hat{\tau}^{\text{fe}}$ is unbiased for $\tau(1/2)$. If, as is likely in the heterogenous treatment effect case, $\tau(s)$ varies by s, $\tau(1/2)$ would differ from τ, and thus that $\hat{\tau}^{\text{fe}}$ estimates something different from τ.

To further illustrate this point, suppose there is a group-level variable $\overline{U}_g$ that is related to the propensity score and the outcomes in the following way:

$$\mathbb{E}[Y_i(1) - Y_i(0)|\overline{U}_{g(i)}] = \overline{U}_{g(i)}^2 \quad \text{and} \quad \text{pr}(W_i = 1|L_{g(i)} = L_g) = \overline{U}_g.$$

Furthermore, suppose $\overline{U}_g$ has the following distribution over the population of groups,

$$\text{pr}(\overline{U}_g = u) = 1/3, \ u \in \{0, 1/2, 1\}.$$

In this set-up, the overall average treatment effect is $\tau = \mathbb{E}[Y_i(1) - Y_i(0)] = E[\overline{U}_{g(i)}^2] = 5/12$. However, $\text{pr}(\overline{U}_{g(i)} = 1/2|\overline{W}_{g(i)} = 1/2) = 1$, so that $\tau(1/2) = 1/4$. Conditioning on groups with one treated and one control observation shifts the distribution of $\overline{U}_g$ from its marginal distribution to a distribution that puts more weight on values of $\overline{U}_g$ that make the observation of both a treated and control unit likely. It is easy to see that the interpretation of the fixed-effect estimator (1.1) changes if we change the sampling scheme, for example, if we sample $N_g = 3$ units for each group, or if the assignment mechanism changes.

2.2. *Covariates*

Next, we consider a setting with a single binary covariate $X_i \in \{0, 1\}$. Define for each group g and each covariate value x the average treatment effect:

$$\tau_l(x) \equiv \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x, L_{g(i)} = l].$$

This quantity is well defined for all (x, l) such that x is in the support of X_i in group L_g.

A natural generalization of the group unconfoundedness assumption in equation (2.1) is the (conditional) group unconfoundedness assumption:

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i, L_{g(i)}. \tag{2.4}$$

Similar to equation (2.2), it implies a different unconfoundedness condition that does not require within-group comparisons:

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i, \overline{W}_{g(i)}, \overline{X}_{g(i)}, \overline{WX}_{g(i)}, \tag{2.5}$$

where $\overline{W}_{g(i)}, \overline{X}_{g(i)}, \overline{WX}_{g(i)}$ are group-level averages of the corresponding variables. Condition (2.5) justifies combining groups that have identical empirical joint distributions of covariates and treatment indicators.

Focusing again on the case with two units per group, we can further unpack condition (2.5). In particular, equation (2.5) allows us to compare units for groups g with variation in the treatment and no variation in the covariate. Formally, let $\underline{W}_g$ be the pair of treatment values (W_i, W_j) for the two units i and j in group g, and let $\underline{X}_g$ be the pair of treatment values (X_i, X_j) for the two units i and j in group g. We can then define the set

$$\mathbb{B} \equiv \left\{ (\underline{W}_g, \underline{X}_g) \middle| \ \underline{W}_g \in \{(0, 1), (1, 0)\}, \underline{X}_g \in \{(0, 0), (1, 1)\} \right\},$$

and consider the average treatment effect

$$\tau_{\mathbb{B}} \equiv \mathbb{E}[\tau_{L_{g(i)}}(X_i) | (\underline{W}_{g(i)}, \underline{X}_{g(i)}) \in \mathbb{B}].$$

Similar to $\tau(1/2)$ from the previous section, $\tau_{\mathbb{B}}$ is generally different from the average treatment effect τ. It depends both on the sampling scheme and the assignment process.

The fixed-effect specification (1.1), in general, does not consistently estimate $\tau_{\mathbb{B}}$ or any other weighted average effect. To see this, observe that the OLS estimator $\hat{\tau}^{\text{fe}}$ for equation (1.1) is equal to the estimated coefficient of a different regression function, namely

$$Y_i = \alpha + W_i \tau + X_i^\top \beta + \overline{W}_{g(i)} \delta + \overline{X}_{g(i)}^\top \gamma + \eta_i. \tag{2.6}$$

The numerical equivalence, first shown in Mundlak (1978), follows from repeated applications of textbook omitted variable bias formulas. In general, least-squares estimation of regression (2.6) combines data from all groups, including those with $\overline{W}_g \in \{0, 1\}$, making $\hat{\tau}^{\text{fe}}$ inconsistent for the average effect.

The alternative version of the fixed-effect regression in representation (2.6) is important because it suggests weakening the conditional group unconfoundedness condition from equation (2.5) by dropping $\overline{WX}_{g(i)}$, leaving the condition as

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i, \overline{X}_{g(i)}, \overline{W}_{g(i)}. \tag{2.7}$$

This condition encapsulates a key feature of the fixed-effect specification that $\overline{W}_{g(i)}$ and $\overline{X}_{g(i)}$ fully capture all relevant differences between groups and that $\overline{WX}_{g(i)}$ is not required for this. Restriction (2.7) implies that we can compare treated and control units as long as they have the same values of X_i and $\overline{S}_{g(i)} \equiv (\overline{W}_{g(i)}, \overline{X}_{g(i)})$, even though they may belong to groups that have different values of $\overline{WX}_g$. With two units per group, by construction $\overline{S}_{g(i)}$ takes 9 possible values, but only 3 of them correspond to groups with variation in treatment status: $\overline{S}_{g(i)} \in \{(\frac{1}{2}, 0), (\frac{1}{2}, \frac{1}{2}), (\frac{1}{2}, 1)\}$. For other values of $\overline{S}_{g(i)}$ we either have only control or only treated units, and thus condition (2.7) is not useful (and neither is equation (2.4)).

The three remaining values of $\overline{S}_{g(i)}$ lead to conceptually different comparisons. Values $\overline{S}_{g(i)} \in \{(\frac{1}{2}, 0), (\frac{1}{2}, 1)\}$ correspond to groups where both units are identical in terms of X_i but vary in their treatment. For these groups equation (2.7) justifies within-group comparisons which were already validated by the more restrictive assumption (2.4). This leaves one remaining value, $\overline{S}_{g(i)} = (\frac{1}{2}, \frac{1}{2})$. If $\overline{S}_{g(i)} = (\frac{1}{2}, \frac{1}{2})$ then by definition the two units in the group have different values of X_i, and within-group comparisons have no causal interpretation: the pair of units in the same group are either $(W_i, X_i) = (0, 0)$ and $(W_j, X_j) = (1, 1)$, or the pair are $(W_i, X_i) = (1, 0)$ and $(W_j, X_j) = (0, 1)$. In both cases, the pairs cannot be compared directly. However, equation (2.7) justifies combining groups where $W_i = X_i$ with those where $W_i = 1 - X_i$. It is useful to consider this comparison in more detail. Suppose that one group has units with $(W_i, X_i) \in \{(0, 0), (1, 1)\}$, and another group has units with $(W_i, X_i) \in \{(0, 1), (1, 0)\}$. In neither of the

groups we can directly compare treated and control units. But under equation (2.7), we can combine the two groups and then compare a treated unit i with $(W_i, X_i) = (1, 1)$ in one group with a control unit j with $(W_j, X_j) = (0, 1)$ in another group because they satisfy the two conditions that (i) they have the same value of the covariates ($X_i = X_j = 1$) and (ii) they belong to groups with the same value for $\overline{W}_g$ and $\overline{X}_g$ (namely $\overline{W}_{g(i)} = \overline{W}_{g(j)} = 1/2$ and $\overline{X}_{g(i)} = \overline{X}_{g(j)} = 1/2$).

Define the conditional expectation of $\tau_l(X_i)$ where the expectation is over all units with $X_i = x$ and over all groups with $\overline{S}_{g(i)} = s$:

$$\tau(s, x) \equiv \mathbb{E}[\tau_{L_{g(i)}}(X_i) | \overline{S}_{g(i)} = s, X_i = x] = \mathbb{E}[Y_i(1) - Y_i(0) | \overline{S}_{g(i)} = s, X_i = x].$$

Similar to $\tau(s)$ and $\tau_{\mathbb{B}}$, this object is defined in terms of the sampling scheme and is not a fixed population characteristic. The discussion above shows that we can consistently estimate $\tau(s, x)$ for certain values of (s, x). Importantly, the linear specification (1.1) relies on comparisons that are not validated by equation (2.7) and thus does not estimate a meaningful average effect in this case.

2.3. *Discussion*

Examples from the previous sections show that we can reduce group unconfoundedness, as in equation (2.1) or equation (2.4), to conditioning on some balancing score $\overline{S}_{g(i)}$ and covariates, as in (2.2) or (2.5). Despite the lack of immediate empirical content, this reduction brings two critical insights. Both insights are about operationalizing the notion that some groups are more similar than others, which is absent in equation (2.1) or equation (2.4).

First, conditioning on an unordered discrete characteristic—group indicator—does not provide a distance metric to make the case that group g is closer to group g' or to group g''. In contrast, equation (2.2) (and similar equation (2.5)) allows for the establishment of such a metric: a group g with $\overline{W}_g = 0.10$ is closer to group g' with $\overline{W}_{g'} = 0.11$ than it is to group g'' with $\overline{W}_{g''} = 0.50$ (see Johannemann *et al.*, 2019 for an application of this idea). The second insight is even more central to the current paper. It allows us to find a middle ground between fully conditioning on the entire set of group indicators and not conditioning on these indicators at all. For example, going from equations (2.5) to (2.7) justifies some, but not all between-group comparisons. Depending on the application, one can go further and assume a stronger condition

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i, \overline{W}_{g(i)}.$$

Alternatively, if we observe more units in each group, we can relax equation (2.7) to improve the comparability of groups without reducing it fully to equation (2.4). We discuss in Section 3 whether and when it is enough to adjust for some subset of these averages.

Another issue illustrated in the previous sections is the relationship between the unconfoundedness conditions and the fixed-effects specification (1.1). In general, this regression does not estimate the average treatment effect, and in the presence of covariates, it might fail to estimate even a weighted average effect. However, under the restriction (2.7) we can go beyond the linear specifications (1.1) and consider different estimation methods. First, we can use more flexible specifications for the regression function as a function of the control variables $(X_i, \overline{W}_{g(i)}, \overline{X}_{g(i)})$ and the treatment W_i:

$$Y_i(w) = \mu(w, X_i, \overline{W}_{g(i)}, \overline{X}_{g(i)}) + \varepsilon_i(w),$$

with a parametric or non-parametric specification for $\mu(w, \cdot)$ that generalizes the linear additive form in equation (1.1). These specifications may include higher-order moments of the control variables, interactions with the treatment, or transformations of the linear index. Given estimates

of $\mu(w, \cdot)$ we can average the difference $\hat{\mu}(1, X_i, \overline{W}_{g(i)}, \overline{X}_{g(i)}) - \hat{\mu}(0, X_i, \overline{W}_{g(i)}, \overline{X}_{g(i)})$ over the sample to estimate the average treatment effect.

Second, we can model the propensity score as a function of $(X_i, \overline{W}_{g(i)}, \overline{X}_{g(i)})$:

$$e(x, \overline{w}, \overline{x}) \equiv \text{pr}(W_i = 1 | X_i = x, \overline{W}_{g(i)} = \overline{w}, \overline{X}_{g(i)} = \overline{x}).$$

Once we have estimates of the propensity score, we can use them to develop inverse propensity score weighting estimators (see Arpino and Mealli, 2011 for an application of this idea to grouped data). In particular, an attractive approach would be to use the inverse propensity score weighting or other forms of balancing in combination with a credible specification of the regression function. For example, one could use a weighted version of equation (2.6) to make the results more robust to misspecification of the regression function. Such doubly robust and balancing methods have been argued in the recent causal inference literature on unconfoundedness to be more robust to misspecification than estimators that rely solely on specifying the conditional mean of the outcome given conditioning variables and treatments (*e.g.* Robins and Rotnitzky, 1995; Zubizarreta, 2015; Chernozhukov *et al.*, 2017; Athey *et al.*, 2018).

Following this discussion, we recommend using three modifications of equation (1.1). First, choose what group characteristics one wishes to include in the analysis beyond the averages of the treatment and the covariates. Second, specify a flexible conditional mean function that allows for dependence on the selected group characteristics. Third, estimate the propensity score as a function of individuals and group characteristics and combine that with the conditional mean specification to obtain a more robust estimator for the average treatment effect.

2.4. *Related literature*

Our approach is related to classic panel data literature, for example, Mundlak (1978), Chamberlain (1984), Chamberlain (2010), and Altonji and Matzkin (2005). Similar to these papers, we develop a conditional approach, where we effectively control for heterogeneity across groups using group-level balancing scores. Our key contribution to this literature is to show how these statistics arise naturally from the design model and connect this approach to modern estimation algorithms developed for cross-sectional data.

Our results are also connected to random effects models studied in the more recent panel data literature (see Arellano and Bonhomme, 2011 for a recent review). For example, Arellano and Bonhomme (2016) explicitly model the distribution of the outcomes and assignments given the individual unobservable characteristics and achieve identification using results from nonparametric non-linear deconvolution literature (Hu and Schennach, 2008). The key conceptual and practical difference between our approach and theirs is that we do not use the outcome data to deal with unobserved heterogeneity and rely only on the distribution of treatments and covariates for identification. In this sense, we follow a tradition used in causal inference literature (Rosenbaum, 2002; Imbens and Rubin, 2015; Abadie *et al.*, 2020).

Our approach is connected to grouped fixed-effects strategies (Hahn and Moon, 2010; Bonhomme and Manresa, 2015; Bonhomme *et al.*, 2022). Similar to this literature, we propose pooling data across groups with similar aggregate statistics. There are two main differences, though. As discussed above, we do not use the outcome data, instead focusing on the assignment process. Also, our identification and inference results are valid for groups of small size, whereas statistical results in the other literature are based on approximations with large groups.

Finally, our results are related to recent literature on regression models with fixed effects (*e.g.* De Chaisemartin and d'Haultfoeuille, 2020; Goodman-Bacon, 2021; Wooldridge, 2021).

Similar to these papers, we discuss the interpretation of the least-squares estimators under heterogeneity in treatment effects and develop an estimator that is robust to the presence of such heterogeneity. Our focus, however, is quite different: instead of restricting the conditional outcome distributions, we focus on the assignment model and achieve robustness by using it to construct the estimator.

3. IDENTIFICATION, ESTIMATION, AND INFERENCE

In this section, we present the main results.

3.1. *Set-up*

We start with the sampling assumption that describes the relationship between groups and unit-level data:

Assumption 3.1 (Grouped sampling). *We randomly sample M groups out of a super-population of groups with $g \in \{1, \ldots, M\}$ being a generic group; we then randomly sample N_g units from each group g, so that $N = \sum_{g=1}^{M} N_g$ is the total sample size. $N_g \geq 2$ and is the same for all $g = 1, \ldots, M$.*

The assumption that the group size is identical for all groups is for expositional reasons only. In the Appendix, we generalize our results to allow for heterogeneity in the group sizes. The key complication is that it requires us to add the group size to the set of conditioning variables that make up the balancing score. We assume that for a generic group g we observe its label L_g. For each unit i define $g(i) \in \{1, \ldots, M\}$—the index of the group this unit belongs to. For each unit i we observe the outcomes Y_i, binary treatment indicators $W_i \in \{0, 1\}$, and covariates $X_i \in \mathbb{X}$. Assumption 3.1 allows us to view the data as M independent and identically distributed copies of random elements $(L_g, \{(W_i, X_i, Y_i)\}_{i:g(i)=g})$, with the additional independence among vectors $\{(W_i, X_i, Y_i)\}_{i:g(i)=g}$.

We interpret the observed outcomes using Neyman–Rubin potential outcomes model (Neyman, 1923/1990; Rubin, 1977; Imbens and Rubin, 2015):

$$Y_i = Y_i(1)W_i + Y_i(0)(1 - W_i), \tag{3.1}$$

which implies the absence of spillovers across units and groups. We impose restrictions on the treatment assignment process:

Assumption 3.2 (Group-level Unconfoundedness).

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i, L_{g(i)}.$$

This assumption restricts the distribution of potential outcomes and treatment indicators within a particular group. It implies that we can give a causal interpretation to comparisons of units with the same characteristics within the group.

3.2. *Identification results*

For the first identification result, we need to introduce additional notation. For each group g, define $\mathbb{P}_g$ to be the empirical measure of (W_i, X_i) in group g. Formally, for any set

$A \subseteq \mathbb{X} \times \{0, 1\}$, $\mathbb{P}_g(A)$ is defined through the following relationship:

$$\mathbb{P}_g(A) = \frac{1}{N_g} \sum_{i:g(i)=g} \mathbf{1}_{(W_i, X_i) \in A}.$$

We use this construction for our first identification result:

Proposition 1 (Unconfoundedness with empirical measure)**.** *Suppose Assumptions* 3.1 *and* 3.2 *hold. Then:*

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i, \mathbb{P}_{g(i)}.$$

For proofs of all results in this section see Appendix A. Proposition 1 states that as long as units have the same characteristics, and they come from groups identical in terms of $\mathbb{P}_g$ they are comparable. To relate this result to the discussions in the previous section, assume that there are no covariates. In this case, $\mathbb{P}_g$ reduces to $\overline{W}_g$ and we get condition (2.2). If we introduce a single binary covariate—we get condition (2.5). This result is a manifestation of a familiar balancing property of the propensity score (*e.g.* Rosenbaum and Rubin, 1983). The subpopulation with the same value for $(X_i, \mathbb{P}_g)$ are balanced: the distribution of treatments is the same for all units within such subpopulations.

However, as discussed in the previous section, the empirical relevance of this result is limited. To make it operational, we need to substitute the empirical distribution of (W_i, X_i) with a low-dimensional object. To do this, we put structure on the joint distribution of (W_i, X_i) and its relation to L_g with an exponential family representation (Cox and Hinkley, 1979).

Assumption 3.3 (Exponential family)**.** *Conditional on $L_g = l$ distribution of (W_i, X_i) belongs to an exponential family with a known sufficient statistic, that is, its conditional density has the following form:*

$$f_{W_i, X_i | L_g}(w, x | l) = h(w, x) \exp \left\{ \eta(l)^\top S(w, x) + \eta_0(l)) \right\}, \tag{3.2}$$

with potentially unknown carrier $h(\cdot)$.

Under relatively weak conditions, distributions can be well approximated by distributions in exponential families (Barron and Sheu, 1991), so as long as we do not restrict the dimension of $S(\cdot)$ this assumption is fairly weak. For example, note that if the distribution of X_i is discrete, one can immediately write the joint distribution of (W_i, X_i) within each group as an exponential family distribution with a group-specific parameter. In addition, we can approximate any distribution arbitrarily well by a discrete distribution. Of course, it is only when the dimension of $S(\cdot)$ is small relative to the number of observations that the assumption will be seen to be useful.

Assumption 3.3 places no restrictions on the distribution of the potential outcomes and thus does not restrict heterogeneity in treatment effects. At the same time, it restricts the joint distribution of (W_i, X_i), not just the conditional distribution of W_i. This is without loss of generality: if X_i is discrete and function $S(W_i, X_i)$ includes indicators for all potential values of X_i, then equation (3.2) places no restrictions on the marginal distribution of X_i given $L_{g(i)}$. For continuous X_i the same argument holds using an infinite set of indicators.

With Assumption 3.3 we can define an unobserved group-level variable that captures all the information about the distribution of (W_i, X_i):

$$\overline{U}_g \equiv \eta(L_g).$$

We use the bar notation to indicate that this is a group-level variable. Our next lemma shows that we can substitute L_g with $\overline{U}_g$ for the purpose of identification.

Lemma 1. *Suppose Assumptions* 3.1–3.3 *hold. Then:* (i)

$$\{(W_i, X_i)\}_{i:g(i)=g} \perp\!\!\!\perp L_g \mid \overline{U}_g$$

and (ii):

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i, \overline{U}_{g(i)}.$$

This result demonstrates that our problem can be converted into a traditional random effects set-up (*e.g.* Chamberlain, 1984): the group labels L_g are not important, only the group-level characteristics $\overline{U}_g$, which are functions of these labels, are. We do not know the group-specific parameters $\overline{U}_g$, nor can we estimate them consistently in settings with a small number of observations per group because of the incidental parameter problem (Neyman and Scott, 1948; Lancaster, 2000). However, and this is a key insight in the paper, it is *not* necessary to estimate them consistently because the sufficient statistic in the exponential family representation in Assumption 3.3, $\overline{S}_g \equiv \sum_{i:g(i)=g} S(W_i, X_i)/N_g$, plays the role of a balancing statistic.

Using Assumption 3.3 we state our main identification result:

Theorem 1. *Suppose Assumptions* 3.1–3.3 *hold. Then:*

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i, \overline{S}_{g(i)}.$$

Theorem 1 is related to Proposition 1, but the key difference is that Theorem 1 has empirical content as long as $\overline{S}_{g(i)}$ is less general than $\mathbb{P}_{g(i)}$. Theorem 1 reduces the high-dimensional object $\mathbb{P}_{g(i)}$ to an average $\overline{S}_{g(i)}$ and allows us to combine multiple groups. This makes cross-sectional methods developed in the last decades feasible for applications with grouped data. In practice, the importance of this point depends on the number of units per group and the dimension of $S(\cdot)$. The higher the dimension of $S(\cdot)$, the closer we are to controlling for $\mathbb{P}_{g(i)}$ and thus relying solely on comparisons within groups.

Theorem 1 becomes useful once we fix a particular low-dimensional $S(W_i, X_i)$. We assume that $S(\cdot)$ is known throughout most of the paper. This makes Assumption 3.3 restrictive and we can no longer argue that it holds for an arbitrary conditional distribution of (W_i, X_i). As we illustrate with the examples in the previous section, this is needed if we want to go beyond within-groups analysis without imposing additional assumptions on the potential outcomes. In the current empirical practice, researchers rely on equation (1.1) or, equivalently, equation (2.6), thus implicitly using $S(W_i, X_i) = (W_i, X_i)$, which can be too coarse for some applications. In particular, this choice assumes that covariates do not interact with group labels in the assignment model. We informally discuss selecting the set of balancing scores in Section 3.3.

Define propensity score:

$$e(x, s) \equiv \mathrm{pr}(W_i = 1 | X_i = x, \overline{S}_{g(i)} = s).$$

Depending on the value of the balancing score, the propensity score $e(x, s)$ might be equal to zero or one for certain values of s. For example, this happens if $\overline{W}_g$ is part of the balancing score. As a result, we cannot expect the overlap assumption (Imbens and Rubin, 2015) to hold. Instead, we are making the following assumption:

Assumption 3.4 (Known overlap). *For a known set* $\mathbb{A}$ *with* $\mathrm{pr}((X_i, \overline{S}_{g(i)}) \in \mathbb{A}) > 0$ *there exists* $\eta > 0$, *such that* (i) *for any* $(x, s) \in \mathbb{A}$ *we have* $\eta < e(x, s) < 1 - \eta$ *and* (ii) *for* $(x, s) \notin \mathbb{A}$ $e(x, s) \in \{0, 1\}$.

This assumption has two parts: the first part restricts $e(x, s)$ to be non-degenerate on a certain set and only that set. This is necessary if we want to identify treatment effects without relying on functional form assumptions. The second part is different: we assume that the set is known to a researcher. This is a generalization of the standard overlap assumption, where we assume that the set $\mathbb{A}$ is equal to the support of the covariate space, see Crump *et al.* (2009).

To provide additional intuition for Assumption 3.4, let us go back to the example from Section 2.2, where $S(W_i, X_i) = (W_i, X_i)$ and we observe two units per cluster. In that case, the propensity score is equal to either zero or one for any $\overline{S}_{g(i)} \notin \{(\frac{1}{2}, 0), (\frac{1}{2}, \frac{1}{2}), (\frac{1}{2}, 1)\}$. By construction $e(0, (\frac{1}{2}, 0)) = e(1, (\frac{1}{2}, 1)) = \frac{1}{2}$, but the values of $e(0, (\frac{1}{2}, \frac{1}{2}))$ and $e(1, (\frac{1}{2}, \frac{1}{2}))$ are not fixed, except for the restriction $e(0, (\frac{1}{2}, \frac{1}{2})) + e(1, (\frac{1}{2}, \frac{1}{2})) = 1$. As a result, if we set $\mathbb{A} \equiv \{(x, s) : (0, (\frac{1}{2}, 0)), (0, (\frac{1}{2}, \frac{1}{2})), (1, (\frac{1}{2}, \frac{1}{2})), (1, (\frac{1}{2}, 1))\}$, then this amounts to assuming $\eta < e(0, (\frac{1}{2}, \frac{1}{2})) < 1 - \eta$.

We can now define our target estimands. We focus on two of them; the first one is similar to the objects considered in the previous section:

$$\tau_{\mathbb{A}} \equiv \mathbb{E}[(Y_i(1) - Y_i(0))|(X_i, \overline{S}_{g(i)}) \in \mathbb{A}]. \tag{3.3}$$

The second object is a sample group-weighted version of equation (3.3):

$$\tilde{\tau}_{\mathbb{A}} \equiv \frac{\sum_{i=1}^{N} \mathbf{1}_{(X_i, \overline{S}_{g(i)}) \in \mathbb{A}} \mathbb{E}[Y_i(1) - Y_i(0)|X_i, \overline{S}_{g(i)}]}{\sum_{i=1}^{N} \mathbf{1}_{(X_i, \overline{S}_{g(i)}) \in \mathbb{A}}}.$$

Theorem 1 together with Assumption 3.4 implies that $\tau_{\mathbb{A}}$ and $\tilde{\tau}_{\mathbb{A}}$ are identified. We formally state this result in the next corollary.

Corollary 1. *Suppose Assumptions 3.1–3.4 hold. Then $\tau_{\mathbb{A}}$ is identified.*

Proof. The results follow from the identity:

$$\mathbb{E}[(Y_i(1) - Y_i(0))|(X_i, \overline{S}_{g(i)}) \in \mathbb{A}]$$
$$= \mathbb{E}\left[\left(\frac{W_i}{e(X_i, \overline{S}_{g(i)})} - \frac{1 - W_i}{1 - e(X_i, \overline{S}_{g(i)})}\right) Y_i \,\middle|\, (X_i, \overline{S}_{g(i)}) \in \mathbb{A}\right],$$

and the fact that $e(X_i, \overline{S}_{g(i)})$ being a conditional expectation of a binary variable is identified under Assumption 3.1. $\qquad\square$

In the remainder of the paper, we primarily focus on the properties of estimators for $\tilde{\tau}_{\mathbb{A}}$. Such estimators can also be used to estimate $\tau_{\mathbb{A}}$, but the variances are different. Our focus on $\tilde{\tau}_{\mathbb{A}}$ is analogous to the focus on the sample average treatment effect in the experimental literature.

3.3. *Discussion*

Our key result—Theorem 1—is related to some well-known identification results. For example, in Altonji and Matzkin (2005) a key assumption (Assumption 2.1) requires that there is an observed variable Z_i such that conditioning on Z_i renders the covariate of interest (the treatment in our case) exogenous. In our setting, the role of this conditioning variable is played by the balancing score $\overline{S}_{g(i)}$. We show how this property can arise from assumptions on the joint distribution of the treatment and the other covariates, and how we can make this more plausible by increasing the dimension of balancing scores.

Assumption 3.3 plays the key role in Theorem 1 and in the previous section we argued that it can be viewed as a natural approximation for the conditional distribution of covariates and treatment indicators. In addition to this statistical justification, it arises in a structural economic model. To this end, consider a set-up where each unit is characterized by (W_i, X_i) and then chooses group membership l out of set L. For example, W_i can be a scholarship or a voucher, of a student with observed characteristics X_i, and $L_{g(i)}$ can be a college or a school this individual has chosen. In this case, $\mathbb{E}[Y_i(1) - Y_i(0)|L_{g(i)}]$ is an average direct effect of treatment unmediated by the choice $L_{g(i)}$.

Suppose the following function is the indirect utility for individual i from group l:

$$V_i(l) = V(W_i, X_i, l) + \epsilon_i(l),$$

where $\epsilon_i(l)$ are i.i.d. extreme-value type-1 shocks. In this case, the probability of selecting a particular option g has the following form:

$$\mathbb{E}[\mathbf{1}_{L_{g(i)}=l}|W_i, X_i] = \frac{\exp(V(W_i, X_i, l))}{\sum_{k \in L} \exp(V(W_i, X_i, k))}.$$

This restricts the joint distribution of $(L_{g(i)}, W_i, X_i)$:

$$f_{W_i, X_i, L_{g(i)}}(w, x, u) = f(w, x) \exp\left(V(w, x, l)\right),$$

and thus as long as $V(w, x, u) = \eta^\top(l) S(w, x) + \eta_0(l)$ Assumption 3.3 is satisfied. This discussion shows that exponential family restriction is a natural approximation in a structural model of choice.

In practice, balancing scores are unknown and need to be selected by researchers. This choice should reflect domain knowledge and a full treatment of this problem is an open question. However, we provide a suggestion for systematically selecting balancing scores in the case where we have a large set of potential balancing scores that includes all the relevant ones but also some that are not relevant. Intuitively we would like a selection procedure to select more balancing scores in settings where we have a lot of units per group and if the distributions vary substantially across groups. Assumption 3.3 implies that the conditional probability that a unit in the sample is from group l, conditional on (W_i, X_i) and conditional on the set of $L_1, \ldots, L_M$, has a multinomial logit form:

$$\mathrm{pr}(L_{g(i)} = l|W_i, X_i, L_1, \ldots, L_M) = \frac{\exp(\eta_0(L_g) + \eta(L_g)^\top S(W_i, X_i))}{\sum_{k=1}^M \exp(\eta_0(L_k) + \eta(L_k)^\top S(W_i, X_i))}.$$

Hence, the problem of selecting the balancing scores is similar to the problem of selecting covariates in a multinomial logistic regression model. Given a large set of potential balancing scores, we can use standard covariate selection methods, such as least absolute shrinkage and selection operator (LASSO, Tibshirani, 1996) to select a sparse set of relevant ones. In practice, balancing scores are likely to be correlated, and other selection methods such as adaptive elastic net (Zou and Zhang, 2009) might be more appropriate. See also De Paula *et al.* (2019) for the application of this algorithm to the identification of unobserved network structure with panel data.

3.4. *Estimation and inference*

This section collects several inference results for a class of semiparametric estimators of $\tau_{\mathbb{A}}$ and $\tilde{\tau}_{\mathbb{A}}$. All proofs can be found in Appendix B. For further use, we use the following notation for the conditional mean, propensity score, and residuals:

$$\begin{cases} \mu(W_i, X_i, \overline{S}_{g(i)}) \equiv \mathbb{E}[Y_i | W_i, X_i, \overline{S}_{g(i)}], \\ e(X_i, \overline{S}_{g(i)}) \equiv \mathrm{pr}(W_i = 1 | X_i, \overline{S}_{g(i)}), \\ \varepsilon_i(w) \equiv Y_i(w) - \mu(w, X_i, \overline{S}_{g(i)}). \end{cases}$$

These expectations are defined using Assumption 3.1, which determines the distribution of $\overline{S}_{g(i)}$. In particular, as we discussed in Section 2, the distribution of $\overline{S}_{g(i)}$ changes with the number of observed units in each group. We restrict moments of the corresponding errors:

Assumption 3.5 (Moment conditions). *For $w \in \{0, 1\}$*

$$\mathbb{E}[\varepsilon_i^2(w) | X_i, \overline{S}_g] < K \ a.s., \quad \mathbb{E}[\varepsilon_i^4(w)] < \infty.$$

These conditions hold for bounded potential outcomes, which is the leading case in practice. In the case of unbounded outcomes, the first condition imposes a restriction on the degree of heteroscedasticity.

We will use $\hat{\mu}_g(\cdot)$ and $\hat{e}_g(\cdot)$ for generic estimators of $\mu(\cdot)$ and $e(\cdot)$. Subscript g is used to allow for cross-fitting (*e.g.* Chernozhukov *et al.*, 2018a); we assume that researchers use K-fold group-level cross-fitting. In particular, researchers split the observed M groups into K random subsamples, and for each group g construct estimators $\hat{\mu}_g(\cdot), \hat{e}_g(\cdot)$ using the data from $K - 1$ subsamples that do not contain group g. Define indicator variable $A_i \equiv \mathbf{1}_{(X_i, \overline{S}_{g(i)}) \in \mathbb{A}}$ and let $\overline{A} \equiv \frac{1}{M} \sum_{g=1}^{M} \frac{1}{N_g} \sum_{i:g(i)=g} A_i$ be the estimate of the share of units for which we have overlap. We assume the generic estimators $\hat{e}_g$ and $\hat{\mu}_g$ satisfy several high-level consistency properties standard in the program evaluation literature:

Assumption 3.6 (High-level conditions). *The following conditions are satisfied for $\hat{e}_g(\cdot)$ and $\hat{\mu}_g(\cdot)$ as M goes to infinity: (i) $\frac{\eta}{2} < \hat{e}_g(X_i, \overline{S}_{g(i)}) < 1 - \frac{\eta}{2}$ on $\mathbb{A}$, (ii) mean-consistency, where $w \in \{0, 1\}$:*

$$\mathbb{E}\left[\frac{1}{M} \sum_{g=1}^{M} \frac{1}{N_g} \sum_{i:g(i)=g} A_i (e(X_i, \overline{S}_{g(i)}) - \hat{e}_g(X_i, \overline{S}_g))^2 \right] = o(1),$$

$$\mathbb{E}\left[\frac{1}{M} \sum_{g=1}^{M} \frac{1}{N_g} \sum_{i:g(i)=g} A_i (\mu(w, X_i, \overline{S}_{g(i)}) - \hat{\mu}_g(w, X_i, \overline{S}_{g(i)}))^2 \right] = o(1),$$

(iii) and product rate condition:

$$\frac{1}{M} \sum_{g=1}^{M} \frac{1}{N_g} \sum_{i:g(i)=g} A_i \left(\hat{\mu}_g(W_i, X_i, \overline{S}_{g(i)}) - \mu(W_i, X_i, \overline{S}_{g(i)}) \right)^2$$

$$\times \frac{1}{M} \sum_{g=1}^{M} \frac{1}{N_g} \sum_{i:g(i)=g} A_i \left(\hat{e}_g(X_i, \overline{S}_{g(i)}) - e(X_i, \overline{S}_{g(i)}) \right)^2 = o_p\left(\frac{1}{M} \right).$$

Under Assumptions 3.1, 3.4, and bounded outcomes, these conditions are satisfied if covariates are discrete. To see this note that (i) and (ii) are weak consistency restrictions, while each term in (iii) is $O_p(\frac{1}{M})$. Under appropriate smoothness restrictions (see Hansen, 2022 for a textbook treatment), these conditions hold for conventional kernel estimators. The use of cross-fitting expands the set of available estimators, allowing researchers to use machine learning estimators (see Newey and Robins, 2018 for details).

For arbitrary functions $m(\cdot)$, $p(\cdot)$, define the following functional (*e.g.* Robins and Rotnitzky, 1995; Hahn, 1998; Chernozhukov *et al.*, 2018a):

$$\psi(y, w, x, s, m(\cdot), p(\cdot)) \equiv m(1, x, s) - m(0, x, s)$$
$$+ \left(\frac{w}{p(x, s)} - \frac{1 - w}{1 - p(x, s)} \right) (y - m(w, x, s),$$

and its group-level aggregate:

$$\rho_g(m(\cdot), p(\cdot)) \equiv \frac{1}{N_g} \sum_{i:g(i)=g} A_i \psi(Y_i, W_i, X_i, \overline{S}_{g(i)}, m(W_i, X_i, \overline{S}_i), p(X_i, \overline{S}_i)).$$

We use $\rho_g(\cdot)$, estimators $\hat{\mu}_g(\cdot)$, $\hat{e}_g(\cdot)$, and $\overline{A}$ to define our proposed GME:

$$\hat{\tau}_{\text{GME}} \equiv \frac{\sum_{g=1}^M \rho_g(\hat{\mu}_g, \hat{e}_g)}{\overline{A}}. \tag{3.4}$$

We also define the influence function:

$$\xi_g \equiv \sum_{i:g(i)=g} \frac{A_i}{N_g} \left(\frac{W_i}{e(X_i, \overline{S}_g)} - \frac{1 - W_i}{1 - e(X_i, \overline{S}_g)} \right) (Y_i - \mu(W_i, X_i, \overline{S}_g)).$$

Our next result shows the properties of $\hat{\tau}_{\text{GME}}$:

Theorem 2 (Inference). *Suppose Assumptions 3.1–3.5 hold. Then as M goes to infinity (i):*

$$\hat{\tau}_{\text{GME}} = \tilde{\tau}_{\mathbb{A}} + o_p(1) = \tau_{\mathbb{A}} + o_p(1), \tag{3.5}$$

and (ii):

$$\sqrt{M}(\hat{\tau}_{\text{GME}} - \tilde{\tau}_{\mathbb{A}}) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}), \quad \text{where } \mathbb{V} = \frac{\mathbb{E}\left[\xi_g^2\right]}{\mathbb{E}^2[A_i]}.$$

This result allows us to conduct inference for $\tilde{\tau}_{\mathbb{A}}$—the sample group-weighted version of our estimand. A similar result holds for $\tau_{\mathbb{A}}$ with a larger variance that reflects additional randomness in $\tilde{\tau}_{\mathbb{A}}$. To estimate the asymptotic variance $\mathbb{V}$, we define the estimated version of ξ_g:

$$\hat{\xi}_g \equiv \sum_{i:g(i)=g} \frac{A_i}{N_g} \left(\left(\frac{W_i}{\hat{e}_g(X_i, \overline{S}_g)} - \frac{1 - W_i}{1 - \hat{e}_g(X_i, \overline{S}_g)} \right) (Y_i - \hat{\mu}_g(W_i, X_i, \overline{S}_g)) \right),$$

and use it to construct the estimator:

$$\hat{\mathbb{V}} := \frac{1}{\overline{A}^2} \frac{1}{M} \sum_{g=1}^M \left(\hat{\xi}_g - \frac{1}{M} \sum_{g'=1}^M \hat{\xi}_{g'} \right)^2. \tag{3.6}$$

Next result shows that this estimator is consistent:

Theorem 3 (Variance consistency). *Suppose Assumptions of Theorem* 2 *hold. Then the variance estimator is consistent:*

$$\hat{\mathbb{V}} = \mathbb{V} + o_p(1).$$

Together Theorems 2 and 3 imply that standard confidence intervals are asymptotically valid. In particular, if z_α is the α-quantile of the standard normal distribution then asymptotically

$$\tilde{\tau}_{\mathbb{A}} \in \hat{\tau}_{\mathrm{GME}} \pm \sqrt{\frac{\hat{\mathbb{V}}}{M}} z_{\alpha/2},$$

with probability $(1 - \alpha)$.

4. EXTENSIONS

In this section, we discuss three extensions of the ideas introduced in this paper.

4.1. *Quantile treatment effects*

Theorem 1 states that conditional on the covariates and the balancing scores we have the unconfoundedness condition:

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i, \overline{S}_{g(i)}.$$

This implies that we can study estimation of effects other than average treatment effects. This is important in applications where we want to estimate, distributional effects controlling for group-level unobserved heterogeneity.

In particular, for any bounded function $f : \mathbb{R} \to \mathbb{R}$ we can estimate $\mathbb{E}[f(Y_i(w))]$ using the inverse propensity score:

$$\mathbb{E}[f(Y_i(w)] = \mathbb{E}\left[\frac{\mathbf{1}_{W_i=w} f(Y_i)}{e(X_i, \overline{S}_{g(i)})}\right].$$

This allows us to identify quantile treatment effects of the type introduced by Leo Lehmann and D'Abrera (2006). If we are interested in qth quantile of the distribution of $Y_i(w)$ then (under appropriate continuity) we can identify it as a solution to the following problem:

$$c(q) : \mathbb{E}\left[\frac{\mathbf{1}_{W_i=w}\mathbf{1}_{Y_i \leq c(q)}}{e(X_i, \overline{S}_{g(i)})}\right] = q.$$

For the standard case under unconfoundedness, Firpo (2007) has developed effective estimation methods that can be adapted to this case.

4.2. *Beyond exponential families*

Modelling the conditional distribution of (W_i, X_i) given $\overline{U}_g$ using an exponential family is natural for the purposes of this paper. Nevertheless, in some applications, other families can be more appropriate. In particular, another operational choice is a discrete mixture. Assume that $\overline{U}_{g(i)}$ can take a finite number of values $\{u_1, \ldots, u_p\}$ with probabilities $\pi_1, \ldots, \pi_p$ and the conditional

distribution of W_i, X_i given $\overline{U}_{g(i)}$ is given by $f(w, x|u)$. Collect all the data that we observe for group g in the following tuple:

$$\mathcal{D}_g \equiv ((X_1, W_1), \ldots, (X_{N_g}, W_{N_g})).$$

The marginal distribution of this object is given by the following expression:

$$f_{\mathcal{D}_g}(x_1, w_1, \ldots, x_{N_g}, w_{N_g}) = \sum_{k=1}^{p} \prod_{j=1}^{N_g} f(x_j, w_j|k)\pi_k.$$

This implies that the conditional distribution of $\overline{U}_g$ given $\mathcal{D}_g$ has the following form:

$$\pi(\overline{U}_g = k|x_1, w_1, \ldots, x_{N_g}, w_{N_g}) = \frac{\prod_{j=1}^{N_g} f(x_j, w_j|k)\pi_k}{\sum_{k=1}^{p} \prod_{j=1}^{N_g} f(x_j, w_j|k)\pi_k}.$$

Define $\overline{S}_g \equiv (\pi(\overline{U}_g = 1|\mathcal{D}_g), \ldots, \pi(\overline{U}_g = p|\mathcal{D}_g))$ and observe that as long as assignment is unconfounded given $(X_i, \overline{U}_{g(i)})$ we have the following:

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i, \overline{S}_{g(i)}.$$

Recent results (*e.g.* Allman *et al.*, 2009) show that $(\pi_1, \ldots, \pi_p)$ and $f(w, x|u)$ are non-parametrically identified under quite general assumptions, and Bonhomme *et al.* (2016) provides a way of estimating these objects. Using these methods, we can construct $\hat{\overline{S}}_g$ and use it as a balancing score.

4.3. *Beyond binary treatments*

So far, we have discussed only the case of binary treatments. Applications with non-binary treatments are also common in empirical work, and our strategy has a natural extension for this case. Following most of the applied work, we consider a linear model for the potential outcomes:

$$Y_i(w) = \alpha_i + \tau_i w,$$

where w does not have to be binary. Here heterogeneity in (α_i, τ_i) can be related to group labels $g(i)$ and covariates X_i. Observed outcomes are defined as values of this function at the observed treatments:

$$Y_i = Y_i(W_i).$$

We focus on this model for its simplicity, an extension to a more general non-linear model can be designed in the same way.

Our assumptions remain the same and imply the following unconfoundedness restriction:

$$(\alpha_i, \tau_i) \perp\!\!\!\perp W_i|X_i, \overline{S}_{g(i)}.$$

Using this restriction, we can consider a general double robust strategy, where we estimate the model for W_i and Y_i using X_i, $\overline{S}_{g(i)}$ as regressors (*e.g.* Chernozhukov *et al.*, 2018a). In particular, let $\hat{m}_{w,i,g} \equiv \hat{m}_{w,g}(X_i, \overline{S}_{g(i)})$, and $\hat{m}_{y,i,g} \equiv \hat{m}_{y,g}(X_i, \overline{S}_{g(i)})$ be the corresponding estimators (with cross-fitting) for the conditional means of W_i and Y_i. Define the residuals:

$$\tilde{Y}_i \equiv Y_i - \hat{m}_{y,i,g}, \quad \tilde{W}_i \equiv W_i - \hat{m}_{w,i,g},$$

and consider the following regression:

$$\tilde{Y}_i = \alpha + \tau \tilde{W}_i + u_i. \tag{4.1}$$

Let $\hat{\tau}$ be the OLS estimator for τ. Using standard techniques one can show that τ is a weighted average of τ_i, where the weights are proportional to the variance of W_i conditional on X_i, $\overline{S}_{g(i)}$.

Regression (4.1) is a direct generalization of the strategy based on equation (1.1). To see this observe that representation (2.6) does not rely on W_i being binary and can be seen as a particular implementation of equation (4.1). Given the wide use of equation (1.1) in applied work, we view equation (4.1) as a natural extension of the standard regression with fixed effects with non-binary treatments.

This discussion demonstrates that, practically, for our strategy, there are few differences between binary and non-binary treatments. Of course, the key restriction here is Assumption 3.3 that models the distribution of (W_i, X_i) conditional on $L_{g(i)}$. With binary treatments, the group average $\overline{W}_g$ is the only function of W_i we can consider. Even with non-binary treatments, $\overline{W}_g$ is the only function that is effectively used in equation (1.1). The choice set is much larger with general treatments, and one can use other moments, for example, $\overline{W^2}$. The economic model discussed in Section 3.3 motivates using such objects and suggests a general way of building a balancing score.

5. ILLUSTRATIONS

5.1. *Empirical example*

To illustrate our method, we use the data from Lacetera *et al.* (2012a), (2012b), (2023) which focuses on the effects of various economic incentives on blood donations. To answer this question, the authors leverage a dataset on blood drives—visits to particular locations—conducted by the American Red Cross (ARC). For each drive, the authors have information on various outcomes, such as the number of present potential donors, blood units collected, and the number of people turned down. Each blood drive has multiple characteristics, such as the time of the year when it was conducted, its host, weather conditions on that day, and the way it was organized. Additionally, the data includes whether a reward for donation was offered, its type, and cost.

We observe multiple drives with the same host with variation in incentives over them. As a result, the dataset has a grouped structure with unit-level variation in outcomes and treatments: each location defines a group g, with a blood drive to this location being a unit-level observation i. We use $W_i \in \{0, 1\}$ for the absence or presence of economic incentive, covariates X_i to describe the date of the drive and weather conditions on that day and Y_i for the outcome of interest—the number of collected blood units.

Assumption 3.2 is natural in this set-up: appealing to institutional knowledge, the authors argue that conditional on the drive characteristics (date), the assignment of incentives is close to being random. Still, the probability of treatment can vary systematically over hosts, and thus conditioning on them—$L_{g(i)}$ in our notation—is crucial to claim unconfoundedness. We can expect hosts to differ in their fundamentals—a potential number of donors and their elasticity with respect to economic incentives, relationship with ARC managers, and organizational abilities—and U_g captures these characteristics. Importantly, proximity in U_g space might not necessarily be connected to the proximity in observed geographic locations. Our approach allows researchers to capture this unobserved proximity through balancing scores.

Finally, we need to specify the set of balancing scores $S(W_i, X_i)$ that would justify Assumption 3.3. In the paper, the authors employ specification (1.1) which is equivalent to (2.6), and

TABLE 1

Empirical results

	Simple	Fixed effects	IPW	Doubly robust
Estimate	5.27	4.12	2.79	3.42
Standard error (S.E.)	0.62	0.32	0.64	0.44

Notes: $M = 2491$, $N = 14,029$, $\overline{A} = 0.45$. The first column corresponds to the OLS estimator without fixed effects, and the second one—to the OLS estimator with fixed effects. The third column corresponds to the IPW estimator using the propensity score estimated by random forest, restricted to groups where the estimated score lies in $[0.05, 0.95]$. The final column corresponds to the doubly robust estimator with the same propensity score model. The standard errors are robust to heteroscedasticity and within-group correlation. This simulation is conducted using x86-64 architecture. The results we get using Apple M1 Pro chips differ slightly. Using the latter, we get for the IPW estimator 2.58 with an S.E. of 0.65, and for the doubly robust estimator 3.50 an S.E. of 0.45. These results are reported as a part of the replication package.

thus implicitly use $S(W_i, X_i) = (W_i, X_i)$. Following their analysis, we include X_i and W_i but also use interactions of functions of X_i—indicators for a particular season—with W_i. If specific locations have strong seasonal patterns, we can expect ARC managers to react by adjusting incentives. Interactions then help us to control for this additional selection channel.

We start with $M = 2491$ unique hosts $N = 14029$ blood drives. We use temperature, amount of rain and snow, indicators for weekends, and seasons as characteristics X_i. For the interactions with W_i we use season indicators. The dimension of the balancing score is thus equal to 11. Using constructed $(X_i, \overline{S}_{g(i)})$, we estimate a propensity score model by random forest (Breiman, 2001) and keep units with estimated probability in $[0.05, 0.95]$. By being very flexible, random forest prediction allows us to drop units from groups without variation in treatment or those where treatment is perfectly predictable by balancing scores. We use this to define the set $\mathbb{A}$, with $\overline{A} = 0.45$. We then use linear regression to estimate conditional mean $\mu(W_i, X_i, \overline{S}_{g(i)})$, which is equivalent to (1.1). Given the relatively small dimension of covariates, we do not use cross-fitting.

We report our results in Table 1, where we compare four different estimators: the OLS estimator for (1.1) without group-level fixed effects, the fixed-effects estimator $\hat{\tau}^{\text{fe}}$, an inverse propensity weights (IPW) estimator, and, finally, $\hat{\tau}_{\text{GME}}$. The IPW estimator is defined analogously to the doubly robust estimator (3.5) but uses $\hat{\mu}_g(\cdot) \equiv 0$. In all cases, we report heteroscedasticity-robust standard errors corrected for within-group correlation.

The results are similar across specifications, with the doubly robust estimator lying between the fixed-effects estimator and the IPW estimator. Balancing scores play an important role in predicting the treatment indicators; see Figure C.1 in Appendix C. There are also efficiency gains between the IPW and the doubly robust estimator, with $\hat{\tau}_{\text{GME}}$ having a smaller variance.

Similar to the original estimator used in the paper, our estimator $\hat{\tau}_{\text{GME}}$ relies on (3.1). In particular, we do not allow for spillovers across groups. In Lacetera *et al.* (2012a), (2012b), (2023) the authors argue that this assumption might be too restrictive since the supply of blood in neighbouring locations can be interdependent. A possible solution for this problem is to aggregate the data at a higher geographical level and proceed as before, but it would be interesting to develop an alternative approach that uses the original data. Since this issue is orthogonal to the subject of the current paper, we do not attempt this development.

5.2. *Monte Carlo experiments*

We base our simulation results on the dataset described above. In particular, we estimate a separate, host-level propensity score logit model for each location with four regressors: intercept and three seasonal indicators. We drop the hosts where these regressors are collinear, which leaves

TABLE 2
Simulation results

	Linear fixed effects	Doubly robust
Bias	0.79	0.38
RMSE	0.82	0.66

Notes: Sample size $M = 1200$, 100 simulations. The first column corresponds to the standard fixed-effects estimator, and the second to the doubly robust estimator. We report average bias over simulations and RMSE.

us with 616 hosts. We then combine the estimated coefficients into 20 groups using the K-means algorithm. This produces a model for $e(X_i, \overline{U}_{g(i)})$, where four-dimensional $\overline{U}_{g(i)}$ represents the estimated coefficients for one of the 20 groups. We sample $M = 1200$ hosts (with replacement) out of the original ones, keeping all characteristics of the blood drives and their total number fixed, and generate treatment assignments for each host and blood drive using constructed $e(X_i, \overline{U}_{g(i)})$. We keep observed outcomes, which is equivalent to assuming a zero treatment effect for each unit. As a result, this simulation has two sources of randomness: sampled hosts and assignments, while both covariates and outcomes are kept fixed. In particular, this means that the simulation neither satisfies Assumption 3.3 with a simple balancing score nor the regression model (1.1), thus making the comparison between the two methods ambiguous ex-ante.

We simulate this model 100 times, and for each simulation, we compute two estimators: the linear regression (1.1) and the doubly robust estimator (3.4). In the regression, we use all available covariates (the same as in the empirical exercise of the previous section), and for the propensity score, we use the standard implementation of random forest with 11-dimensional balancing scores (the same as before). We keep observations with estimated propensity scores between 0.1 and 0.9, treating this set as $\mathbb{A}$. We use equation (3.6) to estimate the asymptotic variance. We report the results in Table 2. One can see that the doubly robust estimator outperforms the linear regression both in terms of bias and root-mean-square error (RMSE). The bias, however, is not negligible, which is partly a manifestation of misspecification and partly a sample size issue. The average over simulations of the estimated standard errors of $\hat{\tau}_{\text{GME}}$ is equal to 0.58, while the standard deviation of $\hat{\tau}_{\text{GME}}$ over simulations is equal to 0.54. This suggests that the variance estimator (3.6) performs well in this simulation.

These results illustrate that our estimator, while not ideal, outperforms the conventional benchmark in the realistic data-driven simulation. We use standard off-the-shelf prediction methods widely available in various statistical packages to construct the estimator. More targeted balancing estimators analogous to those available for cross-sectional data (*e.g.* Athey *et al.*, 2018; Chernozhukov *et al.*, 2018b; Hirshberg and Wager, 2021) might achieve even better results.

6. CONCLUSION

In this work, we proposed a new approach to identification and estimation in observational studies with unobserved group-level heterogeneity. The identification argument is based on the combination of random effects and exponential family assumptions. Given this structure, we can identify a specific average treatment effect even in cases where the observed number of units per group is small. From the operational point of view, our approach allows researchers to utilize all the recently developed machinery from the standard observational studies. In particular, we generalize the doubly robust estimator and prove its consistency and asymptotic normality under common high-level assumptions.

APPENDICES

A. Identification results

In the main text, we assumed that N_g is constant, but our results hold more generally. Formally, redefine $\overline{U}_g \equiv (\eta(L_g), N_g)$, and $\overline{S}_g \equiv (\sum_{i:g(i)=g} S(W_i, X_i)/N_g, N_g)$, use $\overline{S}_{1g} \equiv \sum_{i:g(i)=g} S(W_i, X_i)/N_g$, and assume that $N_g = N(L_g) \geq 2$ for some function N. We then prove the results using this definition. In the main text, there is no need to include N_g in $\overline{S}_g$ as it is constant by Assumption 3.1.

Proof of Proposition 1. Since distribution of (W_i, X_i) conditional on $\mathbb{P}_{g(i)}$ and $L_{g(i)}$ is equal to $\mathbb{P}_{g(i)}$ and thus does not depend on $L_{g(i)}$, we have a conditional independence restriction:

$$(W_i, X_i) \perp\!\!\!\perp L_{g(i)} | \mathbb{P}_{g(i)}. \tag{1.1}$$

To prove the result, we first show the following condition:

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp W_i | X_i, \mathbb{P}_{g(i)}, L_{g(i)}. \tag{1.2}$$

Fix an arbitrary bounded function $f(y_1, y_0)$ and consider the following equalities:

$$\mathbb{E}[f(Y_i(1), Y_i(0)) | W_i, X_i, \mathbb{P}_{g(i)}, L_{g(i)}] = \mathbb{E}[f(Y_i(1), Y_i(0)) | W_i, X_i, L_{g(i)}]$$
$$= \mathbb{E}[f(Y_i(1), Y_i(0)) | X_i, L_{g(i)}], \tag{1.3}$$

where the first one follows from Assumption 3.1 and the fact that $N_g = N(L_g)$, and the second—from Assumption 3.2. We thus have

$$\mathbb{E}[W_i | Y_i(1), Y_i(0), X_i, \mathbb{P}_{g(i)}]$$
$$= \mathbb{E}[\mathbb{E}[W_i | Y_i(1), Y_i(0), X_i, \mathbb{P}_{g(i)}, L_{g(i)}] | Y_i(1), Y_i(0), X_i, \mathbb{P}_{g(i)}]$$
$$= \mathbb{E}[\mathbb{E}[W_i | X_i, \mathbb{P}_{g(i)}, L_{g(i)}] | Y_i(1), Y_i(0), X_i, \mathbb{P}_{g(i)}]$$
$$= \mathbb{E}[\mathbb{E}[W_i | X_i, \mathbb{P}_{g(i)}] | Y_i(1), Y_i(0), X_i, \mathbb{P}_{g(i)}] = \mathbb{E}[W_i | X_i, \mathbb{P}_{g(i)}], \tag{1.4}$$

where the second equality follows from equation (1.2), and the third one—from equation (1.1), since it implies that $W_i \perp\!\!\!\perp L_{g(i)} | X_i, \mathbb{P}_{g(i)}$. Since W_i is binary, equation (1.4) implies the result.[1]   □

Proof of Lemma 1. The first part of the lemma follows by definition of $\overline{U}_{g(i)}$ and Assumptions 3.1, 3.3. To demonstrate this, we view $(\{W_i, X_i\}_{i \in g}, N_g)$ as an element of $(\{0, 1\} \times \mathbb{X})^\infty \times \mathbb{N}$, where $\mathbb{N}$ is the set of natural numbers and fix an arbitrary bounded function $f_0 \in B((\{0, 1\} \times \mathbb{X})^\infty \times \mathbb{N})$, where $B((\{0, 1\} \times \mathbb{X})^\infty \times \mathbb{N})$ is a set of bounded functions defined on $(\{0, 1\} \times \mathbb{X})^\infty \times \mathbb{N}$.[2] For this function f_0, we define the following quantity (we use the fact that $\overline{U}_g$ is a deterministic function of L_g):

$$v(f_0, L_g, \overline{U}_g) \equiv \mathbb{E}\left[f_0\left(\{W_i, X_i\}_{g(i)=g}, N_g\right) | L_g, \overline{U}_g\right]$$
$$= \exp^{N_g}\left(\eta_0(L_g)\right) \int f_0\left(\{w_i, x_i\}_{i=1}^{N_g}, N_g\right) \prod_{i=1}^{N_g} h(w_i, x_i)$$
$$\times \exp^{N_g}\left(\overline{U}_g^\top \frac{1}{N_g} \sum_{i=1}^{N_g} S(w_i, x_i)\right) \prod_{i=1}^{N_g} (d\mu(w_i, x_i))$$
$$\equiv \exp^{N_g}\left(\eta_0(L_g)\right) v_0(f_0, \overline{U}_g). \tag{1.5}$$

---

1. The proof for a non-binary case follows by applying the same chain of equalities to a bounded function of W_i.
2. Formally, we define a mapping that takes an element $(\{W_i, X_i\}_{i \in g}, N_g)$ and maps it into $(\{W_i, X_i\}_{i \in g} \times \{0, x_0\}^\infty, N_g)$, where x_0 is a fixed element of $\mathbb{X}$.

Since $\upsilon(1, L_g, \overline{U}_g) = 1$ and $\frac{\upsilon(f_0, L_g, \overline{U}_g)}{\upsilon(1, L_g, \overline{U}_g)} = \frac{\upsilon_0(f_0, \overline{U}_g)}{\upsilon_0(1, \overline{U}_g)}$ it follows that $\upsilon(f_0, L_g, \overline{U}_g)$ does not depend on L_g proving the first part. The following equalities imply the second part:

$$
\begin{aligned}
\mathbb{E}[W_i|Y_i(1), Y_i(0), X_i, \overline{U}_{g(i)}] &= \mathbb{E}[\mathbb{E}[W_i|Y_i(1), Y_i(0), L_{g(i)}, X_i, \overline{U}_{g(i)}]|Y_i(1), Y_i(0), X_i, \overline{U}_{g(i)}] \\
&= \mathbb{E}[\mathbb{E}[W_i|L_{g(i)}, X_i, \overline{U}_{g(i)}]|Y_i(1), Y_i(0), X_i, \overline{U}_{g(i)}] \\
&= \mathbb{E}[\mathbb{E}[W_i|X_i, \overline{U}_{g(i)}]|Y_i(1), Y_i(0), X_i, \overline{U}_{g(i)}] = \mathbb{E}[W_i|X_i, \overline{U}_{g(i)}], \qquad (1.6)
\end{aligned}
$$

where the second equality follows from Assumption 3.2, and the third inequality follows from the first part of the lemma, thus proving the result.   □

Proof of Theorem 1. Assumptions 3.1 and 3.3 imply the following:

$$
\overline{U}_g \perp\!\!\!\perp \{W_i, X_i\}_{i:g(i)=g}|\overline{S}_g. \qquad (1.7)
$$

Formally, using the same approach as in the proof of Lemma 1 and its first part, we get the following:

$$
\begin{aligned}
\tilde{\upsilon}(f_0, \overline{S}_g, \overline{U}_g) &\equiv \mathbb{E}\left[f_0\left(\{W_i, X_i\}_{g(i)=g}, N_g\right)|\overline{S}_g, \overline{U}_g\right] = \mathbb{E}\left[f_0\left(\{W_i, X_i\}_{g(i)=g}, N_g\right)|\overline{S}_g, \overline{U}_g, L_g\right] \\
&= \exp^{N_g}\left(\eta_0(L_g)\right)\exp^{N_g}\left(\overline{U}_g^\top S_{1g}\right)\int_{\frac{1}{N_g}\sum_{i=1}^{N_g} S(w_i, x_i) = \overline{S}_{1g}} \\
&\quad \times f_0\left(\{w_i, x_i\}_{i=1}^{N_g}, N_g\right)\prod_{i=1}^{N_g} h(w_i, x_i)\prod_{i=1}^{N_g}(d\mu(w_i, x_i)) \\
&\equiv \exp^{N_g}\left(\eta_0(L_g)\right)\exp^{N_g}\left(\overline{U}_g^\top S_{1g}\right)\tilde{\upsilon}_0(f_0, \overline{S}_g). \qquad (1.8)
\end{aligned}
$$

Again, using the fact that $\tilde{\upsilon}(1, \overline{S}_g, \overline{U}_g) = 1$ we get that $\tilde{\upsilon}(f_0, \overline{S}_g, \overline{U}_g)$ does not depend on $\overline{U}_g$. Using this result, we get for an arbitrary bounded function $f(y_1, y_0)$:

$$
\begin{aligned}
\mathbb{E}[f(Y_i(1), Y_i(0))|W_i, X_i, \overline{S}_{g(i)}] \\
= \mathbb{E}[\mathbb{E}[f(Y_i(1), Y_i(0))|W_i, X_i, \overline{S}_{g(i)}, \overline{U}_{g(i)}]|W_i, X_i, \overline{S}_{g(i)}] \\
= \mathbb{E}[\mathbb{E}[f(Y_i(1), Y_i(0))|X_i, \overline{U}_{g(i)}]|W_i, X_i, \overline{S}_{g(i)}] = \mathbb{E}[f(Y_i(1), Y_i(0))|X_i, \overline{S}_{g(i)}], \qquad (1.9)
\end{aligned}
$$

where the second equality follows from Lemma 1 and Assumption 3.1, and the last equality follows from equation (1.7).
□

B. Inference results

Notation: We are using standard notation from the empirical processes literature adapted to our setting. For any *group-level* random vector X_g define $\mathbb{E}_M(X_g) \equiv \frac{1}{M}\sum_{g=1}^M X_g$. Define $B_i = (X_i, \overline{S}_{g(i)})$ and $D_i \equiv (W_i, B_i)$ and the following objects:

$$
\begin{aligned}
\mu_i &\equiv \mathbb{E}[Y_i|D_i], \quad \hat{\mu}_{i,g} \equiv \hat{\mu}_g(D_i), \quad \mu_{w,i} \equiv \mathbb{E}[Y_i|W_i = w, B_i], \quad \hat{\mu}_{w,i,g} \equiv \hat{\mu}_g(w, B_i), \\
e_i &\equiv \mathbb{E}[W_i|B_i], \quad \hat{e}_{i,g} \equiv \hat{e}_g(B_i).
\end{aligned} \qquad (2.1)
$$

Proof of Theorem 2. We separate $\rho_g(m, p)$ into two parts:

$$
\begin{aligned}
\rho_g(m, p) &= \sum_{i:g(i)=g}\frac{1}{N_g}A_i\left(m(1, B_i) + \frac{W_i}{p(B_i)}(Y_i - m(1, B_i))\right) \\
&\quad - \sum_{i:g(i)=g}\frac{1}{N_g}A_i\left(m(0, B_i) + \frac{1-W_i}{1-p(B_i)}(Y_i - m(0, B_i))\right) = \rho_{1g}(m, p) - \rho_{0g}(m, p). \qquad (2.2)
\end{aligned}
$$

We analyse $\rho_{1g}(\hat{\mu}_g, \hat{e}_g)$; the other term is analysed in the same way. We have the following expansion:

$$\rho_{1g}(\hat{\mu}_{1,g}, \hat{e}_g) = \rho_{1g}(\mu, e) + \sum_{i:g(i)=g} \frac{1}{N_g} A_i \left((\hat{\mu}_{1,i,g} - \mu_{1,i}) \left(1 - \frac{W_i}{e_i} \right) \right)$$

$$+ \sum_{i:g(i)=g} \frac{1}{N_g} A_i \left((\hat{\mu}_{i,g} - \mu_i) W_i \left(\frac{\hat{e}_{i,g} - e_i}{e_i \hat{e}_{i,g}} \right) \right) + \sum_{i:g(i)=g} \frac{1}{N_g} A_i (Y_i - \mu_i) W_i \left(\frac{1}{\hat{e}_{i,g}} - \frac{1}{e_i} \right)$$

$$= \rho_{1g}(\mu, e) + R_{1g} + R_{2g} + R_{3g}. \tag{2.3}$$

We analyse the last three terms separately. By cross-fitting, $\hat{\mu}_{k,i,g}$ is constructed without using the data from group g and since $\mathbb{E}[A_i(1 - \frac{W_i}{e_i})|B_i] = 0$, for the first term we have:

$$\mathbb{E}\left[\mathbb{E}_M R_{1g} \right] = 0. \tag{2.4}$$

We then bound its second moment:

$$\mathbb{E}[(\mathbb{E}_M R_{1g})^2] \leq \frac{L}{M} \mathbb{E}\left[\mathbb{E}_M \sum_{i:g(i)=g} \frac{1}{N_g} A_i \left(\hat{\mu}_{1,i,g} - \mu_{1,i} \right)^2 \left(1 - \frac{W_i}{e_i} \right)^2 \right]$$

$$\leq \frac{1-\eta}{\eta} \frac{L}{M} \mathbb{E}\left[\mathbb{E}_M \sum_{i:g(i)=g} \frac{1}{N_g} A_i \left(\hat{\mu}_{1,i,g} - \mu_{1,i} \right)^2 \right] = o\left(\frac{1}{M} \right). \tag{2.5}$$

From Chebyshev inequality it then follows $\mathbb{E}_M R_{1g} = o_p(\frac{1}{\sqrt{M}})$. The second term is bounded in the following way:

$$\left| \mathbb{E}_M R_{2g} \right| \leq \frac{1}{\eta^2} \left| \mathbb{E}_M \sum_{i:g(i)=g} \frac{1}{N_g} A_i \left(\hat{\mu}_{i,g} - \mu_i \right) \left(\hat{e}_{i,g} - e_i \right) \right|$$

$$\leq \frac{1}{\eta^2} \sqrt{\mathbb{E}_M \sum_{i:g(i)=g} \frac{1}{N_g} A_i \left(\hat{\mu}_{i,g} - \mu_i \right)^2} \sqrt{\mathbb{E}_M \sum_{i:g(i)=g} \frac{1}{N_g} A_i \left(\hat{e}_{i,g} - e_i \right)^2} = o_p\left(\frac{1}{\sqrt{M}} \right). \tag{2.6}$$

Similar to the first term, because $e_{i,g}$ is constructed without using data from the group g and $\mathbb{E}[Y_i - \mu_i | D_i] = 0$, we have for the third term

$$\mathbb{E}\left[\mathbb{E}_M R_{3g} \right] = 0, \tag{2.7}$$

and bounded second moment:

$$\mathbb{E}[(\mathbb{E}_M R_{3g})^2] \leq \frac{L}{M} \frac{16}{\eta^4} \mathbb{E}\left[\mathbb{E}_M \sum_{i:g(i)=g} \frac{1}{N_g} A_i (Y_i - \mu_i)^2 \left(\hat{e}_{i,g} - e_i \right)^2 \right]$$

$$\leq \frac{L}{M} \frac{16K}{\eta^4} \mathbb{E}\left[\mathbb{E}_M \sum_{i:g(i)=g} \frac{1}{N_g} A_i \left(\hat{e}_{i,g} - e_i \right)^2 \right] = o\left(\frac{1}{M} \right). \tag{2.8}$$

From Chebyshev inequality it follows that $\mathbb{E}_M R_{3g} = o_p\left(\frac{1}{\sqrt{M}} \right)$. Combining bounds for R_{1g}, R_{2g}, R_{3g} together we get

$$\mathbb{E}_M \rho_{1g}(\hat{\mu}_g, \hat{e}_g) = \mathbb{E}_M \rho_{1g}(\mu, e) + o_p\left(\frac{1}{\sqrt{M}} \right). \tag{2.9}$$

Repeating the same arguments for $\hat{\rho}_{0g}$ and combining the results we get

$$\mathbb{E}_M \rho_g(\hat{\mu}_g, \hat{e}_g) = \mathbb{E}_M \sum_{i:g(i)=g} \frac{1}{N_g} A_i \tau(B_i) + \mathbb{E}_M \xi_g + o_p\left(\frac{1}{\sqrt{M}} \right). \tag{2.10}$$

By the law of large numbers (LLN), we have $\mathbb{E}_M \sum_{i:g(i)=g} \frac{A_i}{N_g} = \mathbb{E}[A_i] + o_p(1)$ and thus we have the following:

$$\hat{\tau}_{dr} - \frac{\mathbb{E}_M \sum_{i:g(i)=g} \frac{1}{N_g} A_i \tau(B_i)}{\mathbb{E}_M \sum_{i:g(i)=g} \frac{A_i}{N_g}} = \frac{\mathbb{E}_M \xi_g}{\mathbb{E}[A_i]} + o_p\left(\frac{1}{\sqrt{M}}\right). \tag{2.11}$$

Since $\mathbb{E}[\xi_g] = 0$ and ξ_g are i.i.d., the two needed results follow from equation (2.11). First, by the LLN, we get consistency:

$$\hat{\tau}_{dr} = \frac{\mathbb{E}_M \sum_{i:g(i)=g} \frac{1}{N_g} A_i \tau(B_i)}{\mathbb{E}_M \sum_{i:g(i)=g} \frac{A_i}{N_g}} + o_p(1) = \mathbb{E}[\tau(B_i)|B_i \in \mathbb{A}] + o_p(1). \tag{2.12}$$

Second, by the central limit theorem (CLT), we get the asymptotic distribution of our estimator around the conditional estimand:

$$\sqrt{M}\left(\hat{\tau}_{dr} - \frac{\mathbb{E}_M \sum_{i:g(i)=g} \frac{1}{N_g} A_i \tau(B_i)}{\mathbb{E}_M \sum_{i:g(i)=g} \frac{A_i}{N_g}}\right) \Rightarrow \mathcal{N}(0, \mathbb{V}), \tag{2.13}$$

where $\mathbb{V} \equiv \frac{\mathbb{E}[\xi_g^2]}{\mathbb{E}^2[A_i]}$. □

Proof of Theorem 3. Similar to the previous proof, we can divide ξ_g into two parts ξ_{1g} and ξ_{0g}. We will analyse ξ_{1g}, the analysis for ξ_{0g} is the same. We have the following decomposition:

$$\hat{\xi}_{1g} - \xi_{1g} = \sum_{i:g(i)=g} \frac{1}{N_g} A_i\left((\mu_i - \hat{\mu}_{i,g})\frac{W_i}{\hat{e}_{i,g}}\right)$$
$$+ \sum_{i:g(i)=g} \frac{1}{N_g} A_i(Y_i - \mu_i)W_i\left(\frac{1}{\hat{e}_{i,g}} - \frac{1}{e_i}\right) = R_{11g} + R_{12g}. \tag{2.14}$$

For the first term, we have the following bound:

$$(\mathbb{P}_n R_{11g})^2 \leq \mathbb{E}_M \frac{1}{N_g} \sum_{i:g(i)=g} A_i\left((\mu_i - \hat{\mu}_{i,g})^2 \frac{W_i}{\hat{e}_{i,g}^2}\right) \leq \frac{4}{\eta^2} \mathbb{E}_M \frac{1}{N_g} \sum_{i:g(i)=g} A_i\left((\mu_{1,i} - \hat{\mu}_{1,i,g})^2\right) = o_p(1), \tag{2.15}$$

where the last inequality follows from part (i) of Assumption 3.6, and the last equality follows from part (ii).

For the second term, we have the following bound:

$$(\mathbb{E}_M R_{12g})^2 \leq \mathbb{E}_M \frac{1}{N_g} \sum_{i:g(i)=g} A_i W_i \varepsilon_i^2 \frac{(\hat{e}_{i,g} - e_i)^2}{\hat{e}_{i,g}^2 e_i^2}$$
$$\leq \sqrt{\left(\mathbb{E}_M \frac{1}{N_g} \sum_{i:g(i)=g} A_i \frac{(\hat{e}_{i,g} - e_i)^4}{\hat{e}_{i,g}^4 e_i^4}\right)\left(\mathbb{E}_M \frac{\sum_{i:g(i)=g} A_i W_i \varepsilon_i^4}{N_g}\right)}$$
$$\leq \frac{4}{\eta^4} \sqrt{\left(\mathbb{E}_M \frac{1}{N_g} \sum_{i:g(i)=g} A_i(\hat{e}_{i,g} - e_i)^2\right)\left(\mathbb{E}_M \frac{1}{N_g} \sum_{i:g(i)=g} A_i W_i \varepsilon_i^4\right)} = o_p(1) O_p(1) = o_p(1), \tag{2.16}$$

where for the last inequality, we used $\frac{(\hat{e}_{i,g}-e_i)^4}{\hat{e}_{i,g}^4 e_i^4} \leq \frac{(\hat{e}_{i,g}-e_i)^2}{\hat{e}_{i,g}^4 e_i^4} \leq \frac{16}{\eta^8}(\hat{e}_{i,g} - e_i)^2$, which follows from part (i) of Assumption 3.6. For the last equality, we used $\mathbb{E}_M \frac{1}{N_g} \sum_{i:g(i)=g} A_i W_i \varepsilon_i^4 \leq \mathbb{E}_M \frac{1}{N_g} \sum_{i:g(i)=g} \varepsilon_i(1)^4 = \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i(1)^4 = O_p(1)$ by Assumption 3.5.

Putting these results together, we have the following:

$$\mathbb{E}_M(\hat{\xi}_{1g} + \hat{\xi}_{2g})^2 - \mathbb{E}_M(\xi_{1g} + \xi_{2g})^2 = \mathbb{E}_M(\xi_{1g} + \xi_{2g} + R_{11g} + R_{12g} + R_{01g} + R_{02g})^2 - \mathbb{E}_M(\xi_{1g} + \xi_{2g})^2$$

$$= \mathbb{E}_M(\xi_{1g} + \xi_{2g})(R_{11g} + R_{12g} + R_{01g} + R_{02g}) + \mathbb{E}_M(R_{11g} + R_{12g} + R_{01g} + R_{02g})^2$$

$$\leq \sqrt{\mathbb{E}_M(\xi_{1g} + \xi_{2g})^2 4\mathbb{E}_M(R_{11g}^2 + R_{12g}^2 + R_{01g}^2 + R_{02g}^2)} + \mathbb{E}_M(R_{11g}^2 + R_{12g}^2 + R_{01g}^2 + R_{02g}^2)$$

$$= \sqrt{O_p(1)o_p(1)} + o_p(1) = o_p(1). \tag{2.17}$$

This argument also implies that $\mathbb{E}_M(\hat{\xi}_{1g}) = \mathbb{E}_M(\xi_{1g}) = o_p(1)$ and thus we have the final result:

$$\frac{1}{(\overline{A})^2}\left(\mathbb{E}_M(\hat{\xi}_{1g} + \hat{\xi}_{2g})^2 - \left(\mathbb{E}_M(\hat{\xi}_{1g} + \hat{\xi}_{2g})\right)^2\right) - \frac{1}{\mathbb{E}^2[A_i]}\mathbb{E}_M(\xi_{1g} + \xi_{2g})^2$$

$$= \frac{1}{(\overline{A})^2}\left(\mathbb{E}_M(\hat{\xi}_{1g} + \hat{\xi}_{2g})^2 - \mathbb{E}_M(\xi_{1g} + \xi_{2g})^2\right) + \left(\frac{1}{(\overline{A})^2} - \frac{1}{\mathbb{E}^2[A_i]}\right)\mathbb{E}_M(\xi_{1g} + \xi_{2g})^2 + O_p(1)o_p(1)$$

$$= O_p(1)o_p(1) + o_p(1)O_p(1) + O_p(1)o_p(1) = o_p(1). \tag{2.18}$$

$\square$

C. Estimation details
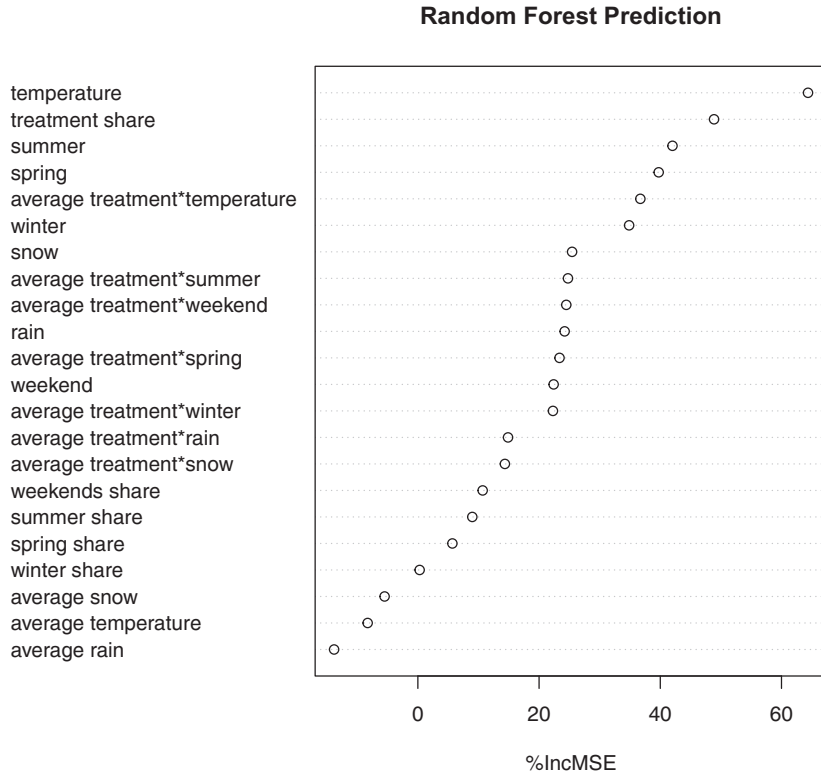
Random Forest Prediction

FIGURE C.1

Variable importance. $M = 2491$, $N = 14{,}029$. Each row is the % increase in the mean-squared error if the corresponding factor was dropped from the propensity score model estimated by the random forest

Data Availability Statement

The data and code underlying this research are available on Zenodo at https://doi.org/10.5281/zenodo.8226636.

REFERENCES

ABADIE, A., ATHEY, S., IMBENS, G. W., *et al.* (2020), "Sampling-Based Versus Design-Based Uncertainty in Regression Analysis", *Econometrica*, **88**, 265–296.

ABADIE, A. and CATTANEO, M. D. (2018), "Econometric Methods for Program Evaluation", *Annual Review of Economics*, **10**, 465–503.

ALLMAN, E. S., MATIAS, C. and RHODES, J. A. (2009), "Identifiability of Parameters in Latent Structure Models with Many Observed Variables", *Annals of Statistics*, **37**, 3099–3132.

ALTONJI, J. G. and MATZKIN, R. L. (2005), "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors", *Econometrica*, **73**, 1053–1102.

ANGRIST, J. D. and PISCHKE, J.-S. (2008), *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton University Press). ISBN 9780691120355.

ARELLANO, M. (2003), *Panel Data Econometrics* (Oxford University Press). ISBN 9780199245291.

ARELLANO, M. and BONHOMME, S. (2011), "Nonlinear Panel Data Analysis", *Annual Review of Economics*, **3**, 395–424.

——, —— (2016), "Nonlinear Panel Data Estimation via Quantile Regressions", *The Econometrics Journal*, **19**, C61–C94.

ARKHANGELSKY, D. and IMBENS, G. (2018), "The Role of the Propensity Score in Fixed Effect Models" (Technical Report, National Bureau of Economic Research).

ARPINO, B. and MEALLI, F. (2011), "The Specification of the Propensity Score in Multilevel Observational Studies", *Computational Statistics & Data Analysis*, **55**, 1770–1780.

ATHEY, S., IMBENS, G. W. and WAGER, S. (2018), "Approximate Residual Balancing: Debiased Inference of Average Treatment Effects in High Dimensions", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80**, 597–623.

BARRON, A. R. and SHEU, C.-H. (1991), "Approximation of Density Functions by Sequences of Exponential Families", *Annals of Statistics*, **19**, 1347–1369.

BONHOMME, S., JOCHMANS, K. and ROBIN, J.-M. (2016), "Estimating Multivariate Latent-Structure Models", *Annals of Statistics*, **44**, 540–563.

BONHOMME, S., LAMADON, T. and MANRESA, E. (2022), "Discretizing Unobserved Heterogeneity", *Econometrica*, **90**, 625–643.

BONHOMME, S. and MANRESA, E. (2015), "Grouped Patterns of Heterogeneity in Panel Data", *Econometrica*, **83**, 1147–1184.

BREIMAN, L. (2001), "Random Forests", *Machine Learning*, **45**, 5–32.

CHAMBERLAIN, G. (1984), "Panel Data", *Handbook of Econometrics*, **2**, 1247–1318.

—— (2010), "Binary Response Models for Panel Data: Identification and Information", *Econometrica*, **78**, 159–168.

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., *et al.* (2017), "Double/Debiased/Neyman Machine Learning of Treatment Effects", *American Economic Review*, **107**, 261–265.

——, ——, —— (2018a), "Double/Debiased Machine Learning for Treatment and Structural Parameters", *The Econometrics Journal*, **21**, C1–C68.

CHERNOZHUKOV, V., NEWEY, W. K. and ROBINS, J. (2018b), "Double/De-Biased Machine Learning Using Regularized Riesz Representers" (Technical Report, CeMMAP Working Paper).

COX, D. R. and HINKLEY, D. V. (1979), *Theoretical Statistics* (CRC Press). https://doi.org/10.1201/b14832.

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W., *et al.* (2009), "Dealing with Limited Overlap in Estimation of Average Treatment Effects", *Biometrika*, **96**, 187–199.

DE CHAISEMARTIN, C. and D'HAULTFOEUILLE, X. (2020), "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects", *American Economic Review*, **110**, 2964–2996.

DE PAULA, A., RASUL, I. and SOUZA, P. (2019), "Identifying Network Ties From Panel Data: Theory and an Application to Tax Competition" (arXiv preprint arXiv:1910.07452).

FIRPO, S. (2007), "Efficient Semiparametric Estimation of Quantile Treatment Effects", *Econometrica*, **75**, 259–276.

GOODMAN-BACON, A. (2021), "Difference-In-Differences with Variation in Treatment Timing", *Journal of Econometrics*, **225**, 254–277.

HAHN, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects", *Econometrica*, **66**, 315–331.

HAHN, J. and MOON, H. R. (2010), "Panel Data Models with Finite Number of Multiple Equilibria", *Econometric Theory*, **26**, 863–881.

HANSEN, B. (2022), *Econometrics* (Princeton University Press). ISBN 9780691235899.

HIRSHBERG, D. A. and WAGER, S. (2021), "Augmented Minimax Linear Estimation", *Annals of Statistics*, **49**, 3206–3227.

HU, Y. and SCHENNACH, S. M. (2008), "Instrumental Variable Treatment of Nonclassical Measurement Error Models", *Econometrica*, **76**, 195–216.

IMAI, K. and KIM, I. S. (2019), "When Should we use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?", *American Journal of Political Science*, **63**, 467–490.

IMBENS, G. W. and RUBIN, D. B. (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge University Press). ISBN 9781139025751.

JOHANNEMANN, J., HADAD, V., ATHEY, S., *et al.* (2019), "Sufficient Representations for Categorical Variables" (arXiv preprint arXiv:1908.09874).

LACETERA, N., MACIS, M. and SLONIM, R. (2012a), "Replication Data for: Will There Be Blood? Incentives and Displacement Effects in Pro-Social Behavior" https://www.openicpsr.org/openicpsr/project/114774/version/V1/view.

——, ——, —— (2012b), "Will There be Blood? Incentives and Displacement Effects in Pro-Social Behavior", *American Economic Journal: Economic Policy*, **4**, 186–223.

——, ——, —— (2023), "Replication Data for: Will There Be Blood? Incentives and Displacement Effects in Pro-Social Behavior", https://drive.google.com/file/d/1D0h8d29oOt7kpuH9bBBweAmTeIDd6IS7/view.

LANCASTER, T. (2000), "The Incidental Parameter Problem Since 1948", *Journal of Econometrics*, **95**, 391–413.

LEO LEHMANN, E. and D'ABRERA, H. J. M. (2006), *Nonparametrics: Statistical Methods Based on Ranks* (New York: Springer).

MUNDLAK, Y. (1978), "On the Pooling of Time Series and Cross Section Data", *Econometrica: Journal of the Econometric Society*, **46**, 69–85.

NEWEY, W. K. and ROBINS, J. R. (2018), "Cross-Fitting and Fast Remainder Rates for Semiparametric Estimation" (arXiv preprint arXiv:1801.09138).

NEYMAN, J. (1923/1990), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9", *Statistical Science*, **5**, 465–472.

NEYMAN, J. and SCOTT, E. L. (1948), "Consistent Estimates Based on Partially Consistent Observations", *Econometrica: Journal of the Econometric Society*, **16**, 1–32.

ROBINS, J. M. and ROTNITZKY, A. (1995), "Semiparametric Efficiency in Multivariate Regression Models with Missing Data", *Journal of the American Statistical Association*, **90**, 122–129.

ROSENBAUM, P. R. (2002), *Observational Studies* (Springer). https://doi.org/10.1007/978-1-4757-3692-2.

ROSENBAUM, P. R. and RUBIN, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, **70**, 41–55.

RUBIN, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology*, **66**, 688–701.

—— (1977), "Assignment to Treatment Group on the Basis of a Covariate", *Journal of Educational Statistics*, **2**, 1–26.

TIBSHIRANI, R. (1996), "Regression Shrinkage and Selection via the Lasso", *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**, 267–288.

WOOLDRIDGE, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data* (MIT Press). ISBN 9780262232586.

—— (2021), "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators" (Available at SSRN 3906345).

ZOU, H. and ZHANG, H. H. (2009), "On the Adaptive Elastic-Net with a Diverging Number of Parameters", *Annals of Statistics*, **37**, 1733.

ZUBIZARRETA, J. R. (2015), "Stable Weights That Balance Covariates for Estimation with Incomplete Outcome Data", *Journal of the American Statistical Association*, **110**, 910–922.