Estimating and predicting treatment-effect heterogeneity across sites, in multi-site randomized experiments with few randomization units per site.

Clément de Chaisemartin(r) [†] Antoine Deeb[‡]

First version: October 31st 2023
This version: October 25, 2024

Abstract

We seek to estimate and predict treatment-effect heterogeneity across sites, in multi-site randomized controlled trials, with a large number of sites but few randomization units per site. As is well-known, an Empirical-Bayes (EB) estimator can be used to estimate the variance of the treatment effect across sites. We propose consistent estimators of the coefficients from ridge and OLS regressions of site-level effects on site-level characteristics that are unobserved but can be unbiasedly estimated, such as sites' average outcome without treatment, or site-specific treatment effects on mediator variables. In experiments with imperfect compliance, we also propose a non-parametric and partly testable assumption under which the variance of local average treatment effects (LATEs) across sites can be estimated. We revisit Behaghel et al. (2014), who study the effect of counseling programs on job seekers job-finding rate, in 200 job placement agencies in France. We find considerable treatment-effect heterogeneity, both for intention to treat and LATE effects, and the treatment effect is negatively correlated with sites' job-finding rate without treatment.

*We are very grateful to Xavier D'Haultfoeuille and Peng Ding for their helpful comments. Clément de Chaisemartin was funded by the European Union (ERC, REALLYCREDIBLE,GA N°101043899).

[†]Department of Economics, Sciences Po

[‡]Development Impact Evaluation, World Bank

1 Introduction

Research question. From 2014 to 2016, "AEJ: Applied Economics" published 12 multi-site RCTs with treated and control units within each site, thus making it possible to estimate the treatment effect in each site. Typically, those RCTs are conducted in dozens, and sometimes hundreds, of different neighborhoods, villages, or regions, but they have a small number of randomization units per site. Those RCTs also often have imperfect compliance with the initial treatment assignment. Few of these 12 papers investigate the treatment-effect's heterogeneity across sites.[1] Estimating the treatment-effect's variance across sites can be helpful to assess if the effect is context- and implementation-dependent. Regressing site-specific effects on some predictors can be helpful to provide suggestive evidence as to the mechanisms underlying the treatment's effect. For instance, in a multi-site job-search counseling RCT, it can be interesting to study whether sites that have the largest effects on job-seekers' job finding rate are also the sites that have the largest effect on their search effort, as a "predictive mediation analysis" of whether the job-finding effect can be "explained" by the job-search effect.

Set-up. We consider an RCT stratified at the site level. We allow for imperfect compliance with treatment assignment, and consider both the heterogeneity of intention-to-treat effects (ITTs) and local-average-treatment-effects (LATEs) across sites. We assume that each site has at least two treated and two control units, so that ITT_s, the ITT effect of site s, can be unbiasedly estimated, using an estimator $\widehat{\text{ITT}}_s$ whose variance can also be unbiasedly estimated. Finally, in our asymptotic analysis, we assume that the number of randomization units in each site n_s is fixed, while the number of sites S goes to infinity, hereafter referred to as a "large S small n_s" sequence. This thought experiment seems well suited to the multi-site RCTs in our survey: 10 out of 12 have at least 40 sites, while the median number of units per site is 12.5.

Estimating the variance of ITTs across sites. As is well-known, to non-parametrically estimate the ITTs' variance across sites, one can use the Empirical Bayes (EB) variance estimator (Morris, 1983). In a multi-site RCT, the EB estimator is equal the variance of $\widehat{\text{ITT}}_s$ across sites,

[1] One paper estimates the treatment-effect's variance across sites, and two more estimate the average treatment effect separately for different subgroups of geographical locations.

minus the average of robust variance estimators of the $\widehat{\text{ITT}}_s$ estimators.

Predicting site-specific ITT effects. Then, we turn attention to $\beta_X^{\text{ITT}}(\lambda)$, the coefficient from a ridge regression of the site-specific ITTs on $\mathbf{X}_s$, a vector of covariates, with hyper-parameter λ. OLS is a special case of a ridge regression, with $\lambda = 0$. Ridge regressions lead to more precisely estimated regression coefficients than OLS when the number of regressors is not negligible with respect to the sample size. This makes them an appealing alternative to OLS in multi-site RCTs, where one typically has a few dozens to a few hundreds of sites. Some of the elements of $\mathbf{X}_s$ might be unobserved variables that can be unbiasedly estimated. For instance, one may want to regress sites' ITTs on sites' outcomes without a treatment offer, to assess if treatment offers reduce or increase inequalities across sites. One could also be interested in regressing the ITTs for the main outcome variable on sites' ITTs for mediator variables, like in the aforementioned job finding/job search example. To consistently estimate $\beta_X^{\text{ITT}}(\lambda)$, one cannot simply regress the estimated ITTs on the estimated covariates $\widehat{\mathbf{X}}_s$: one needs to account for the fact that the regression's dependent and explanatory variables are estimators. This can be achieved easily, given that in an RCT one can estimate the variances of those estimators, and the covariances between them (Li and Ding, 2017). We show that the resulting estimator $\widehat{\beta}_X^{\text{ITT}}(\lambda)$ is asymptotically normal, and we provide a conservative estimator of its asymptotic variance. We also show that the "optimal" hyper-parameter, based on the generalized cross-validation method of Golub et al. (1979), can be consistently estimated.

Estimating and predicting LATEs' heterogeneity. As shown by Walters (2015), in multi-site RCTs with imperfect compliance, a naive EB estimator using site-specific 2SLS estimators as building blocks is often negative and therefore uninformative on the LATEs' variance, because sites with first-stages (FSs) close to zero have large variances. Moreover, as the site-specific LATE estimators are not unbiased, this naive EB estimator is not necessarily consistent in the "large S small n_s" sequence we consider. To bypass this issue, we propose a testable non-parametric assumption under which the LATEs' variance can be written as a function of sites' ITTs and FSs, thus allowing us to use an EB estimator leveraging only unbiased ITTs and FSs estimators. This non-parametric assumption requires that sites' FSs and LATEs are independent. Its testable implication is that the LATE is equal to the coefficient from a regression of sites' ITTs on their

FSs. Finally, we show that for a binary covariate X_s, the coefficient β_X^{LATE} from a regression of the site-specific LATEs on X_s can be consistently estimated, again under our assumption that LATEs and FSs are independent. Thus, techniques to estimate and predict ITTs heterogeneity can be parly extended to estimate and predict LATEs heterogeneity, at the expense of imposing a strong but partly testable assumption.

Extensions to other RCT designs. While the estimators described above assume that the RCT is stratified at the site level, they readily extend to multi-site RCTs stratified at a finer level. On the other hand, and though this may still be a feasible extension, it is less immediate to extend them to RCTs where the site-specific estimators are correlated, as can for instance happen in unstratified multi-site RCTs or in RCTs stratified at a coarser level than the sites.

Application We use our results to revisit Behaghel et al. (2014), who conducted an RCT to study the effect of intensive counseling programs on job seekers' employment, in more than 200 local public employment offices in France. In each site, job seekers are randomly assigned to either the control group, or to a program ran by the local public employment service, or to a program ran by a local private provider. Both programs increase job seekers' job finding rate by around 2 percentage points. Using the EB estimator, we find that the standard deviation of the ITT effects across sites is equal to 469% of the ITT estimate for the public program, and to 389% of the ITT estimate for the private one. Assuming for illustrative purposes that site-specific ITTs follow a normal distribution, the public and private programs respectively have a *negative* effect in 42% and 40% of the sites. Surprisingly, sites ITT effects are not significantly correlated with their FS effects. On the other hand, ITT effects are negatively correlated with sites' average job-finding rate without treatment. Thus, to increase the programs' effectiveness, one could target them to the sites where earlier cohorts of job seekers had the lowest job finding rate. Finally, we find that the ITTs of the public and private programs are strongly positively correlated, and we can actually not rule out a perfect positive correlation. In the same site, the public and private programs are delivered by different providers. Thus, this suggests that effects' heterogeneity is not entirely driven by providers' effects. Turning to LATEs, our test of the assumption that FSs and LATEs are independent is not rejected. Under that assumption, we estimate that the standard deviation of the effects across sites is equal to 314% of the LATE

estimate for the public program, and to 377% for the private one.

1.1 Related literature and contributions

Estimating the variance of ITTs across sites. In the evaluation literature, the EB variance estimator has already been put forward as a tool to estimate the variance of ITT effects across sites in multi-site RCTs (see Equation (16) in Raudenbush and Bloom, 2015).

Predicting site-specific ITT effects. In the literature on multi-site RCTs, the most closely related paper is again Raudenbush and Bloom (2015), who propose an estimator of the covariance between sites' ITTs and their average outcome without treatment (see their Equation (18)). Here, our contribution is to propose an estimator of coefficients from multivariate ridge or OLS regressions of ITTs on a vector of estimated variables, and to study its asymptotic distribution. On the applied side, the idea of correlating site-specific ITT effects on a main outcome and on mediators might also be a contribution of this paper. Ideas similar to those we use to predict site-specific ITTs have appeared in the teacher value-added literature, see in particular Kline et al. (2020) and Rose et al. (2022). In that literature, the most closely related paper is Rose et al. (2022), who propose an estimator of coefficients from a multivariate OLS regression of teachers' long-run value added on their value-added on a vector of short-run outcomes. There as well, one can account for the fact that the regression's dependent and explanatory variables are estimators, by using unbiased estimators of their variances and covariances, but those variances and covariances estimators differ in multi-site RCTs and in teacher value-added models (see the discussion surrounding Equations (5) and (8) in Rose et al., 2022), so estimators are not numerically equivalent after some relabeling.[2] More generally, note that in the setting we consider, standard results on two-steps estimators cannot be used to propose a consistent estimator of $\beta_X^{\text{ITT}}(\lambda)$, because those results assume that the first-step estimators are consistent (see e.g. Theorem 6.1 in Newey and McFadden, 1994), which is not the case in the "large S small n_s" sequence we consider. Also, a vast literature in meta-analysis studies meta-regressions, namely

[2]In our application, we re-estimate the variance of the ITT effects of the public program using the estimator of Kline et al. (2020), and find a sizeably different estimator.

regressions of study-specific effects on moderators (see e.g. Stanley and Doucouliagos, 2012, for a textbook treatment). However, this literature has mostly considered the case where the moderators are observed variables that do not need to be estimated. Moreover, this literature often assumes that estimated effects are normally, or at least symmetrically distributed, which is not warranted in the "large S small n_s" setting we consider.

Estimating and predicting LATEs' heterogeneity. Other papers have tried to bypass the issue that a naive EB estimator cannot be used to estimate the variance of LATEs in "large S small n_s" multi-site RCTs. Walters (2015) estimates a parametric random-coefficient model, while Adusumilli et al. (2024) estimate a parametric grouped-random-effect model. Instead, we pursue a different and complementary route, where we try to estimate that variance under a testable non-parametric assumption, namely that LATEs and FSs are independent. Our work is also related to that of Angrist and Meager (2023), who study the heterogeneity of ITTs and LATEs of an educational intervention, in a setting with a small number of large studies ("small S large n_s"). Using frequentist and Bayesian meta-analytic techniques, they find a much larger variance of ITTs than LATEs, because the variance in ITTs is mostly driven by the variance in FSs and implementation quality. In a very different context, we find that the variability in FSs and implementer effects do not seem to explain the variability in ITTs.

2 Set-up

Completely randomized experiment, with at least two units assigned to treatment and control per site. We consider a stratified RCT with S sites. Site s has n_s units, and let $n = \sum_{s=1}^{S} n_s$ denote the total number of units in the RCT. Let Z_{is} be an indicator for whether unit i in site s is assigned to treatment. $\mathbf{Z}_s$ stacks all assignment indicators in site s.

Assumption 1 *For all s, there exists $n_{1s} \in \{2, ..., n_s - 2\}$ such that for every $(z_1, ..., z_{n_s}) \in \{0,1\}^{n_s}$ such that $z_1 + ... + z_{n_s} = n_{1s}$, $P(\boldsymbol{Z}_s = (z_1, ..., z_{n_s})) = \frac{1}{\binom{n_s}{n_{1s}}}$.*

Potential treatments, outcomes, and mediators. For all $(i, s) \in \{1, ..., n_s\} \times \{1, ..., S\}$, the potential treatments of unit i in site s without and with assignment to treatment are de-

noted $D_{is}(0)$ and $D_{is}(1)$. Similarly, their potential outcomes without and with treatment are denoted $Y_{is}(0)$ and $Y_{is}(1)$.[3] Furthermore, we let $\mathbf{M}_{is}(0)$ denote a vector stacking the values of m intermediate outcomes, or mediators, without treatment, while $\mathbf{M}_{is}(1)$ denotes the values of the mediators with treatment. Then to simplify notation let us introduce "reduced-form" potential outcome and mediators, that are functions of the assignment to treatment: $Y^r_{is}(0) = Y_{is}(D_{is}(0))$, $Y^r_{is}(1) = Y_{is}(D_{is}(1))$, $\mathbf{M}^r_{is}(0) = \mathbf{M}_{is}(D_{is}(0))$, and $\mathbf{M}^r_{is}(1) = \mathbf{M}_{is}(D_{is}(1))$. Finally, let $D_{is} = Z_{is}D_{is}(1)+(1-Z_{is})D_{is}(0)$, $Y_{is} = Z_{is}Y^r_{is}(1)+(1-Z_{is})Y^r_{is}(0)$, and $\mathbf{M}_{is} = Z_{is}\mathbf{M}^r_{is}(1)+(1-Z_{is})\mathbf{M}^r_{is}(0)$ denote the units' observed treatment, outcome, and mediators. We assume that potential treatments, outcomes, and mediators are independent and identically distributed (iid) in each site, independent of the treatment assignment in each site, and that potential treatments, outcomes, and mediators, as well as assignments, are independent across sites.

Assumption 2 *1. For all s, the vectors $(D_{is}(0), D_{is}(1), Y_{is}(0), Y_{is}(1), \mathbf{M}_{is}(0), \mathbf{M}_{is}(1))$ are independent and identically distributed across i.*

2. For all s, $(D_{is}(0), D_{is}(1), Y_{is}(0), Y_{is}(1), \mathbf{M}_{is}(0), \mathbf{M}_{is}(1))_{i\in\{1,\dots,n_s\}} \perp\!\!\!\perp \boldsymbol{Z}_s$.

3. The random vectors $((D_{is}(0), D_{is}(1), Y_{is}(0), Y_{is}(1), \mathbf{M}_{is}(0), \mathbf{M}_{is}(1))_{i\in\{1,\dots,n_s\}}, \boldsymbol{Z}_s)$ are mutually independent across s.

Assumption 2 for instance holds if in each site, the units included in the experiment are randomly drawn from a larger population. When units are not effectively drawn from a larger population, one can assume that such sampling took place. Then, all effects below apply to this hypothetical larger population, rather than to the study sample only. Assuming random sampling is convenient to avoid the well-known issue that in RCTs conducted in convenience samples, the variance of treatment-effect estimators is not identified (Neyman, 1923). As potential treatments and outcomes are assumed to be iid in each site, for all s let $(D_s(0), D_s(1), Y_s(0), Y_s(1), \mathbf{M}_s(0), \mathbf{M}_s(1))$ denote a vector with the same probability distribution as the vectors $(D_{is}(0), D_{is}(1), Y_{is}(0), Y_{is}(1), \mathbf{M}_{is}(0), \mathbf{M}_{is}(1))$.

[3]This notation implicitly assumes that assignment to treatment has no direct effect on the outcome, the so-called exclusion restriction, see Angrist et al. (1996).

First-stage and intention-to-treat effects. For all s let

$$\text{FS}_s = E(D_s(1) - D_s(0))$$

denote the first-stage (FS) effect in site s, and let

$$\text{FS} = \sum_s w_s \text{FS}_s$$

be a weighted average of the FSs across sites, for some non-negative and non-stochastic weights w_s that sum to one. With $w_s = n_s/n$, FS is the FS effect across units. With $w_s = 1/S$, FS is the FS effect across sites.[4] Similarly, for all s let

$$\text{ITT}_s = E(Y_s^r(1) - Y_s^r(0))$$

denote the intention-to-treat effect in site s, and let

$$\text{ITT} = \sum_s w_s \text{ITT}_s.$$

Finally, for all s let

$$\mathbf{ITT}_{\text{M},s} = E(\mathbf{M}_s^r(1) - \mathbf{M}_s^r(0))$$

denote the intention-to-treat effects on the mediators in site s, and let

$$\mathbf{ITT}_{\text{M}} = \sum_s w_s \mathbf{ITT}_{\text{M},s}.$$

Local average treatment effects. As in Imbens and Angrist (1994), we assume that monotonicity holds and that the first-stage is strictly positive:

Assumption 3 *For all s $D_s(1) \geq D_s(0)$, and $FS > 0$.*

Then, for all s such that $\text{FS}_s > 0$, let

$$\text{LATE}_s = \frac{\text{ITT}_s}{\text{FS}_s}$$

denote the local average treatment effect (LATE) in site s, and let

$$\text{LATE} = \frac{\text{ITT}}{\text{FS}}.$$

[4] If the analysis is at a more disaggregated level than randomization units (e.g. the randomization is at the village level and stratified at the region level, but the analysis is at the villager level), w_s could be proportional to the number of observations in site s.

FS, ITT, and LATE estimators. For all s, let $n_{0s} = n_s - n_{1s}$ denote the number of untreated units in site s. For any generic variable x_{is} defined for all $i \in \{1, ..., n_s\}$ and $s \in \{1, ..., S\}$, let $\overline{x}_s = \frac{1}{n_s}\sum_{i=1}^{n_s} x_{is}$ denote the average of x_{is} in site s, let $\overline{x}_{1s} = \frac{1}{n_{1s}}\sum_{i=1}^{n_s} Z_{is}x_{is}$ and $\overline{x}_{0s} = \frac{1}{n_{0s}}\sum_{i=1}^{n_s}(1 - Z_{is})x_{is}$ respectively denote the average of x_{is} among the treated and untreated units in site s, and let $\overline{x} = \frac{1}{S}\sum_{s=1}\overline{x}_s$ denote the average of x_s across sites. Then, let $\tilde{w}_s = Sw_s$ denote the weights re-scaled by the number of sites. For example, if $w_s = \frac{1}{S}$ then $\tilde{w}_s = 1$ and if $w_s = \frac{n_s}{n}$ $\tilde{w}_s = \frac{n_s}{\overline{n}}$ where $\overline{n}$ is the average number of units per site. Finally, let

$$\begin{aligned}
\widehat{\mathrm{FS}}_s &= \overline{D}_{1s} - \overline{D}_{0s} \\
\widehat{\mathrm{ITT}}_s &= \overline{Y}_{1s} - \overline{Y}_{0s} \\
\widehat{\mathbf{ITT}}_{\mathrm{M},s} &= \overline{\mathbf{M}}_{1s} - \overline{\mathbf{M}}_{0s}, \\
\widehat{\mathrm{LATE}}_s &= \widehat{\mathrm{ITT}}_s / \widehat{\mathrm{FS}}_s
\end{aligned}$$

respectively denote the FS, ITTs, and LATE estimators in site s, and let

$$\begin{aligned}
\widehat{\mathrm{FS}} &= \frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\widehat{\mathrm{FS}}_s \\
\widehat{\mathrm{ITT}} &= \frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\widehat{\mathrm{ITT}}_s \\
\widehat{\mathbf{ITT}}_{\mathrm{M}} &= \frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\widehat{\mathbf{ITT}}_{\mathrm{M},s} \\
\widehat{\mathrm{LATE}} &= \widehat{\mathrm{ITT}}/\widehat{\mathrm{FS}}
\end{aligned}$$

respectively denote the FS, ITTs, and LATE estimators across sites. Under Assumptions 1 and 2, $\widehat{\mathrm{FS}}_s$, $\widehat{\mathrm{ITT}}_s$, and $\widehat{\mathbf{ITT}}_{\mathrm{M},s}$ are unbiased, so $\widehat{\mathrm{FS}}$, $\widehat{\mathrm{ITT}}$, and $\widehat{\mathbf{ITT}}_{\mathrm{M}}$ are also unbiased.

Robust site-specific variance estimators. For all $s \in \{1, ..., S\}$, for any variable x_{is} defined for every $i \in \{1, ..., n_s\}$, let $r^2_{x,s} = \frac{1}{n_s - 1}\sum_{i=1}^{n_s}(x_{is} - \overline{x}_s)^2$ denote the variance of x_{is} in site s, and let $r^2_{x,1,s} = \frac{1}{n_{1s}-1}\sum_{i=1}^{n_s} Z_{is}(x_{is} - \overline{x}_{1s})^2$ and $r^2_{x,0,s} = \frac{1}{n_{0s}-1}\sum_{i=1}^{n_s}(1 - Z_{is})(x_{is} - \overline{x}_{0s})^2$ respectively denote the variance of x_{is} among the treated and untreated units in site s. Then let,

$$\widehat{V}_{rob}\left(\widehat{\mathrm{ITT}}_s\right) = \frac{1}{n_{1s}}r^2_{Y,1,s} + \frac{1}{n_{0s}}r^2_{Y,0,s} \tag{1}$$

denote the robust estimator of the variance of $\widehat{\mathrm{ITT}}_s$ (Eicker et al., 1963; Huber et al., 1967; White et al., 1980). As is well-known (see, e.g., Equation (6.17) in Imbens and Rubin, 2015),

under Assumptions 1 and 2,

$$E\left(\widehat{V}_{rob}\left(\widehat{\text{ITT}}_s\right)\right) = V\left(\widehat{\text{ITT}}_s\right). \tag{2}$$

Robust site-specific covariance estimators. For all $s \in \{1, ..., S\}$, for any variables q_{is} and x_{is} defined for every $i \in \{1, ..., n_s\}$, let $c_{q,x,s} = \frac{1}{n_s - 1}\sum_{i=1}^{n_s}(q_{is} - \overline{q}_s)(x_{is} - \overline{x}_s)$ denote the covariance between q_{is} and x_{is} in site s, and let $c_{q,x,1,s} = \frac{1}{n_{1s}-1}\sum_{i=1}^{n_s} Z_{is}(q_{is} - \overline{q}_{1s})(x_{is} - \overline{x}_{1s})$ and $c_{q,x,0,s} = \frac{1}{n_{0s}-1}\sum_{i=1}^{n_s}(1 - Z_{is})(q_{is} - \overline{q}_{1s})(x_{is} - \overline{x}_{0s})$ respectively denote the covariance between q_{is} and x_{is} among the treated and untreated units in site s.

Variances across sites. As many of our target parameters are variances or covariances of vectors of real numbers across sites, we introduce a dedicated notation. Let A^T denote the transpose of a matrix A. For any site-specific $K \times 1$ vector of real numbers $(\mathbf{U}_s)_{s\in\{1,...,S\}}$, let

$$\sigma^2\left[\mathbf{U}\right] = \sum_{s=1}^{S} w_s \left(\mathbf{U}_s - \sum_{s'=1}^{S} w_{s'}\mathbf{U}_{s'}\right)\left(\mathbf{U}_s - \sum_{s'=1}^{S} w_{s'}\mathbf{U}_{s'}\right)^T$$

denote the weighted variance matrix of those vectors across sites.

3 Application: the effects of publicly- and privately-provided counseling for job seekers.

Behaghel et al. (2014) conduct a large-scale RCT, in 216 local Public Employment Service (PES) offices in France, to compare the public and private provision of counseling to job seekers. During their first interview at the local PES office, 43,977 job seekers are randomly assigned to one of three groups, with assignment probabilities varying locally. The first group is a control group, where they receive the standard services provided by the PES. The second group is assigned to an intensive counseling program provided by the PES, and the third is assigned to an intensive counseling program provided by a private provider. Our framework is applicable to this RCT, with local public employment offices as sites and job seekers as randomization units. A first slight difference is that each unemployed has two assignment variables $Z_{1,is}$ and $Z_{2,is}$, respectively equal to one if they are assigned to the PES-operated and to the privately-operated program. This difference is immaterial for our results. For instance, if one is interested in the heterogeneous

effects of the PES-provided program, in the estimators defined below one lets Z_{is} stand for $Z_{1,is}$, and one drops job seekers assigned to the privately-operated program from the sample.[5] A second slight difference is that for the private program, 12 offices have less than two treated or two control units: they have to be dropped from our analysis. For the public program, 16 offices have to be dropped for the same reason. Compliance with randomized assignment is imperfect. While almost no job seekers unassigned to the counseling programs gets access to them, only 32% (resp. 43%) of job seekers assigned to the public (resp. private) counseling program took it up. The outcome we consider is an indicator for holding any employment 6 months after randomization, one of the three main employment outcomes considered by the authors. Results are similar if we consider the authors' two other outcomes. Unfortunately, the authors' data set does not contain mediators, such as measures of workers' job-search effort.

4 Estimating and predicting ITTs' and FSs' heterogeneity.

4.1 Estimating the variance of ITTs and FSs across sites.

Target parameters. In this section, our target parameter is $\sigma^2\left[\text{ITT}\right]$, the variance of the ITTs across sites. The variance of the FS effects and the variances of the ITT effects on the mediators can be estimated similarly.

Estimating $\sigma^2\left[\text{ITT}\right]$ using an Empirical Bayes estimator. Let

$$\widehat{\sigma}^2\left[\text{ITT}\right] = \sum_{s=1}^{S} w_s \left[\left(\widehat{\text{ITT}}_s - \widehat{\text{ITT}}\right)^2 - \widehat{V}_{rob}\left(\widehat{\text{ITT}}_s\right)\right].$$

$\widehat{\sigma}^2\left[\text{ITT}\right]$ is the standard Empirical Bayes (EB) variance estimator (Morris, 1983), applied to multi-site RCTs. $\widehat{\sigma}^2\left[\text{ITT}\right]$ can easily be extended to multi-site RCTs stratified at a finer level than sites. Then, one needs to define strata-specific robust variance estimators, and redefine $\widehat{V}_{rob}\left(\widehat{\text{ITT}}_s\right)$ as a weighted sum of those estimators, across all strata of site s.

[5] In particular, it follows from Theorem 3 in Li and Ding (2017) that the formulas in Neyman (1923) for the variance of comparisons of treated and control observations still apply to RCTs with more than two treatments.

Asymptotic distribution of the EB estimator. Let $\phi_{s,1} = \tilde{w}_s \left[\left(\widehat{\text{ITT}}_s - \text{ITT} \right)^2 - \widehat{V}_{rob} \left(\widehat{\text{ITT}}_s \right) \right]$.

Assumption 4 *Sufficient conditions under which $\widehat{\sigma}^2 [ITT]$ is asymptotically normal.*

1. *The sequences $\left(\tilde{w}_s \widehat{ITT}_s \right)_{s \geq 1}$ and $(\phi_{s,1})_{s \geq 1}$ satisfy the Lyapunov condition.*
2. *For all s, $\tilde{w}_s < N$ for some $N > 0$ and $N < +\infty$.*
3. *ITT, $\frac{1}{S} \sum_{s=1}^{S} V(\phi_{s,1})$, $\frac{1}{S} \sum_{s=1}^{S} E(\phi_{s,1})$, $\frac{1}{S} \sum_{s=1}^{S} E(\phi_{s,1}^2)$ converge towards finite limits when $S \to \infty$.*

Point 1 of Assumption 4 requires that one can apply the Lyapunov central limit theorem to $\widehat{\text{ITT}}$ and to an infeasible version of $\widehat{\sigma}^2 [\text{ITT}]$ where $\widehat{\text{ITT}}$ is replaced by ITT. Point 2 of Assumption 4 requires that the rescaled weights for each site be bounded. Finally, Point 3 requires that certain deterministic averages have finite limits. Under Assumption 4, let

$$V_{\sigma^2[\text{ITT}]} = \lim_{S \to \infty} \frac{1}{S} \sum_{s=1}^{S} V(\phi_{s,1}),$$

and let $\widehat{\phi}_{s,1} = \tilde{w}_s \left[\left(\widehat{\text{ITT}}_s - \widehat{\text{ITT}} \right)^2 - \widehat{V}_{rob} \left(\widehat{\text{ITT}}_s \right) \right]$ and

$$\widehat{V}_{\sigma^2[\text{ITT}]} = \frac{1}{S} \sum_{s=1}^{S} \left[\widehat{\phi}_{s,1} - \overline{\widehat{\phi}_1} \right]^2.$$

Theorem 1 *If Assumptions 1 and 4 hold,*

$$\sqrt{S} \left(\widehat{\sigma}^2 [ITT] - \sigma^2 [ITT] \right) \xrightarrow{d} N(0, V_{\sigma^2[ITT]}),$$

and $\widehat{V}_{\sigma^2[ITT]} \xrightarrow{\mathbb{P}} \overline{v}$, where $\overline{v}$ is a real number larger than $V_{\sigma^2[ITT]}$ defined in the proof.

Theorem 1 shows that in the "large S fixed n_s" asymptotic sequence we consider, $\widehat{\sigma}^2 [\text{ITT}]$ is asymptotically normal for $\sigma^2 [\text{ITT}]$, and $\widehat{V}_{\sigma^2[\text{ITT}]}$ is a conservative estimator of its asymptotic variance. Thus, Theorem 1 can be used to obtain conservative confidence intervals for $\sigma^2 [\text{ITT}]$.

Application: the variance across sites of the ITT effects of publicly- and privately-provided counseling. In Table 1, we start by estimating the ITT effect of each treatment. On average across all sites, both programs increase job seekers' employment rate after six months by slightly less than two percentage points (pp), an effect that is insignificant for the public

program.[6] However, this hides very substantial heterogeneity across sites. $\widehat{\sigma}^2\left[\text{ITT}\right]$ is large and significantly different from zero for the private program. For the public program, $\sigma^2\left[\text{ITT}\right]$ is not significantly different from zero at conventional levels (p-value=0.106), but this is due to the larger standard error of $\widehat{\sigma}^2\left[\text{ITT}\right]$ for that treatment, not to a lower point estimate. In each local office, the majority of job seekers were assigned to the private program, so the effects of the public program are less precisely estimated. $\sqrt{\widehat{\sigma}^2\left[\text{ITT}\right]}/\widehat{\text{ITT}} = 469\%$ for the public program, and 389% for the private one. This is a very substantial amount of treatment effect heterogeneity. For instance, assuming for illustrative purposes that site-specific ITTs follow a truncated normal,[7] where the underlying untruncated distribution has a mean equal to $\widehat{\text{ITT}}$ and a standard deviation equal to $\sqrt{\widehat{\sigma}^2\left[\text{ITT}\right]}$, the public program has a negative effect in 42% of the sites, while the private program has a negative effect in 40% of them. We also re-estimate the variance of the ITT effects of the public program using the estimator of Kline et al. (2020), in the special case described in their Example 2 with a single binary regressor, in which case the target parameter coincides with $\sigma^2\left[\text{ITT}\right]$. Doing so, we obtain an estimator around 20% smaller than our estimator, thus showing that the two approaches do not coincide after some relabeling.[8]

[6] In the paper, that effect is significant at the 10% level, owing to the slightly different estimation sample.

[7] The outcome is binary so ITTs have to belong to $[-1, 1]$.

[8] In our calculations, we divided $\tilde{z}_i$ by T_g in their covariance representation equation page 1868, as we interpreted the missingness of T_g as a typo. Without that change, their variance estimator is around 50 times smaller than ours.

Table 1: Estimating the variance across sites of the ITT effect of counseling on job seekers' probability of having a job after 6 months

	$\widehat{\text{ITT}}$	$\widehat{\sigma}^2\left[\text{ITT}\right]$	$\sqrt{\widehat{\sigma}^2\left[\text{ITT}\right]}/\widehat{\text{ITT}}$	N
	(1)	(2)	(3)	(4)
Public Counseling	0.017	0.0061	4.688	7,198
	(0.012)	(0.0038)		
Private Counseling	0.019	0.0057	3.894	34,768
	(0.009)	(0.0022)		

Results are based on data from the RCT in Behaghel et al. (2014). The outcome variable is an indicator equal to 1 if the jobseeker holds a job 6 months after the randomization. In Column (1), we estimate the average ITT effect across sites, with a robust standard error in parentheses beneath it. In Column (2), we compute $\widehat{\sigma}^2\left[\text{ITT}\right]$, the estimator of the variance of ITT effects across sites, with a robust standard error in parentheses beneath it, computed following Theorem 1. In Column (3), we show $\sqrt{\widehat{\sigma}^2\left[\text{ITT}\right]}/\widehat{\text{ITT}}$. Our estimation sample slightly differs from that in the paper: PESs with less than two treated or two control units have to be dropped from our analysis. The estimation is weighted, using the weights of the paper.

4.2 Predicting site-specific ITT and FS effects

Target parameter. Let $\mathbf{X}_s$ denote a $K \times 1$ vector of site-level variables, which we want to use to predict sites' ITTs. $\mathbf{X}_s$ may include observed variables, like some baseline covariates of site s. $\mathbf{X}_s$ may also include unobserved variables that have to be estimated. Let $\mathbf{I}_K$ denote the $K \times K$ identity matrix. Assuming that $\sigma^2[\mathbf{X}] + \lambda \mathbf{I}_K$ is invertible, our main target is

$$\beta_X^{\text{ITT}}(\lambda) \equiv \left(\sigma^2[\mathbf{X}] + \lambda \mathbf{I}_K\right)^{-1} \sum_{s=1}^{S} w_s \left(\mathbf{X}_s - \mu(\mathbf{X})\right) \left(\text{ITT}_s - \text{ITT}\right),$$

the coefficients on $\mathbf{X}_s$ in a Ridge regression of the demeaned ITT_s on the demeaned $\mathbf{X}_s$, weighted by w_s, and with hyper-parameter λ. $\beta_X^{\text{ITT}}(0)$ is a standard OLS regression coefficient, denoted β_X^{ITT}. When $\lambda = 0$, an auxiliary target is

$$\text{R}_X^{\text{ITT}} \equiv \frac{\left(\beta_X^{\text{ITT}}\right)^T \sigma^2[\mathbf{X}] \beta_X^{\text{ITT}}}{\sigma^2[\text{ITT}]},$$

the R-squared of the OLS regression.

Leading examples of unobserved variables one might want to include in $\mathbf{X}_s$. We have three leading examples in mind of potentially interesting unobserved variables one might want to include in $\mathbf{X}_s$. The first one is FS_s, the first-stage effect in site s. For instance, one can use the regression of ITT_s on FS_s to test the null that LATEs do not vary across sites: this null holds if and only if the regression's intercept is equal to zero while its R-squared is equal to one. The second unobserved variable one might want to include in $\mathbf{X}_s$ is $E(Y_s^r(0))$, the average outcome in the control group. Regressing ITT_s on $E(Y_s^r(0))$ is a way to assess if ITTs are larger or lower in sites with the lowest control outcomes, to assess if treatment offers reduce or increase inequalities across sites. The third one is $\mathbf{ITT}_{\text{M},s}$, the site-specific ITT effects on mediator variables. Regressing ITT_s on $\mathbf{ITT}_{\text{M},s}$ is a way to do "predictive mediation" analysis, by assessing if sites with large effects on the mediators also tend to have large effects on the final outcome. Of course, this type of mediation analysis remains predictive and not causal: larger effects in sites with larger mediator effects could be due to omitted variables rather than the mediator themselves.

Unbiased estimators of $\mathbf{X}_s$. As explained above, $\mathbf{X}_s$ may include unobserved variables, that need to be estimated. Then, we assume that we have an unbiased estimator of $\mathbf{X}_s$, denoted $\widehat{\mathbf{X}}_s$, that is a function of $((D_{is}(0), D_{is}(1), Y_{is}(0), Y_{is}(1), \mathbf{M}_{is}(0), \mathbf{M}_{is}(1))_{i\in\{1,...,n_s\}}, \mathbf{Z}_s)$ and known real numbers. Of course, for all coordinates $X_{k,s}$ of $\mathbf{X}_s$, that are observed and do not need to be estimated, $\widehat{X}_{k,s} = X_{k,s}$, so $\widehat{X}_{k,s}$ is non-stochastic. We let $\mu(\mathbf{X}) = \sum_{s=1}^S w_s \mathbf{X}_s$ and $\widehat{\mu}(\mathbf{X}) = \sum_{s=1}^S w_s \widehat{\mathbf{X}}_s$. Letting $\widehat{X}_{k,s}$ denote the kth coordinate of $\widehat{\mathbf{X}}_s$, we assume that for all $k \in \{1, ..., K\}$ we also have unbiased estimators of $\text{Cov}\left(\widehat{X}_{k,s}, \widehat{\text{ITT}}_s\right)$, denoted $\widehat{\text{Cov}}\left(\widehat{X}_{k,s}, \widehat{\text{ITT}}_s\right)$, and we let $\widehat{\text{Cov}}\left(\widehat{\mathbf{X}}_s, \widehat{\text{ITT}}_s\right)$ denote a vector stacking those estimators. Finally, we assume that we have an unbiased estimator of $V\left(\widehat{\mathbf{X}}_s\right)$, denoted $\widehat{V}\left(\widehat{\mathbf{X}}_s\right)$. The next lemma shows that those conditions are satisfied in our three leading examples.

Lemma 1 *If Assumptions 1 and 2 hold,*

1. $E\left(\widehat{FS}_s\right) = FS_s$, $E\left(\frac{c_{D,Y,0,s}}{n_{0,s}} + \frac{c_{D,Y,1,s}}{n_{1,s}}\right) = Cov\left(\widehat{FS}_s, \widehat{ITT}_s\right)$, *and* $E\left(\frac{r^2_{D,0,s}}{n_{0,s}} + \frac{r^2_{D,1,s}}{n_{1,s}}\right) = V\left(\widehat{FS}_s\right)$.
2. $E\left(\overline{Y}_{0s}\right) = E(Y_s^r(0))$, $E\left(-\frac{r^2_{Y,0,s}}{n_{0,s}}\right) = Cov\left(\overline{Y}_{0s}, \widehat{ITT}_s\right)$, *and* $E\left(\frac{r^2_{Y,0,s}}{n_{0,s}}\right) = V\left(\overline{Y}_{0s}\right)$.

3. $E\left(\widehat{\mathbf{ITT}}_{M,s}\right) = \mathbf{ITT}_{M,s}$, *for all* $k \in \{1, ..., K\}$ $E\left(\frac{c_{M_k,Y,0,s}}{n_{0,s}} + \frac{c_{M_k,Y,1,s}}{n_{1,s}}\right) = Cov\left(\widehat{ITT}_{M_k,s}, \widehat{ITT}_s\right)$, *and for all* $(k, k') \in \{1, ..., K\}^2$ $E\left(\frac{c_{M_k,M_{k'},0,s}}{n_{0,s}} + \frac{c_{M_k,M_{k'},1,s}}{n_{1,s}}\right) = Cov\left(\widehat{ITT}_{M_k,s}, \widehat{ITT}_{M_{k'},s}\right)$.

Consistent and asymptotically normal estimator of $\beta_X^{\mathbf{ITT}}(\lambda)$. We let

$$\widehat{\beta}_X^{\mathrm{ITT}}(\lambda) = \left(\sigma^2\left[\widehat{\mathbf{X}}\right] - \sum_{s=1}^{S} w_s \widehat{V}\left(\widehat{\mathbf{X}}_s\right) + \lambda \mathbf{I}_K\right)^{-1} \left(\sum_{s=1}^{S} w_s \left(\left(\widehat{\mathbf{X}}_s - \widehat{\mu}(\mathbf{X})\right)\left(\widehat{\mathrm{ITT}}_s - \widehat{\mathrm{ITT}}\right) - \widehat{\mathrm{Cov}}\left(\widehat{\mathbf{X}}_s, \widehat{\mathrm{ITT}}_s\right)\right)\right).$$

Without the $\widehat{V}\left(\widehat{\mathbf{X}}_s\right)$ and $\widehat{\mathrm{Cov}}\left(\widehat{\mathbf{X}}_s, \widehat{\mathrm{ITT}}_s\right)$ terms, $\widehat{\beta}_X^{\mathrm{ITT}}(\lambda)$ would just be the coefficient on $\widehat{\mathbf{X}}_s$ in a Ridge regression of the demeaned $\widehat{\mathrm{ITT}}_s$ on the demeaned $\widehat{\mathbf{X}}_s$. Those terms account for the fact that $\widehat{\mathbf{X}}_s$ is unbiased but not consistent for $\mathbf{X}_s$. Similarly, when $\lambda = 0$, we let

$$\widehat{\mathrm{R}}_X^{\mathrm{ITT}} = \frac{\left(\widehat{\beta}_X^{\mathrm{ITT}}\right)^T \widehat{\sigma}^2[\mathbf{X}]\widehat{\beta}_X^{\mathrm{ITT}}}{\widehat{\sigma}^2[\mathrm{ITT}]}$$

denote the estimator of $\mathrm{R}_X^{\mathrm{ITT}}$. Then, let

$$\begin{aligned}
A(\lambda) =& \sum_{s=1}^{S} w_s \left(\mathbf{X}_s - \mu(\mathbf{X})\right)\left(\mathbf{X}_s - \mu(\mathbf{X})\right)^T + \lambda \mathbf{I}_K \\
B =& \sum_{s=1}^{S} w_s \left(\mathbf{X}_s - \mu(\mathbf{X})\right)\left(\mathrm{ITT}_s - \mathrm{ITT}\right) \\
\widehat{A}(\lambda) =& \sum_{s=1}^{S} w_s \left(\left(\widehat{\mathbf{X}}_s - \widehat{\mu}(\mathbf{X})\right)\left(\widehat{\mathbf{X}}_s - \widehat{\mu}(\mathbf{X})\right)^T - \widehat{V}\left(\widehat{\mathbf{X}}_s\right)\right) + \lambda \mathbf{I}_K \\
\widehat{B} =& \sum_{s=1}^{S} w_s \left(\left(\widehat{\mathbf{X}}_s - \widehat{\mu}(\mathbf{X})\right)\left(\widehat{\mathrm{ITT}}_s - \widehat{\mathrm{ITT}}\right) - \widehat{\mathrm{Cov}}\left(\widehat{\mathbf{X}}_s, \widehat{\mathrm{ITT}}_s\right)\right),
\end{aligned}$$

$$\begin{aligned}
\phi_{s,2} =& \tilde{w}_s \left(\left(\widehat{\mathbf{X}}_s - \mu(\mathbf{X})\right)\left(\widehat{\mathbf{X}}_s - \mu(\mathbf{X})\right)^T - \widehat{V}\left(\widehat{\mathbf{X}}_s\right)\right) + \lambda \mathbf{I}_K \\
\phi_{s,3} =& \tilde{w}_s \left(\left(\widehat{\mathbf{X}}_s - \mu(\mathbf{X})\right)\left(\widehat{\mathrm{ITT}}_s - \mathrm{ITT}\right) - \widehat{\mathrm{Cov}}\left(\widehat{\mathbf{X}}_s, \widehat{\mathrm{ITT}}_s\right)\right) \\
\phi_{s,4} =& -\left[A(\lambda)\right]^{-1} \phi_{s,2} \left[A(\lambda)\right]^{-1} B + \left[A(\lambda)\right]^{-1} \phi_{s,3},
\end{aligned}$$

and let $V_{\beta_X^{\mathrm{ITT}}(\lambda)}$ denote the limit of $\frac{1}{S}\sum_{s=1}^{S} V\left(\phi_{s,4}\right)$, which is assumed to exist in Assumption 7 in the Web Appendix.

Theorem 2 *Suppose that Assumptions 1 and 2 hold, and that the technical conditions in Assumption 7 in the Web Appendix hold. Then,*

$$\widehat{\beta}_X^{ITT}(\lambda) - \beta_X^{ITT}(\lambda) \xrightarrow{\mathbb{P}} 0,$$

and

$$\sqrt{S}\left(\widehat{\beta}_X^{ITT}(\lambda) - \beta_X^{ITT}(\lambda)\right) \xrightarrow{d} N(0, V_{\beta_X^{ITT}(\lambda)}).$$

Let

$$\begin{aligned}
\widehat{\phi}_{s,4} =& -\left[\widehat{A}(\lambda)\right]^{-1} \widehat{\phi}_{s,2} \left[\widehat{A}(\lambda)\right]^{-1} \widehat{B} + \left[\widehat{A}(\lambda)\right]^{-1} \widehat{\phi}_{s,3} \\
\widehat{\phi}_{s,2} =& \tilde{w}_s \left(\left(\widehat{\mathbf{X}}_s - \widehat{\mu}(\mathbf{X})\right)\left(\widehat{\mathbf{X}}_s - \widehat{\mu}(\mathbf{X})\right)^T - \widehat{V}\left(\widehat{\mathbf{X}}_s\right)\right) + \lambda \mathbf{I}_K \\
\widehat{\phi}_{s,3} =& \tilde{w}_s \left(\left(\widehat{\mathbf{X}}_s - \widehat{\mu}(\mathbf{X})\right)\left(\widehat{\mathrm{ITT}}_s - \widehat{\mathrm{ITT}}\right) - \widehat{\mathrm{Cov}}\left(\widehat{\mathbf{X}}_s, \widehat{\mathrm{ITT}}_s\right)\right).
\end{aligned}$$

We conjecture that using similar steps as in the proof of Theorem 1, one can show that $\widehat{V}_{\beta_X^{\mathrm{ITT}}(\lambda)}$, the sample variance of $\widehat{\phi}_{s,4}$, is a conservative estimator of $V_{\beta_X^{\mathrm{ITT}}(\lambda)}$.[9]

Choice of hyper-parameter. Golub et al. (1979) propose to use a generalized cross-validation (GCV) method to choose λ. Applying their Equation (1.4) to our multi-site RCT setting, rewriting explicitly the inner product in the numerator and using the linearity and cyclicality of the trace operator to rewrite the denominator, GCV amounts to using λ^*, the minimizer of

$$V(\lambda) = \frac{\sigma^2[ITT] + B'\left([A(\lambda)]^{-1}\sigma^2[\mathbf{X}][A(\lambda)]^{-1} - 2[A(\lambda)]^{-1}\right)B}{\left(1 - \frac{1}{S}\mathrm{Tr}\left([A(\lambda)]^{-1}\sigma^2[\mathbf{X}]\right)\right)^2}, \tag{3}$$

where Tr(.) denotes the trace operator. (3) makes it clear that for any λ, $V(\lambda)$ can be consistently estimated, replacing $\sigma^2[ITT]$, B, $A(\lambda)$, and $\sigma^2[\mathbf{X}]$ by their estimators. Accordingly, we propose to use $\widehat{\lambda}^*$, the minimizer of $\widehat{V}(\lambda)$. While it should be feasible to derive the asymptotic variance of $\widehat{\beta}_X^{\mathrm{ITT}}\left(\widehat{\lambda}^*\right)$ using standard results from M-estimation, for now we rely on the bootstrap.

Application: predicting site-specific ITT effects of the publicly- and privately-provided counseling programs. Table 2 reports several univariate OLS regressions of sites ITT effects on predictors. Future versions of this paper will also show some ridge regressions. In Column (1), we regress sites' ITTs on their FSs. While FSs varies across sites (sd = 12 pp for both programs, see Table 4 below), FSs are not significantly correlated with ITTs. In Column (2), we investigate if the heterogeneity in ITT effects could be due to differences in the

[9]For a vector, a conservative variance estimator means that for any $K \times 1$ vector of real numbers θ, $\theta' \widehat{V}_{\beta_X^{\mathrm{ITT}}(\lambda)}\theta$ converges to a limit weakly larger than that of $\theta' V_{\beta_X^{\mathrm{ITT}}(\lambda)}\theta$.

populations of job seekers across sites. For that purpose, we regress sites' ITTs on the average employability of their job seekers.[10] While employability varies across sites (sd = 3.5 pp), it is not correlated with sites' ITTs. In Column (3), we investigate if the heterogeneity in ITT effects could be due to differences in the local labor market conditions across sites. For that purpose, we regress ITTs on sites' local unemployment rate, which we could retrieve for all but one site.[11] Again, while the unemployment rate varies across sites (sd = 4.4 pp), it is not correlated with sites' ITTs. In Column (4), we find that the ITT effects of the public and private programs are strongly negatively correlated with the control group job finding rate, which might indicate that counseling is more effective in less tight labor markets. The regressions in Column (3) could fail to detect that, because the local unemployment rate might be an imperfect proxy of the labor market conditions faced by the job seekers eligible for this RCT, namely those at high risk of long-term employment, and the average job finding rate of the experimental control group may be a better proxy. This correlation may also be used to better target the program, using for instance the average job finding rate of an earlier cohort of job seekers in each site as a proxy for $E(Y_s^r(0))$. Finally, in Column (5) of Panel A, we find a very strong positive correlation between the ITTs of the public and private programs. In each site, the two programs are delivered by different providers. Therefore, this suggests that the heterogeneity in sites' ITT effects is unlikely to be entirely driven by providers' effects.[12]

Application: comparing our regression coefficients $\widehat{\beta}_X^{\mathbf{ITT}}$ to naive ones. At the bottom of each column of Table 2, we show naive OLS regression coefficients of ITTs on characteristics, that do not account for the estimation of the independent variable, and of the explanatory variable for some of the regressions. When the explanatory variable is estimated (Columns (1), (4), and (5)), the naive regression leads to a point estimate that differs from $\widehat{\beta}_X^{\text{ITT}}$, and to much smaller standard errors. When the characteristic is not estimated (Columns (2), (3)), the naive regression leads to identical point estimates and only very slightly different standard errors.

[10] Employability is a job seeker's probability of finding a job in less than six months, according to a logistic regression with a rich set of covariates estimated in the control group by Behaghel et al. (2014).

[11] Specifically, we matched the data of Behaghel et al. (2014) to a dataset produced by the French National Office of Statistics, with unemployment rates at the city level in 2007, the year when the RCT was conducted.

[12] Of course, this claim remains suggestive, as one cannot rule out a perfect correlation between the quality of the local public and private providers, but that scenario does not sound very plausible.

Table 2: Predicting site-specific ITTs

Panel A: Public Counseling					
	FS_s	Employability	Local Unemp Rate	$E(Y_s^r(0))$	ITT_s^{priv}
	(1)	(2)	(3)	(4)	(5)
$\widehat{\beta}_X^{ITT}$	-0.113	0.266	-0.089	-0.626	1.043
	(0.213)	(0.391)	(0.271)	(0.206)	(0.282)
$\widehat{R}_X^{ITT}$	0.025	0.015	0.002	0.451	1.063
Naive estimator	-0.026	0.266	-0.089	-0.848	0.768
	(0.084)	(0.393)	(0.271)	(0.080)	(0.071)
Number of sites	200	200	199	200	200
Panel B: Private Counseling					
	FS_s	Employability	Local Unemp Rate	$E(Y_s^r(0))$	
	(1)	(2)	(3)	(4)	
$\widehat{\beta}_X^{ITT}$	-0.048	0.399	0.089	-0.854	
	(0.094)	(0.295)	(0.242)	(0.090)	
$\widehat{R}_X^{ITT}$	0.005	0.032	0.003	0.957	
Naive estimator	-0.035	0.399	0.089	-0.939	
	(0.073)	(0.296)	(0.243)	(0.036)	
Number of sites	204	204	203	204	

Results are based on data from the RCT in Behaghel et al. (2014). In Panel A, we estimate univariate regressions of the site-level ITTs of the public counseling program on the following site-level variables: the program take-up rate, job seekers average employability, the local unemployment rate, job seekers' job finding rate without the program, and the ITT effect of the private counseling program. Panel B shows the same regressions, except for the last one, for the ITT effects of the private program. The estimator $\widehat{\beta}_X^{ITT}$ and it standard error are computed as described in the text. The naive estimator and its standard error are computed by running a linear regression of the ITTs on the site-level variable under consideration, using robust standard errors. The estimation is weighted, using the weights of the paper.

Application: regressing the job finding rate in the public counseling group on the job finding rates in the control group and in the private counseling group. Table 2 shows that both the job finding rate in the control group and the ITT of the private program

predict the ITT of the public one. We now assess if the job finding rates in the control group and in the private counseling group significantly predict the job finding rate in the public counseling group. Table 3 below shows that both coefficients are positive and significant.

Table 3: Regressing the job finding rate in the public counseling group on the job finding rates in the control group and in the private counseling group

	$E(Y_s^r(0))$	$E(Y_s^r(\text{priv}))$
	(1)	(2)
$\widehat{\beta}_X^{E(Y_s^r(\text{pub}))}$	0.322	0.508
	(0.191)	(0.205)
Naive estimator	0.156	0.295
	(0.094)	(0.164)
Number of sites	200	200

Results are based on data from the RCT in Behaghel et al. (2014). We estimate a multivariate regression of the job finding rate in the public counseling group on the job finding rates in the control group and in the private counseling group. The estimator $\widehat{\beta}_X^{\text{ITT}}$ and it standard error are computed as described in the text. The naive estimator and its standard error are computed by running a linear regression, using heteroscedasticity-robust standard errors. The estimation is weighted, using the weights of the paper.

5 Estimating and predicting LATEs' heterogeneity.

Studying LATEs' heterogeneity when FSs are homogeneous. If $\sigma^2[\text{FS}] = 0$, then $\text{LATE}_s = \text{ITT}_s/\text{FS}$, and to study LATEs' heterogeneity one can merely study ITTs' heterogeneity using the techniques presented in the previous section.

Application: the publicly- and privately-provided counseling programs have heterogeneous FS effects across sites. Table 4 shows that in Behaghel et al. (2014), first-stage effects vary across sites, both for the public and for the private program. The estimated standard deviation of FS effects is around 12 pp, namely 38% of the average FS effect of the public program, and around 30% of the average FS effect of the private program.

Table 4: Estimating the variance across sites of the FS effect of receiving an offer for the counseling programs

	$\widehat{\text{FS}}$	$\widehat{\sigma}^2[\text{FS}]$	$\sqrt{\widehat{\sigma}^2[\text{FS}]}/\widehat{\text{FS}}$	N
	(1)	(2)	(3)	(4)
Public Counseling	0.312	0.012	0.379	7,198
	(0.007)	(0.003)		
Private Counseling	0.402	0.013	0.305	34,768
	(0.004)	(0.002)		

Results are based on data from the RCT in Behaghel et al. (2014). The outcome variable is an indicator equal to 1 if the jobseeker enrolled for the public (resp. private) counseling program. In Column (1), we estimate the average FS effect across sites, with a robust standard error in parentheses beneath it. In Column (2), we compute $\widehat{\sigma}^2[\text{FS}]$, the estimator of the variance of FS effects across sites, with a robust standard error in parentheses beneath it, computed following Theorem 1. In Column (3), we show $\sqrt{\widehat{\sigma}^2[\text{FS}]}/\widehat{\text{FS}}$. The estimation is weighted, using the weights of the paper.

Studying LATEs' heterogeneity with heterogeneous FS effects. In the remainder of this section, we propose techniques to estimate the variance of LATEs across sites, or the coefficient from a regression of LATE_s on some covariates, allowing for heteogeneous FS effects. Doing so is more difficult than in the ITT case, because unlike $\widehat{\text{ITT}}_s$, $\widehat{\text{LATE}}_s$ is not unbiased. Thus, we will propose testable assumptions under which our targets can be written as functions of ITT_s, FS_s, and other objects that can be unbiasedly estimated.

5.1 Estimating the variance of LATEs across sites.

Target parameter. In this section, our target parameter is

$$\sigma^2[\text{LATE}] \equiv \sum_{s=1}^{S} \frac{w_s \text{FS}_s}{\text{FS}} [\text{LATE}_s - \text{LATE}]^2,$$

a weighted variance of LATEs where the weight assigned to site s corresponds to the weight assigned to that site in LATE.[13]

[13]With a slight abuse of notation, we keep the same $\sigma^2[.]$ notation as in the previous section, despite the difference in the weights.

Identification of σ^2 [LATE] assuming independent FSs and LATEs.

Assumption 5 *For any functions f and g,*

$$\sum_{s=1}^{S} w_s f(LATE_s) \times g(FS_s) = \left(\sum_{s=1}^{S} w_s f(LATE_s)\right) \times \left(\sum_{s=1}^{S} w_s g(FS_s)\right).$$

Assumption 5 requires that sites' FSs and LATEs be independent.

Theorem 3 *If Assumption 5 holds, then*

$$\sigma^2 [LATE] = \frac{\sum_{s=1}^{S} w_s (ITT_s - FS_s \times LATE)^2}{\sum_{s=1}^{S} w_s FS_s^2}.$$

Once noted that under Assumption 5,

$$\sigma^2 [\text{LATE}] = \sum_{s=1}^{S} w_s [\text{LATE}_s - \text{LATE}]^2,$$

Theorem 3 follows from applying Assumption 5 to $f(x) = (x - \text{LATE})^2$ and $g(x) = x^2$.

A lower bound for σ^2 [LATE] based on Theorem 3. Heuristically, let

$$\nu_s = \text{ITT}_s - \text{FS}_s \text{LATE}.$$

As $\sum_{s=1}^{S} w_s \nu_s = 0$, the numerator of σ^2 [LATE] in Theorem 3 is equal to the variance of ν_s across sites. Then, one may be able to show that an EB variance estimator with outcome variable

$$\widehat{\nu}_{is} = Y_{is} - D_{is} \times \widehat{\text{LATE}}$$

converges to the same limit as $\sum_{s=1}^{S} w_s (\text{ITT}_s - \text{FS}_s \times \text{LATE})^2$. As Jensen's inequality implies that $\sum_{s=1}^{S} w_s \widehat{\text{FS}}_s^2$ converges to a limit larger than that of $\sum_{s=1}^{S} w_s \text{FS}_s^2$, dividing this EB estimator by $\sum_{s=1}^{S} w_s \widehat{\text{FS}}_s^2$ would yield an estimator of a lower bound of σ^2 [LATE]. Let

$$\begin{aligned}
\widehat{\nu}_s =& \widehat{\text{ITT}}_s - \widehat{\text{FS}}_s \times \widehat{\text{LATE}} \\
\tilde{\nu}_s =& \widehat{\text{ITT}}_s - \widehat{\text{FS}}_s \times \text{LATE} \\
\widehat{V}_{rob}(\widehat{\nu}_s) =& \frac{1}{n_{1s}} r^2_{\widehat{\nu},1,s} + \frac{1}{n_{0s}} r^2_{\widehat{\nu},0,s} \\
\widehat{V}_{rob}(\tilde{\nu}_s) =& \frac{1}{n_{1s}} r^2_{\tilde{\nu},1,s} + \frac{1}{n_{0s}} r^2_{\tilde{\nu},0,s}.
\end{aligned}$$

Let

$$\phi_{s,5} = \frac{\tilde{w}_s \tilde{\nu}_s}{\text{FS}},$$

and let

$$\phi_{s,6} = \frac{\tilde{w}_s \left((\tilde{\nu}_s)^2 - \widehat{V}_{rob}\left(\tilde{\nu}_s\right) \right) - 2(C_1 + C_2)\phi_{s,5} - \tilde{w}_s \widehat{FS}_s^2 C_3}{C_4},$$

where C_1, C_2, C_3, and C_4 respectively denote the limits of $\frac{1}{S}\sum_{s=1}^{S} \tilde{w}_s E(\widehat{\text{FS}}_s \tilde{\nu}_s)$, $\frac{1}{S}\sum_{s=1}^{S} \tilde{w}_s E\left(\frac{\text{LATE}\times r_{D,1,s}^2 - c_{D,Y,1,s}}{n_{1s}} + \frac{\text{LATE}\times r_{D,0,s}^2 - c_{D,Y,0,s}}{n_{0s}}\right)$, $\underline{\sigma}^2[\text{LATE}]$, and $\frac{1}{S}\sum_{s=1}^{S} \tilde{w}_s E\left(\widehat{\text{FS}}_s^2\right)$, which are assumed to exist in Assumption 6 below. Let $V_{\sigma^2[\text{LATE}]}$ denote the limit of $\frac{1}{S}\sum_{s=1}^{S} V\left(\phi_{s,6}\right)$, which is also assumed to exist below. Finally, let

$$\begin{aligned}
\widehat{\underline{\sigma}}^2[\text{LATE}] =& \frac{\frac{1}{S}\sum_{s=1}^{S} \tilde{w}_s \left[(\widehat{\nu}_s)^2 - \widehat{V}_{rob}\left(\widehat{\nu}_s\right) \right]}{\frac{1}{S}\sum_{s=1}^{S} \tilde{w}_s \widehat{\text{FS}}_s^2} \\
\underline{\sigma}^2[\text{LATE}] =& \frac{\frac{1}{S}\sum_{s=1}^{S} \tilde{w}_s \left[E((\tilde{\nu}_s)^2) - E\left(\widehat{V}_{rob}(\tilde{\nu}_s)\right) \right]}{\frac{1}{S}\sum_{s=1}^{S} \tilde{w}_s E\left(\widehat{\text{FS}}_s^2\right)}.
\end{aligned}$$

Assumption 6 *1. The sequence $(\phi_{s,6})_{s\geq 1}$ satisfies the Lyapunov condition.*

2. *The limits of the following sequences exist: i) $\frac{1}{S}\sum_{s=1}^{S} \tilde{w}_s E\left(\widehat{FS}_s^2\right)$; ii) $\frac{1}{S}\sum_{s=1}^{S} \tilde{w}_s E(\widehat{FS}_s \tilde{\nu}_s)$; iii) $\underline{\sigma}^2[LATE]$; iv) $\frac{1}{S}\sum_{s=1}^{S} \tilde{w}_s E\left(\frac{LATE\times r_{D,1,s}^2 - c_{D,Y,1,s}}{n_{1s}} + \frac{LATE\times r_{D,0,s}^2 - c_{D,Y,0,s}}{n_{0s}}\right)$; v) $\frac{1}{S}\sum_{s=1}^{S} \tilde{w}_s FS_s^2$; vi) $\frac{1}{S}\sum_{s=1}^{S} V\left(\phi_{s,6}\right)$; vii) $\sigma^2[LATE]$.*

3. $\lim_{S\to+\infty} \frac{1}{S}\sum_{s=1}^{S} \tilde{w}_s E\left(\widehat{FS}_s^2\right) > 0$.

Theorem 4 *Suppose that Assumptions 1-6 hold. Then,*

$$\lim_{S\to+\infty} \underline{\sigma}^2[LATE] \leq \lim_{S\to+\infty} \sigma^2\left[LATE\right],$$

and

$$\sqrt{S}(\widehat{\underline{\sigma}}^2[LATE] - \underline{\sigma}^2[LATE]) \xrightarrow{d} N(0, V_{\sigma^2[LATE]}).$$

Theorem 4 shows that under Assumption 5, $\widehat{\underline{\sigma}}^2[\text{LATE}]$ is an asymptotically normal estimator of a lower bound of $\sigma^2\left[\text{LATE}\right]$. We conjecture that using similar steps as in the proof of Theorem 1, one can show that the sample variance of $\widehat{\phi}_{s,6}$, a variable where all the population quantites in $\phi_{s,6}$ are replaced by their sample equivalents, converges to a limit weakly larger than $V_{\sigma^2[\text{LATE}]}$, and can thus be used as a conservative variance estimator.

Application: estimating the variance of the LATEs of the publicly- and privately-provided counseling programs. Though we do not report them, naive EB estimators using site-specific 2SLS estimators as building blocks are negative and therefore uninformative on the LATEs' variance, as was already found by Walters (2015) in a different context. Instead, in Table 5 we estimate our lower bound for the variance of LATEs across sites, under Assumption 5. We find evidence of very heterogeneous LATEs across sites for both programs. Our lower bound on LATEs' standard deviation across sites is equal to 377% of the LATE estimate for the private program, and to 314% of the LATE estimate for the public one. Our lower bound is significantly different from zero for both programs (at the 10% level for the public program).

Table 5: Lower bound for the variance of LATEs across sites

	$\widehat{LATE}$	$\underline{\widehat{\sigma}}^2[\text{LATE}]$	$\sqrt{\underline{\widehat{\sigma}}^2[\text{LATE}]}/\widehat{LATE}$	N
	(1)	(2)	(3)	(4)
Public Counseling	0.070	0.0486	3.138	7,198
	(0.040)	(0.0296)		
Private Counseling	0.048	0.0326	3.773	34,768
	(0.023)	(0.0125)		

Results are based on data from the RCT in Behaghel et al. (2014). In Column (1), we show the average LATE effect across sites, with a robust standard error in parentheses beneath it. In Column (2), we show an estimator of the variance of LATE effects across sites and a robust standard error in parentheses beneath it, both computed following Theorem 4. In Column (3), we show the estimated standard deviation of LATEs divided by $\widehat{LATE}$. The estimation is weighted, using the weights of the paper.

An upper bound for σ^2 [LATE]. While upper bounding σ^2 [LATE] is of less interest than lower bounding it, it still worth noting that under Assumption 5, it follows from the law of total variance that

$$\sigma^2[\text{ITT}] = \sigma^2\left[\text{LATE}\right]\sum_{s=1}^{S} w_s \text{FS}_s^2 + \sigma^2\left[\text{FS}\right]\text{LATE}^2,$$

thus implying that

$$\sigma^2\left[\text{LATE}\right] = \frac{\sigma^2[\text{ITT}] - \sigma^2[\text{FS}]\text{LATE}^2}{\sum_{s=1}^{S} w_s \text{FS}_s^2} \leq \frac{\sigma^2[\text{ITT}] - \sigma^2[\text{FS}]\text{LATE}^2}{\text{FS}^2},$$

an upper bound that can be consistently estimated. The previous display also implies that

$$\sigma^2\left[\text{LATE}\right] \leq \frac{\sigma^2[\text{ITT}]}{\text{FS}^2},$$

so that

$$\sqrt{\sigma^2\left[\text{LATE}\right]}/\text{LATE} \leq \sqrt{\sigma^2[\text{ITT}]}/\text{ITT}:$$

LATEs cannot be more heterogeneous than ITTs under Assumption 5.

Testing Assumption 5. We now show that Assumption 5 has a testable implication.

Theorem 5 *Suppose that Assumptions 1- 3 hold. If Assumption 5 further holds,*

$$LATE = \beta_{FS}^{ITT}.$$

Then, to test Assumption 5, one just needs to estimate LATE and $\beta_{\text{FS}}^{\text{ITT}}$, and test that the two estimators are significantly different.

Application: the test that the LATEs and FSs effects of the publicly- and privately-provided counseling programs are independent is not rejected. In Table 6, we test the null that LATE = $\beta_{\text{FS}}^{\text{ITT}}$. The test is not rejected, either for the private or for the public program. In this application, the test is not very powerful, especially for the public program, and it could fail to detect meaningful differences between the two parameters.

Table 6: Testing if sites' first-stage and LATE effects are independent

	$\widehat{\text{LATE}} - \widehat{\beta}_{\text{FS}}^{\text{ITT}}$	s.e.	N
	(1)	(2)	(3)
Public Counseling	0.160	(0.207)	7,198
Private Counseling	0.096	(0.099)	34,768

Results are based on data from the RCT in Behaghel et al. (2014). We follow Theorem 5 to test the assumption that sites' LATE and FS effects are not correlated. Column (1) shows $\widehat{\text{LATE}} - \widehat{\beta}_{\text{FS}}^{\text{ITT}}$, the test's statistic. Column (2) shows its standard error, obtained using linearizations of $\widehat{\beta}_{\text{FS}}^{\text{ITT}}$ and $\widehat{\text{LATE}}$ that can be found in the proofs. The estimation is weighted, using the weights of the paper.

5.2 Predicting site-specific LATEs

Target parameter. Let x_s be a scalar and binary site-level covariate. Letting $S_{x,1} = \sum_{s=1}^{S} x_s w_s$ and $S_{x,0} = \sum_{s=1}^{S}(1-x_s)w_s$, our target parameter is

$$\beta_x^{\text{LATE}} \equiv \frac{1}{S_{x,1}} \sum_{s=1}^{S} x_s w_s \text{LATE}_s - \frac{1}{S_{x,0}} \sum_{s=1}^{S} (1-x_s) w_s \text{LATE}_s,$$

the difference between the average LATEs of sites with $x_s = 1$ and $x_s = 0$. Let $\text{LATE}_{x,1}$ and $\text{LATE}_{x,0}$ respectively denote the average LATE across sites with $x_s = 1$ and $x_s = 0$.

Identification. Assume that for any functions f and g,

$$\begin{aligned}
\frac{1}{S_{x,0}} \sum_{s=1}^{S} (1-x_s) w_s f(\text{LATE}_s) \times g(\text{FS}_s) &= \left(\frac{1}{S_{x,0}} \sum_{s=1}^{S} (1-x_s) w_s f(\text{LATE}_s) \right) \times \left(\frac{1}{S_{x,0}} \sum_{s=1}^{S} (1-x_s) w_s g(\text{FS}_s) \right) \\
\frac{1}{S_{x,1}} \sum_{s=1}^{S} x_s w_s f(\text{LATE}_s) \times g(\text{FS}_s) &= \left(\frac{1}{S_{x,1}} \sum_{s=1}^{S} x_s w_s f(\text{LATE}_s) \right) \times \left(\frac{1}{S_{x,1}} \sum_{s=1}^{S} x_s w_s g(\text{FS}_s) \right), \qquad (4)
\end{aligned}$$

meaning that sites' FSs and LATEs are independent within the subsample of sites with $x_s = 1$ and within the subsample of sites with $x_s = 0$, a conditional version of Assumption 5.

Theorem 6 *If* (4) *holds, then*

$$\beta_x^{LATE} = \frac{\frac{1}{S_{x,1}} \sum_{s=1}^{S} x_s \left(ITT_s - FS_s LATE_{x,0} \right) - \frac{1}{S_{x,0}} \sum_{s=1}^{S} (1-x_s) \left(ITT_s - FS_s LATE_{x,1} \right)}{FS}.$$

To consistently estimate β_x^{LATE}, one can just replace ITT_s, FS_s, $\text{LATE}_{x,0}$, $\text{LATE}_{x,1}$ and FS by their estimators in the previous display.

6 Conclusion

In multi-site randomized controlled trials, with a large number of sites but few randomization units per site, an Empirical-Bayes (EB) estimator can be used to estimate the variance of the treatment effect across sites. We propose a consistent estimator of the coefficient from a ridge regression of site-level effects on site-level characteristics that are unobserved but can be unbiasedly estimated, such as sites' average outcome without treatment, or site-specific treatment effects on mediator variables. For instance, in a multi-site job-search counseling RCT, it can

be interesting to study whether sites that have the largest effects on job-seekers' job finding rate are also the sites that have the largest effect on their search effort, as a "predictive mediation analysis" of whether the job-finding effect can be "explained" by the job-search effect. In experiments with imperfect compliance, we also propose a non-parametric and partly testable assumption under which the variance of local average treatment effects (LATEs) across sites can be estimated. We revisit Behaghel et al. (2014), who study the effect of counseling programs on job seekers job-finding rate, in more than 200 job placement agencies in France. We find considerable treatment-effect heterogeneity, both for intention to treat and LATE effects, and the treatment effect is negatively correlated with sites' job-finding rate without treatment.

References

Adusumilli, K., F. Agostinelli, and E. Borghesan (2024). Heterogeneity and endogenous compliance: Implications for scaling class size interventions. Technical report, National Bureau of Economic Research.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association 91*(434), 444–455.

Angrist, N. and R. Meager (2023). Implementation matters: Generalizing treatment effects in education. edworkingpaper no. 23-802. *Annenberg Institute for School Reform at Brown University*.

Behaghel, L., B. Crépon, and M. Gurgand (2014). Private and public provision of counseling to job seekers: Evidence from a large controlled experiment. *American economic journal: applied economics 6*(4), 142–174.

De Chaisemartin, C. and X. d'Haultfoeuille (2018). Fuzzy differences-in-differences. *The Review of Economic Studies 85*(2), 999–1028.

Eicker, F. et al. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *The Annals of Mathematical Statistics 34*(2), 447–456.

Golub, G. H., M. Heath, and G. Wahba (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics 21*(2), 215–223.

Huber, P. J. et al. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 221–233. University of California Press.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), pp. 467–475.

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

Kline, P., R. Saggio, and M. Sølvsten (2020). Leave-out estimation of variance components. *Econometrica 88*(5), 1859–1898.

Li, X. and P. Ding (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association 112*(520), 1759–1769.

Liu, R. Y. et al. (1988). Bootstrap procedures under some non-iid models. *The Annals of Statistics 16*(4), 1696–1708.

Morris, C. N. (1983). Parametric empirical bayes inference: theory and applications. *Journal of the American statistical Association 78*(381), 47–55.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics 4*, 2111–2245.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9. translated in Statistical Science 5(4), 465-472, 1990.

Raudenbush, S. W. and H. S. Bloom (2015). Learning about and from a distribution of program impacts using multisite trials. *American Journal of Evaluation 36*(4), 475–499.

Rose, E. K., J. T. Schellenberg, and Y. Shem-Tov (2022). The effects of teacher quality on adult criminal justice contact. Technical report, National Bureau of Economic Research.

Stanley, T. D. and H. Doucouliagos (2012). *Meta-regression analysis in economics and business*. routledge.

Walters, C. R. (2015). Inputs in the production of early childhood human capital: Evidence from head start. *American Economic Journal: Applied Economics 7*(4), 76–102.

White, H. et al. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *econometrica 48*(4), 817–838.

Web Appendix, not for publication

7 Proofs

7.1 Proof of Theorem 1

Asymptotic normality.

Let

$$\tilde{\sigma}^2\left[\mathrm{ITT}\right]=\sum_{s=1}^{S}w_s\left[\left(\widehat{\mathrm{ITT}}_s-\mathrm{ITT}\right)^2-\widehat{V}_{rob}\left(\widehat{\mathrm{ITT}}_s\right)\right].$$

$$\begin{aligned}\sqrt{S}\left(\widehat{\sigma}^2\left[\mathrm{ITT}\right]-\tilde{\sigma}^2\left[\mathrm{ITT}\right]\right)=&\frac{1}{\sqrt{S}}\sum_{s=1}^{S}\tilde{w}_s\left[\left(\widehat{\mathrm{ITT}}_s-\widehat{\mathrm{ITT}}\right)^2-\left(\widehat{\mathrm{ITT}}_s-\mathrm{ITT}\right)^2\right]\\=&-\sqrt{S}\left(\widehat{\mathrm{ITT}}-\mathrm{ITT}\right)\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\left[2\widehat{\mathrm{ITT}}_s-\widehat{\mathrm{ITT}}-\mathrm{ITT}\right]\\=&-\sqrt{S}\left(\widehat{\mathrm{ITT}}-\mathrm{ITT}\right)\left[\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\widehat{\mathrm{ITT}}_s-\mathrm{ITT}\right]\\=&-\sqrt{S}\left(\widehat{\mathrm{ITT}}-\mathrm{ITT}\right)o_P(1)\\=&o_P(1).\end{aligned}\tag{5}$$

The fourth equality follows from the fact $\widehat{\mathrm{ITT}}$ is unbiased for ITT, from applying the law of large numbers in Lemma 1 of Liu et al. (1988) to the sequence of independent and bounded random variables $\tilde{w}_s\widehat{\mathrm{ITT}}_s$, and from Point 3 of Assumption 4. The fifth equality follows from applying the Lyapunov CLT to $\left(\tilde{w}_s\widehat{\mathrm{ITT}}_s\right)_{s\geq 1}$. Then, as

$$\begin{aligned}E\left(\phi_{s,1}\right)=&\tilde{w}_s\left[E\left(\left(\widehat{\mathrm{ITT}}_s-\mathrm{ITT}\right)^2\right)-E\left(\widehat{V}_{rob}\left(\widehat{\mathrm{ITT}}_s\right)\right)\right]\\=&\tilde{w}_s\left[E\left(\left(\widehat{\mathrm{ITT}}_s-\mathrm{ITT}_s\right)^2\right)+\left(\mathrm{ITT}_s-\mathrm{ITT}\right)^2-2\left(\mathrm{ITT}_s-\mathrm{ITT}\right)E\left(\widehat{\mathrm{ITT}}_s-\mathrm{ITT}_s\right)-V(\widehat{\mathrm{ITT}}_s)\right]\\=&\tilde{w}_s\left(\mathrm{ITT}_s-\mathrm{ITT}\right)^2,\end{aligned}$$

$$\sqrt{S}\left(\tilde{\sigma}^2\left[\mathrm{ITT}\right]-\sigma^2\left[\mathrm{ITT}\right]\right)=\frac{1}{\sqrt{S}}\sum_{s=1}^{S}\left(\phi_{s,1}-E(\phi_{s,1})\right).\tag{6}$$

The result follows from (5) and (6), from applying the Lyapunov CLT to $(\phi_{s,1})_{s\geq 1}$, and from the Slutsky lemma.

Asymptotically conservative variance estimator.

Let

$$\widehat{V}^I_{bound} = \frac{1}{S}\sum_{s=1}^{S}\left[\phi_{s,1} - \overline{\phi}_1\right]^2.$$

$$\begin{aligned}
&\widehat{V}_{\sigma^2[\mathrm{ITT}]} - \widehat{V}^I_{bound} \\
=&\frac{1}{S}\sum_{s=1}^{S}\left[\widehat{\phi}^2_{s,1} - \phi^2_{s,1}\right] - \left(\left(\frac{1}{S}\sum_{s=1}^{S}\phi_{s,1} + \frac{1}{S}\sum_{s=1}^{S}\left[\widehat{\phi}_{s,1} - \phi_{s,1}\right]\right)^2 - \left(\frac{1}{S}\sum_{s=1}^{S}\phi_{s,1}\right)^2\right). \qquad (7)
\end{aligned}$$

Let $(x,y,z) \mapsto g(x,y,z) = \tilde{w}_s\left[(x-y)^2 - z\right]$.
$\phi_{s,1} = g\left(\widehat{\mathrm{ITT}}_s, \mathrm{ITT}, \widehat{V}_{rob}\left(\widehat{\mathrm{ITT}}_s\right)\right)$, and $\widehat{\phi}_{s,1} = g\left(\widehat{\mathrm{ITT}}_s, \widehat{\mathrm{ITT}}, \widehat{V}_{rob}\left(\widehat{\mathrm{ITT}}_s\right)\right)$. Under Points 1 and 2 of Assumption 4, $(\widehat{\mathrm{ITT}}_s, \mathrm{ITT}, \widehat{V}_{rob}\left(\widehat{\mathrm{ITT}}_s\right))$ belongs to a compact subset Θ of $\mathbb{R}^3$, and as g is continuously differentiable, there exists a real number C such that $\left|\frac{\partial g}{\partial y}(x,y,z)\right| \leq C$ for all $(x,y,z) \in \Theta$.

$$\begin{aligned}
&\left|\frac{1}{S}\sum_{s=1}^{S}\left[\widehat{\phi}_{s,1} - \phi_{s,1}\right]\right| \\
\leq&\frac{1}{S}\sum_{s=1}^{S}\left|\widehat{\phi}_{s,1} - \phi_{s,1}\right| \\
=&\frac{1}{S}\sum_{s=1}^{S}\left|\left(\widehat{\mathrm{ITT}} - \mathrm{ITT}\right)\frac{\partial g}{\partial y}\left(\widehat{\mathrm{ITT}}_s, \tilde{a}_s, \widehat{V}_{rob}\left(\widehat{\mathrm{ITT}}_s\right)\right)\right|, \text{ for } \tilde{a}_s \in \left[\min(\widehat{\mathrm{ITT}}, \mathrm{ITT}), \max(\widehat{\mathrm{ITT}}, \mathrm{ITT})\right] \\
\leq&\left|\widehat{\mathrm{ITT}} - \mathrm{ITT}\right| C.
\end{aligned}$$

The first inequality follows from the triangle inequality, the equality follows from the mean value theorem. Then, as $\widehat{\mathrm{ITT}} - \mathrm{ITT} = o_P(1)$, the previous display implies that

$$\frac{1}{S}\sum_{s=1}^{S}\left[\widehat{\phi}_{s,1} - \phi_{s,1}\right] = o_P(1). \qquad (8)$$

One can use similar steps to show that

$$\frac{1}{S}\sum_{s=1}^{S}\left[\widehat{\phi}^2_{s,1} - \phi^2_{s,1}\right] = o_P(1). \qquad (9)$$

Finally, it follows from (7)-(9), the fact that under Assumptions 1 and 4 $\frac{1}{S}\sum_{s=1}^{S}\phi_{s,1} \xrightarrow{\mathbb{P}} \lim_{S\to+\infty}\frac{1}{S}\sum_{s=1}^{S}E\left(\phi_{s,1}\right)$, and the continuous mapping theorem, that

$$\widehat{V}_{\sigma^2[\mathrm{ITT}]} - \widehat{V}^I_{bound} = o_P(1). \qquad (10)$$

Finally, under Assumptions 1 and 4,

$$\widehat{V}^I_{bound} \xrightarrow{\mathbb{P}} \overline{v} \equiv \lim_{S\to+\infty} \frac{1}{S}\sum_{s=1}^{S} E\left(\phi_{s,1}^2\right) - \left(\lim_{S\to+\infty} \frac{1}{S}\sum_{s=1}^{S} E\left(\phi_{s,1}\right)\right)^2 \geq V_{\sigma^2[\mathrm{ITT}]},$$

where the inequality follows by convexity of $x \mapsto x^2$. The result follows from (10) and the previous display.

7.2 Proof of Lemma 1

Proof of Point 1

The first and last equalities are well-known results. The proof of the second one is similar to the proof of the second and third equalities in Point 3 below.

Proof of Point 2

$E\left(\overline{Y}_{0s}\right) = E(Y_s^r(0))$ is a well-known result. Conditional on $(Y_{is}^r(0))_{i\in\{1,\ldots,n_s\}}$, the only source of randomness in $\overline{Y}_{0s}$ is the random sampling, without replacement, of $n_{0,s}$ units out of n_s assigned to the control group. Then, as is well-known,

$$V\left(\overline{Y}_{0s}|(Y_{is}^r(0))_{i\in\{1,\ldots,n_s\}}\right) = r^2_{Y_s^r(0),s}\left(\frac{1}{n_{0,s}} - \frac{1}{n_s}\right).$$

Then, from the law of total variance and the fact that $E\left(r^2_{Y_s^r(0),s}\right) = V(Y_s^r(0))$, it follows that

$$V\left(\overline{Y}_{0s}\right) = \frac{V(Y_s^r(0))}{n_{0,s}}. \tag{11}$$

Then,

$$\begin{aligned}
&\mathrm{Cov}\left(\overline{Y}_{0s}, \overline{Y}_{1s}\right)\\
=&1/2\left(V\left(\overline{Y}_{0s}\right) + V\left(\overline{Y}_{1s}\right) - V\left(\widehat{\mathrm{ITT}}_s\right)\right)\\
=&1/2\left(\frac{V(Y_s^r(0))}{n_{0,s}} + \frac{V(Y_s^r(1))}{n_{1,s}} - \frac{V(Y_s^r(0))}{n_{0,s}} - \frac{V(Y_s^r(1))}{n_{1,s}}\right)\\
=&0,
\end{aligned} \tag{12}$$

The first equality follows from the fact that for any random variables A and B, $V(A-B) = V(A) + V(B) - 2\mathrm{Cov}(A,B)$. The second equality follows from (11), an equivalent equality for

$V\left(\overline{Y}_{1s}\right)$, and the fact that under Assumptions 1 and 2, $V\left(\widehat{\text{ITT}}_s\right) = \frac{V(Y_s^r(0))}{n_{0,s}} + \frac{V(Y_s^r(1))}{n_{1,s}}$ (see, e.g., Equation (6.17) in Imbens and Rubin, 2015). (12) directly implies that

$$\text{Cov}\left(\overline{Y}_{0s}, \widehat{\text{ITT}}_s\right) = -V\left(\overline{Y}_{0s}\right). \tag{13}$$

Finally, the result follows from (11), (13), and the fact that under Assumptions 1 and 2, $r^2_{Y,0,s}$ is unbiased for $V(Y_s^r(0))$.

Proof of Point 3

$E\left(\widehat{\mathbf{ITT}}_{\text{M},s}\right) = \mathbf{ITT}_{\text{M},s}$ is a well-known result. We only prove that $E\left(\frac{c_{M_k,Y,0,s}}{n_{0,s}} + \frac{c_{M_k,Y,1,s}}{n_{1,s}}\right) = \text{Cov}\left(\widehat{\text{ITT}}_{\text{M}_k,s}, \widehat{\text{ITT}}_s\right)$, the proof that $E\left(\frac{c_{M_k,M_{k'},0,s}}{n_{0,s}} + \frac{c_{M_k,M_{k'},1,s}}{n_{1,s}}\right) = \text{Cov}\left(\widehat{\text{ITT}}_{\text{M}_k,s}, \widehat{\text{ITT}}_{\text{M}_{k'},s}\right)$ is similar. Let $\mathcal{T}_s = (Y_{is}^r(0), Y_{is}^r(1), M_{k,is}^r(0), M_{k,is}^r(1))_{i\in\{1,\ldots,n_s\}}$. Under Assumptions 1 and 2, we can apply Theorem 3 in Li and Ding (2017) conditional on $\mathcal{T}_s$, to show that

$$\text{Cov}\left(\widehat{\text{ITT}}_{\text{M}_k,s}, \widehat{\text{ITT}}_s \middle| \mathcal{T}_s\right) = \frac{c_{M_k^r(0),Y^r(0),s}}{n_{0,s}} + \frac{c_{M_k^r(1),Y^r(1),s}}{n_{1,s}} - \frac{c_{M_k^r(1)-M_k^r(0),Y^r(1)-Y^r(0),s}}{n_s}. \tag{14}$$

Then,

$$\begin{aligned}
\text{Cov}\left(\widehat{\text{ITT}}_{\text{M}_k,s}, \widehat{\text{ITT}}_s\right) =& E\left(\text{Cov}\left(\widehat{\text{ITT}}_{\text{M}_k,s}, \widehat{\text{ITT}}_s \middle| \mathcal{T}_s\right)\right) + \text{Cov}\left(E\left(\widehat{\text{ITT}}_{\text{M}_k,s} \middle| \mathcal{T}_s\right), E\left(\widehat{\text{ITT}}_s \middle| \mathcal{T}_s\right)\right) \\
=& E\left(\frac{c_{M_k^r(0),Y^r(0),s}}{n_{0,s}} + \frac{c_{M_k^r(1),Y^r(1),s}}{n_{1,s}} - \frac{c_{M_k^r(1)-M_k^r(0),Y^r(1)-Y^r(0),s}}{n_s}\right) \\
&+ \text{Cov}\left(\frac{1}{n_s}\sum_{i=1}^{n_s}(M_{k,is}^r(1) - M_{k,is}^r(0)), \frac{1}{n_s}\sum_{i=1}^{n_s}(Y_{is}^r(1) - Y_{is}^r(0))\right) \\
=& \frac{\text{Cov}(M_{k,s}^r(0), Y_s^r(0))}{n_{0,s}} + \frac{\text{Cov}(M_{k,s}^r(1), Y_s^r(1))}{n_{1,s}} \\
&- \frac{\text{Cov}(M_{k,s}^r(1) - M_{k,s}^r(0), Y_s^r(1) - Y_s^r(0))}{n_s} \\
&+ \frac{\text{Cov}(M_{k,s}^r(1) - M_{k,s}^r(0), Y_s^r(1) - Y_s^r(0))}{n_s} \\
=& \frac{\text{Cov}(M_{k,s}^r(0), Y_s^r(0))}{n_{0,s}} + \frac{\text{Cov}(M_{k,s}^r(1), Y_s^r(1))}{n_{1,s}}.
\end{aligned} \tag{15}$$

The first equality follows from the law of total covariance. The second equality follows from (14), and the fact that $\widehat{\text{ITT}}_{\text{M}_k,s}$ and $\widehat{\text{ITT}}_s$ are conditionally unbiased for the sample ITT effects on the outcome and the mediator. The third equality follows from the fact that the vectors $(Y_{is}^r(0), Y_{is}^r(1), M_{k,is}^r(0), M_{k,is}^r(1))$ are iid across i. The result follows from the previous display, and the fact that under Assumptions 1 and 2, $c_{M_k,Y,0,s}$ and $c_{M_k,Y,1,s}$ are respectively unbiased for $\text{Cov}(M_{k,s}^r(0), Y_s^r(0))$, and $\text{Cov}(M_{k,s}^r(1), Y_s^r(1))$.

Assumption 7 *1. There exists real numbers M_0 and M_1 such that $|\widehat{\mathbf{X}}_s| \leq M_0$ and $\tilde{w}_s \leq M_1$, and the sequence $(\phi_{s,4})_{s\geq 1}$ satisfies the Lyapunov condition.*

2. The limits of the following sequences, when $S \to +\infty$, exist:

(a) $\sum_{s=1}^{S} w_s \mathbf{X}_s \mathbf{X}_s^T$

(b) $\mu(\mathbf{X})$

(c) $\sum_{s=1}^{S} w_s \mathbf{X}_s ITT_s$

(d) $1/S \sum_{s=1}^{S} V(\phi_{s,4})$.

7.3 Proof of Theorem 2

Proof of consistency.

We have

$$\beta_X^{\mathrm{ITT}}(\lambda) = \left(\sum_{s=1}^{S} w_s \mathbf{X}_s \mathbf{X}_s^T - \mu(\mathbf{X})\mu(\mathbf{X})^T + \lambda \mathbf{I}_K\right)^{-1} \left(\sum_{s=1}^{S} w_s \mathbf{X}_s \mathrm{ITT}_s - \mu(\mathbf{X})\mathrm{ITT}\right), \tag{16}$$

and

$$\begin{aligned}\widehat{\beta}_X^{\mathrm{ITT}}(\lambda) = &\left(\sum_{s=1}^{S} w_s \left(\widehat{\mathbf{X}}_s \widehat{\mathbf{X}}_s^T - \widehat{V}\left(\widehat{\mathbf{X}}_s\right)\right) - \widehat{\mu}(\mathbf{X})\widehat{\mu}(\mathbf{X})^T + \lambda \mathbf{I}_K\right)^{-1} \\ &\times \left(\sum_{s=1}^{S} w_s \left(\widehat{\mathbf{X}}_s \widehat{\mathrm{ITT}}_s - \widehat{\mathrm{Cov}}\left(\widehat{\mathbf{X}}_s, \widehat{\mathrm{ITT}}_s\right)\right) - \widehat{\mu}(\mathbf{X})\widehat{\mathrm{ITT}}\right).\end{aligned} \tag{17}$$

Moreover,

$$E\left(\widehat{\mathbf{X}}_s \widehat{\mathbf{X}}_s^T - \widehat{V}\left(\widehat{\mathbf{X}}_s\right)\right) = E\left(\widehat{\mathbf{X}}_s\right) E\left(\widehat{\mathbf{X}}_s^T\right) = \mathbf{X}_s \mathbf{X}_s^T. \tag{18}$$

The first equality follows from the fact $\widehat{V}\left(\widehat{\mathbf{X}}_s\right)$ is unbiased for $V\left(\widehat{\mathbf{X}}_s\right) = E\left(\widehat{\mathbf{X}}_s \widehat{\mathbf{X}}_s^T\right) - E\left(\widehat{\mathbf{X}}_s\right) E\left(\widehat{\mathbf{X}}_s^T\right)$. The second equality follows from the fact $\widehat{\mathbf{X}}_s$ is unbiased.

Similarly,

$$E\left(\widehat{\mathbf{X}}_s \widehat{\mathrm{ITT}}_s - \widehat{\mathrm{Cov}}\left(\widehat{\mathbf{X}}_s, \widehat{\mathrm{ITT}}_s\right)\right) = E\left(\widehat{\mathbf{X}}_s\right) E\left(\widehat{\mathrm{ITT}}_s\right) = \mathbf{X}_s \mathrm{ITT}_s. \tag{19}$$

The first equality follows from the fact $\widehat{\mathrm{Cov}}\left(\widehat{\mathbf{X}}_s, \widehat{\mathrm{ITT}}_s\right)$ is unbiased for $\mathrm{Cov}\left(\widehat{\mathbf{X}}_s, \widehat{\mathrm{ITT}}_s\right) = E\left(\widehat{\mathbf{X}}_s \widehat{\mathrm{ITT}}_s\right) - E\left(\widehat{\mathbf{X}}_s\right) E\left(\widehat{\mathrm{ITT}}_s\right)$. The second equality follows from the fact $\widehat{\mathbf{X}}_s$ and $\widehat{\mathrm{ITT}}_s$ are unbiased.

Finally, the result follows from (16)-(19), the fact that $\widehat{\mathbf{X}}_s$ and the normalized weights $\tilde{w}_s$ are bounded, the fact that random variables are independent across sites, the law of large numbers for independent variables in Lemma 1 of Liu et al. (1988), Point 2 of Assumption 7, and repeated uses of the continuous mapping theorem.

Proof of asymptotic normality.

Let

$$\begin{aligned}\tilde{A}(\lambda) &= \sum_{s=1}^{S} w_s \left(\left(\widehat{\mathbf{X}}_s - \mu(\mathbf{X}) \right) \left(\widehat{\mathbf{X}}_s - \mu(\mathbf{X}) \right)^T - \widehat{V}\left(\widehat{\mathbf{X}}_s \right) \right) + \lambda \mathbf{I}_K \\ \tilde{B} &= \sum_{s=1}^{S} w_s \left(\left(\widehat{\mathbf{X}}_s - \mu(\mathbf{X}) \right) \left(\widehat{\mathrm{ITT}}_s - \mathrm{ITT} \right) - \widehat{\mathrm{Cov}}\left(\widehat{\mathbf{X}}_s, \widehat{\mathrm{ITT}}_s \right) \right).\end{aligned}$$

As $E\left(\sum_{s=1}^{S} w_s \left(\widehat{\mathbf{X}}_s - \mu(\mathbf{X}) \right) \right) = 0$, it follows from a Taylor expansion that

$$\sqrt{S}\left(\widehat{A}(\lambda) - \tilde{A}(\lambda) \right) = \sqrt{S} \left(\hat{\mu}(\mathbf{X}) - \mu(\mathbf{X}) \right) o_P(1) + o_P(1) = o_P(1). \tag{20}$$

Similarly,

$$\sqrt{S}\left(\widehat{B} - \widetilde{B} \right) = o_P(1). \tag{21}$$

Using the same arguments as in the proof of Theorem 2, one can show that $A(\lambda) = \frac{1}{S} \sum_{s=1}^{S} E(\phi_{s,2})$. Combined with (20), this implies that

$$\sqrt{S}\left(\widehat{A}(\lambda) - A(\lambda) \right) = \frac{1}{\sqrt{S}} \sum_{s=1}^{S} \left(\phi_{s,2} - E(\phi_{s,2}) \right) + o_P(1). \tag{22}$$

Similarly, one can show that

$$\sqrt{S}\left(\widehat{B} - B \right) = \frac{1}{\sqrt{S}} \sum_{s=1}^{S} \left(\phi_{s,3} - E(\phi_{s,3}) \right) + o_P(1). \tag{23}$$

Finally, using the fact that

$$\sqrt{S}\left(\widehat{A}^{-1}(\lambda)\widehat{B} - [A(\lambda)]^{-1} B \right) = \sqrt{S}\left(-[A(\lambda)]^{-1} \left(\widehat{A}(\lambda) - A(\lambda) \right) [A(\lambda)]^{-1} B + [A(\lambda)]^{-1} \left(\widehat{B} - B \right) \right) + o_P(1), \tag{24}$$

it follows from (22) and (23) that

$$\sqrt{S}\left(\widehat{\beta}_X^{\mathrm{ITT}}(\lambda) - \beta_X^{\mathrm{ITT}}(\lambda) \right) = \frac{1}{\sqrt{S}} \sum_{s=1}^{S} \left(\phi_{s,4} - E(\phi_{s,4}) \right) + o_P(1).$$

The result follows from applying the Lyapunov CLT to $(\phi_{s,4})_{s \geq 1}$, and from the Slutsky lemma.

7.4 Proof of Theorem 4

It follows from Jensen's inequality that the denominator of $\underline{\sigma}^2[\text{LATE}]$ is larger than that of $\sigma^2[\text{LATE}]$, and it follows from Assumption 6 that both denominators have a finit limit. Then, one can use arguments similar to those used to show Theorem 1, the fact that $\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\nu_s = 0$, and Assumption 6, to show that their numerators have the same finite limit.

We now show the asymptotic normality result. It follows from, e.g., (A28) in De Chaisemartin and d'Haultfoeuille (2018) and the fact that $\frac{1}{S}\sum_{s=1}^{S}E(\phi_{s,5}) = 0$ that

$$\widehat{\text{LATE}} - \text{LATE} = \frac{1}{S}\sum_{s=1}^{S}\phi_{s,5} + o_P\left(\frac{1}{\sqrt{S}}\right). \tag{25}$$

As the variables $\phi_{s,5}$ are independent and bounded, it then follows from the law of large numbers in Lemma 1 of Liu et al. (1988) that

$$\widehat{\text{LATE}} - \text{LATE} = o_P(1). \tag{26}$$

Then, letting $\tilde{\nu}_s(x) = \widehat{\text{ITT}}_s - x \times \widehat{\text{FS}}_s$,

$$\begin{aligned}
&\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\left(\widehat{\nu}_s\right)^2\\
=&\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s(\tilde{\nu}_s)^2 + \frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\left[\left(\widehat{\nu}_s\right)^2 - (\tilde{\nu}_s)^2\right]\\
=&\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s(\tilde{\nu}_s)^2 + \left(\widehat{\text{LATE}} - \text{LATE}\right)\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\frac{\partial\left(\tilde{\nu}_s^2\right)}{\partial x}(\tilde{\text{LATE}}_s)\\
=&\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s(\tilde{\nu}_s)^2 + \left(\widehat{\text{LATE}} - \text{LATE}\right)\left(\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\frac{\partial\left(\tilde{\nu}_s^2\right)}{\partial x}(\text{LATE}) + \frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\frac{\partial^2\left(\tilde{\nu}_s^2\right)}{\partial x^2}(\bar{\text{LATE}}_s)(\tilde{\text{LATE}}_s - \text{LATE})\right),
\end{aligned}$$

where the second and third equalities follow from the mean-value theorem, for some $\tilde{\text{LATE}}_s$ included between LATE and $\widehat{\text{LATE}}$, and for some $\bar{\text{LATE}}_s$ included between LATE and $\tilde{\text{LATE}}_s$. As $\frac{\partial\left(\tilde{\nu}_s^2\right)}{\partial x}(x) = -2\widehat{\text{FS}}_s\left(\widehat{\text{ITT}}_s - \widehat{\text{FS}}_s x\right)$ and $\frac{\partial^2\left(\tilde{\nu}_s^2\right)}{\partial x^2}(x) = 2\widehat{\text{FS}}_s^2$,

$$\begin{aligned}
&\left|\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\frac{\partial^2\left(\tilde{\nu}_s^2\right)}{\partial x^2}(\bar{\text{LATE}}_s)(\widehat{\text{LATE}}_s - \text{LATE})\right|\\
=&\left|\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s 2\widehat{\text{FS}}_s^2(\widehat{\text{LATE}}_s - \text{LATE})\right|\\
\leq&\left|\widehat{\text{LATE}} - \text{LATE}\right| 2\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\widehat{\text{FS}}_s^2\\
=&o_P(1),
\end{aligned}$$

where the last equality follows from (26), from applying the law of large numbers in Lemma 1 of Liu et al. (1988) to the sequence of independent and bounded random variables $\tilde{w}_s\widehat{\mathrm{FS}}_s^2$, and from Point 2i) of Assumption 6. Therefore,

$$\begin{aligned}\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\left(\widehat{\nu}_s\right)^2 =&\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s(\tilde{\nu}_s)^2 - 2\left(\widehat{\mathrm{LATE}}-\mathrm{LATE}\right)\left(\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\widehat{\mathrm{FS}}_s\tilde{\nu}_s+o_P(1)\right)\\ =&\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s(\tilde{\nu}_s)^2 - 2\left(\widehat{\mathrm{LATE}}-\mathrm{LATE}\right)\left(C_1+o_P(1)\right)\\ =&\frac{1}{S}\sum_{s=1}^{S}\left(\tilde{w}_s(\tilde{\nu}_s)^2-2C_1\phi_{s,5}\right)+o_P\left(\frac{1}{\sqrt{S}}\right).\end{aligned} \tag{27}$$

The second equality follows from applying the law of large numbers in Lemma 1 of Liu et al. (1988) to the sequence of independent and bounded random variables $\tilde{w}_s\widehat{\mathrm{FS}}_s\tilde{\nu}_s$ and from Point 2ii) of Assumption 6. The third equality follows from (25).

Similarly, let

$$\begin{aligned}\tilde{\nu}_{is}(x) =&Y_{is}-D_{is}\times x\\ v\left(x\right) =&\frac{1}{n_{1s}}r^2_{\tilde{\nu}(x),1,s}+\frac{1}{n_{0s}}r^2_{\tilde{\nu}(x),0,s}\\ =&\frac{1}{n_{1s}(n_{1s}-1)}\sum_{i=1}^{n_s}Z_{is}\left(Y_{is}-\overline{Y}_{1s}-\left(D_{is}-\overline{D}_{1s}\right)x\right)^2\\ &+\frac{1}{n_{0s}(n_{0s}-1)}\sum_{i=1}^{n_s}(1-Z_{is})\left(Y_{is}-\overline{Y}_{0s}-\left(D_{is}-\overline{D}_{0s}\right)x\right)^2.\end{aligned}$$

One has

$$\begin{aligned}\frac{\partial v}{\partial x}(x) =&2\left(\frac{1}{n_{1s}}(x\times r^2_{D,1,s}-c_{D,Y,1,s})+\frac{1}{n_{0s}}(x\times r^2_{D,0,s}-c_{D,Y,0,s})\right)\\ \frac{\partial^2 v}{\partial x^2}(x) =&2\left(\frac{1}{n_{1s}}r^2_{D,1,s}+\frac{1}{n_{0s}}r^2_{D,0,s}\right).\end{aligned}$$

Then, using arguments similar to those used to show (27),

$$\begin{aligned}
&\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\widehat{V}_{rob}\left(\widehat{\nu}_s\right)\\
=&\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\widehat{V}_{rob}\left(\tilde{\nu}_s\right)+\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\left[\widehat{V}_{rob}\left(\widehat{\nu}_s\right)-\widehat{V}_{rob}\left(\tilde{\nu}_s\right)\right]\\
=&\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\widehat{V}_{rob}\left(\tilde{\nu}_s\right)+\left(\widehat{\text{LATE}}-\text{LATE}\right)\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\frac{\partial v}{\partial x}(\tilde{\text{LATE}}_s)\\
=&\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\widehat{V}_{rob}\left(\tilde{\nu}_s\right)+2\left(\widehat{\text{LATE}}-\text{LATE}\right)\left(C_2+o_P(1)\right)\\
=&\frac{1}{S}\sum_{s=1}^{S}\left(\tilde{w}_s\widehat{V}_{rob}\left(\tilde{\nu}_s\right)+2C_2\phi_{s,5}\right)+o_P\left(\frac{1}{\sqrt{S}}\right). \qquad (28)
\end{aligned}$$

Then, it follows from (27) and (28) that

$$\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\left[(\widehat{\nu}_s)^2-\widehat{V}_{rob}\left(\widehat{\nu}_s\right)\right]=\frac{1}{S}\sum_{s=1}^{S}\left(\tilde{w}_s\left(\tilde{\nu}_s\right)^2-\widehat{V}_{rob}\left(\tilde{\nu}_s\right)\right)-2(C_1+C_2)\phi_{s,5}\Big)+o_P\left(\frac{1}{\sqrt{S}}\right). \qquad (29)$$

Let

$$\tilde{\underline{\sigma}}^2[\text{LATE}]=\frac{\frac{1}{S}\sum_{s=1}^{S}\left(\tilde{w}_s\left(\tilde{\nu}_s\right)^2-\widehat{V}_{rob}\left(\tilde{\nu}_s\right)\right)-2(C_1+C_2)\phi_{s,5}\Big)}{\frac{1}{S}\sum_{s=1}^{S}\widehat{\text{FS}}_s^2}.$$

It follows from, e.g., (A28) in De Chaisemartin and d'Haultfoeuille (2018), and from the fact that $\frac{1}{S}\sum_{s=1}^{S}E(\phi_{s,5})=0$, that

$$\sqrt{S}(\tilde{\underline{\sigma}}^2[\text{LATE}]-\underline{\sigma}^2[\text{LATE}])=\frac{1}{\sqrt{S}}\sum_{s=1}^{S}\left(\phi_{s,6}-E(\phi_{s,6})\right)+o_P\left(1\right). \qquad (30)$$

Then, it follows from (29), (30) and Point 3 of Assumption 6 that

$$\sqrt{S}(\widehat{\underline{\sigma}}^2[\text{LATE}]-\underline{\sigma}^2[\text{LATE}])=\frac{1}{\sqrt{S}}\sum_{s=1}^{S}\left(\phi_{s,6}-E(\phi_{s,6})\right)+o_P\left(1\right). \qquad (31)$$

The result follows from applying the Lyapunov CLT to $(\phi_{s,6})_{s\geq 1}$, and from the Slutsky lemma.

7.5 Proof of Theorem 5

$$\begin{aligned}
\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\text{ITT}_s(\text{FS}_s-\text{FS})=&\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\text{LATE}_s\text{FS}_s(\text{FS}_s-\text{FS})\\
=&\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\text{LATE}_s\times\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\text{FS}_s(\text{FS}_s-\text{FS})\\
=&\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\text{LATE}_s\sigma^2\left[\text{FS}\right],
\end{aligned}$$

where the second equality follows from Assumption 5. Therefore,

$$\text{LATE} = \frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\text{LATE}_s = \frac{\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s\text{ITT}_s(\text{FS}_s - \text{FS})}{\sigma^2\left[\text{FS}\right]} = \beta_{\text{FS}}^{\text{ITT}}, \tag{32}$$

where the first equality follows again from Assumption 5.

7.6 Proof of Theorem 6

$$\begin{aligned}
&\frac{1}{S_{x,1}}\sum_{s=1}^{S}x_s w_s\left(\text{ITT}_s - \text{FS}_s\text{LATE}_{x,0}\right) - \frac{1}{S_{x,0}}\sum_{s=1}^{S}(1-x_s)w_s\left(\text{ITT}_s - \text{FS}_s\text{LATE}_{x,1}\right)\\
=&\frac{1}{S_{x,1}}\sum_{s=1}^{S}x_s w_s\text{FS}_s\left(\text{LATE}_s - \text{LATE}_{x,0}\right) - \frac{1}{S_{x,0}}\sum_{s=1}^{S}(1-x_s)w_s\text{FS}_s\left(\text{LATE}_s - \text{LATE}_{x,1}\right)\\
=&\left(\frac{1}{S_{x,1}}\sum_{s=1}^{S}x_s w_s\text{FS}_s\right)\left(\frac{1}{S_{x,1}}\sum_{s=1}^{S}x_s w_s\left(\text{LATE}_s - \text{LATE}_{x,0}\right)\right)\\
-&\left(\frac{1}{S_{x,0}}\sum_{s=1}^{S}(1-x_s)w_s\text{FS}_s\right)\left(\frac{1}{S_{x,0}}\sum_{s=1}^{S}(1-x_s)w_s\left(\text{LATE}_s - \text{LATE}_{x,1}\right)\right)\\
=&\beta_x^{\text{LATE}}\text{FS}.
\end{aligned}$$

8 Survey of Multi-Site RCTs

Table 7: Multi-site RCTs in AEJ: Applied Economics 2014-2016

Title	Units of Observation	Units of Randomization	Sites
Keeping It Simple: Financial Literacy and Rules of Thumb	Individual Clients	1,193 Individual Clients	107 Barrio
Improving Educational Quality through Enhancing Community Participation: Results from a Randomized Field Experiment in Indonesia	Students	520 Schools	44 Subdistricts
The Demand for Medical Male Circumcision	Individuals	1,634 Individuals	29 Enumeration Areas
Should Aid Reward Performance? Evidence from a Field Experiment on Health and Education in Indonesia	Individuals	300 Kecamatan	20 Kabupaten
Private and Public Provision of Counseling to Job Seekers: Evidence from a Large Controlled Experiment	Individuals	43,977 Individuals	216 Employment Offices
Estimating the Impact of Microcredit on Those Who Take It Up: Evidence from a Randomized Experiment in Morocco	Households	Villages (81 pairs)	47 Branches
Microcredit Impacts: Evidence from a Randomized Microcredit Program Placement Experiment by Compartamos Banco	Households	250 Geographic Clusters	Superclusters of 4 Adjacent Clusters
The Impacts of Microcredit: Evidence from Bosnia and Herzegovina	Individuals	1,196 Individuals	282 City/Towns or 14 Branches
Social Networks and the Decision to Insure	Households	5,300 Households	185 Villages
Inputs in the Production of Early Childhood Human Capital: Evidence from Head Start	Individuals	4,442 Individuals	353 Head Start Centers
The Returns to Microenterprise Support among the Ultrapoor: A Field Experiment in Postwar Uganda[14]	Individuals	904 Individuals	60 Villages
The Impact of High School Financial Education: Evidence from a Large-Scale Evaluation in Brazil	Student	892 Schools (in matched pairs)	Municipalities

"The Returns to Microenterprise Support among the Ultrapoor: A Field Experiment in Postwar Uganda" corresponds to the Phase 2 experiment.

"Social Networks and the Decision to Insure" corresponds to the household level randomization and analysis.