



ROCKWOOL Foundation Berlin

Institute for the Economy and the Future of Work (RFBerlin)

DISCUSSION PAPER SERIES

254/26

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

Clement de Chaisemartin, Antoine Deeb

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

Authors

Clement de Chaisemartin, Antoine Deeb

Reference

JEL Codes: C21

Keywords: Measurement error, multi-site randomized controlled trials, multi-centre randomized controlled trials, heterogeneous treatment effects

Recommended Citation: Clement de Chaisemartin, Antoine Deeb (2026): Estimating treatment-effect heterogeneity across sites, in multi-site randomized experiments with few units per site.. RFBerlin Discussion Paper No. 254/26

Access

Papers can be downloaded free of charge from the RFBerlin website: <https://www.rfberlin.com/discussion-papers>

Discussion Papers of RFBerlin are indexed on RePEc: <https://ideas.repec.org/s/crm/wpaper.html>

Disclaimer

Opinions and views expressed in this paper are those of the author(s) and not those of RFBerlin. Research disseminated in this discussion paper series may include views on policy, but RFBerlin takes no institutional policy positions. RFBerlin is an independent research institute.

RFBerlin Discussion Papers often represent preliminary or incomplete work and have not been peer-reviewed. Citation and use of research disseminated in this series should take into account the provisional nature of the work. Discussion papers are shared to encourage feedback and foster academic discussion.

All materials were provided by the authors, who are responsible for proper attribution and rights clearance. While every effort has been made to ensure proper attribution and accuracy, should any issues arise regarding authorship, citation, or rights, please contact RFBerlin to request a correction.

These materials may not be used for the development or training of artificial intelligence systems.

Imprint

RFBerlin
ROCKWOOL Foundation Berlin –
Institute for the Economy
and the Future of Work

Gormannstrasse 22, 10119 Berlin
Tel: +49 (0) 151 143 444 67
E-mail: info@rfberlin.com
Web: www.rfberlin.com



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

Clément de Chaisemartin[†]

Antoine Deeb[‡]

September 18, 2026

Abstract

Multi-site randomized experiments, which are common in economics and medicine, can reveal how treatment effects vary across settings. However, differences between site-specific estimates reflect both treatment-effect heterogeneity and estimation noise. We build on established measurement-error corrections and develop new ones to study this heterogeneity in experiments with many sites and few units per site. In particular, our results can be used to estimate the variance and centered third moment of treatment effects across sites, relate site effects to untreated outcomes, conduct mediation analyses, assess the roles of provider effectiveness and site responsiveness, and study heterogeneity in local average treatment effects (LATEs) under imperfect compliance. We apply our methods to counseling programs delivered by public and private providers in about 200 French employment offices. Average employment gains of about 2 percentage points conceal cross-site standard deviations of intention-to-treat effects of 8 to 9 percentage points. We also find substantial LATE heterogeneity. Effects are negatively associated with untreated employment rates. The two programs also tend to be more effective in the same sites, consistent with a role for heterogeneous site responsiveness.

Keywords: Measurement error, multi-site randomized controlled trials, multi-centre randomized controlled trials, heterogeneous treatment effects

*We are very grateful to Manuel Arellano, Dmitry Arkhangelsky, Xavier D'Haultfoeuille, Peng Ding, Roberto Galbiati, Grégory Jolivet, Julien Monardo, Timo Schenk, Liyang Sun, Christine Valente, Yanos Zylberberg, and seminar participants at the 2025 Encounters in Econometric Theory Workshop, CEMFI, KU Leuven, the University of Bristol, and the World Bank for their helpful comments. Clément de Chaisemartin was funded by the European Union (ERC, REALLYCREDIBLE, GA N°101043899).

AI disclosure statement: Versions 1–3 of the arXiv manuscript were written without AI assistance. The current version was prepared with the assistance of ChatGPT for the theory and Claude for the empirical application and simulations. Among the paper's main results, Propositions 2, 3, 4, and 6, and Point 1 of Proposition 5 were conjectured and proven by us with little AI assistance. Equation (4) and Proposition 1 were conjectured by us and proven by us with some assistance from ChatGPT. Proposition 7 was conjectured by us and proven by ChatGPT with little assistance from us. Point 2 of Proposition 5 was conjectured and proven by ChatGPT with little assistance from us. We checked all AI-generated arguments and take responsibility for the final manuscript.

Estimated CO₂e emissions generated by AI use: Our interactions with ChatGPT in preparing this paper involved approximately 220.1 million input tokens—of which 205.9 million were cached—and 0.92 million output tokens, corresponding to an estimated 1.92 kg CO₂e under GPT's reference scenario; this estimate covers only the operational emissions associated with the marginal electricity consumed in processing our prompts and excludes emissions from training the models and their predecessors. Our interactions with Claude in preparing this paper involved approximately 94.6 million input tokens, of which 87.7 million were cached, and 0.35 million output tokens, corresponding to an estimated 0.83 kg CO₂e under the same reference scenario.

[†]Department of Economics, Sciences Po

[‡]Development Impact, World Bank

“Vérité au-deçà des Pyrénées, erreur au-delà.”
“*Truth on this side of the Pyrenees, error on the other.*”
—Blaise Pascal, *Pensées*

1 Introduction

In multi-site randomized controlled trials (RCTs), on top of learning whether the treatment works on average, researchers may want to know how much treatment effects vary across sites, whether effects are larger where untreated outcomes are worse, and whether sites’ effects on a mediator predict their effects on the final outcome. When compliance with random assignment is imperfect, researchers may also want to know whether differences in intention-to-treat effects reflect only differences in take-up rates or also differences in local average treatment effects (LATEs). Answering these questions is difficult when each site contains few units. Due to estimation error, site-specific estimates can differ substantially even when the underlying effects are similar. Moreover, estimation error can create misleading relationships between estimated effects and other estimated site characteristics. Accordingly, while experiments with many sites and few units per site are common in economics¹ and medicine,² researchers rarely systematically investigate effect heterogeneity across sites, and those that do so generally use parametric random-effects models.³

We develop methods for studying treatment-effect heterogeneity in multi-site RCTs with many sites and few units per site. Section 2 develops the statistical framework. It introduces classical measurement-error corrections, which subtract the contribution of estimation noise from observed cross-site variances and covariances (Fuller, 1987), and also proposes a novel corrected estimator of the centered third moment of site effects. Section 3 then shows that under i.i.d. sampling within sites, random assignment provides the unbiased estimators needed to implement these corrections.

The remainder of Section 3 shows how these results can be used to conduct six types of heterogeneity analyses. First, to assess whether effects vary across sites, we estimate the variance of average treatment effects (ATEs) across sites. Second, to assess whether heterogeneity is asymmetric, we estimate ATEs’ centered third moment, which measures whether unusually large treatment effects are concentrated above or below the average.

Third, to assess if effects are larger where untreated outcomes are worse, we regress sites’ ATEs on their average untreated outcome. This regression may inform targeting to new sites outside the RCT, if the relationship between treatment effects and untreated outcomes carries over to the new sites, and if outcomes among previous cohorts provide a useful proxy for new sites’ untreated outcomes. A central difficulty is that the control-group mean enters positively in the estimated untreated outcome and negatively in the estimated ATE. Our regression estimator corrects for the resulting negative covariance between the estimators, and for measurement error in the regressor.

¹From 2014 to 2016, 12 multi-site RCTs were published in *AEJ: Applied Economics*, with a median of 46 sites and a median of 15 randomization units per site (see Table 8 in the appendix).

²In medicine, such experiments are usually called multi-center RCTs. Reviewing 206 multi-center RCTs published in four leading medical journals, Kahan and Harhay (2015) report medians of 20 centers and 24 patients per center; 30% of the trials included more than 50 centers.

³In Table 8, the only paper that systematically investigates effect heterogeneity across sites is Walters (2015), who does so using a parametric random-coefficient model. In medicine, Hemming, Hughes, McKenzie, and Forbes (2021) note that analyses of treatment-effect heterogeneity across centers remain uncommon. Exceptions exist (see, e.g. Lingsma, Roozenbeek, Perel, Roberts, Maas, and Steyerberg, 2011), though they usually rely on parametric random-effects models (Hemming, Hughes, McKenzie, and Forbes, 2021).

Fourth, we study predictive mediation regressions: do sites with larger treatment effects on a mediator also have larger effects on the final outcome? With a binary mediator, under monotonicity and a strong cross-site independence assumption, the regression’s intercept identifies the average direct effect, while its slope identifies the average across sites of the mediator’s effect among each site’s mediator compliers. We derive a testable implication of homogeneous direct and mediator effects across sites, a sufficient condition for the cross-site independence assumption.

Fifth, when different providers deliver two similar treatments within the same sites, we use the variance of the difference between their site-specific effects to assess the roles of provider effectiveness and site responsiveness. We give conditions under which a constant difference across sites rules out heterogeneous provider effectiveness as a source of effect heterogeneity. More generally, this variance can help assess the roles of sites’ responsiveness and providers’ effectiveness, although it does not separately identify their contributions.

Sixth, we consider imperfect compliance. Site-specific local average treatment effects (LATEs) are ratios and generally lack unbiased estimators, so the usual measurement-error corrections cannot be applied directly to estimated LATEs. Nevertheless, we identify and estimate a variance of LATEs that weights sites proportionally to their squared first stages. This target depends only on estimable second moments of intention-to-treat (ITT) effects and first stages. We also identify and estimate a lower bound for the complier-weighted variance. These results allow researchers to test whether differences in take-up alone explain ITT heterogeneity. We also identify and estimate the covariance between LATEs and first stages under complier weights. This reveals whether sites with larger shares of compliers also tend to have compliers who benefit more from treatment; a positive covariance is consistent with Roy selection across sites, whereby take-up is higher where compliers’ gains are larger. Finally, we identify and estimate weighted regressions relating site-specific LATEs on different outcomes.

Section 4 applies these tools to the experiment of Behaghel, Crépon, and Gurgand (2014), which compared publicly and privately provided intensive counseling for job seekers in French public-employment offices. Average effects on employment after six months are 2.4 percentage points for public counseling and 1.9 percentage points for private counseling. These averages conceal substantial heterogeneity: the estimated standard deviations of site-specific ITTs are approximately 9.2 and 8.5 percentage points, respectively. We also find substantial LATE heterogeneity. Both programs’ estimated effects are negatively associated with employment rates without treatment, with particularly precise evidence for private counseling. This association is concentrated in the component of employment rate without treatment not predicted by job seekers’ observed characteristics. The programs also tend to be more effective in the same sites, despite being delivered by different providers. This positive association is consistent with a role for site responsiveness.

Related literature and paper outline

The corrected-moment estimators we use build directly on the measurement-error literature. Subtracting estimation-error variance from the observed variance of estimated effects is familiar from classical measurement-error models (Fuller, 1987), random-effects meta-analysis (DerSimonian and Laird, 1986), and empirical work on industry wage effects (Krueger and Summers, 1988), teacher effects (Aaronson, Barrow, and Sander, 2007), and employers’ propensities to discriminate (Kline, Rose, and Walters, 2022). A particularly close connection is to the leave-out variance-component estimators of Kline, Saggio, and Sølvesten (2020). Specializing their random-coefficient model to

a binary regressor yields a leave-one-out variance estimator for a group-specific slope that coincides with the conventional variance estimator for a difference in means used below. Similarly, the regression estimator we consider specializes the corrected measurement-error estimator in Equation (3.1.7) of Fuller (1987). Related estimators have been used with survey averages (Deaton, 1985) and teachers’ value added (Rose, Schellenberg, and Shem-Tov, 2022) as regressors. Another related approach regresses outcomes on empirical-Bayes posterior means of estimated group effects (see, e.g., Kline, Rose, and Walters, 2022). Chen, Gu, and Kwon (2025) show that this approach generally requires additional restrictions that corrected-moment estimators do not require.

To our knowledge, our centered-third-moment estimator is new. The closest precedents are Rimoldini (2014), whose correction assumes independent Gaussian measurement errors with known variances, and Ben-Moshe and Genesove (2026), who derive unbiased estimators of higher central moments in unbalanced multilevel models with mutually independent additive random effects and within-level i.i.d. errors. Our framework instead allows the sampling-error distribution to vary with the site-specific conditional mean.

Our paper’s main focus is to apply corrected-moment estimators to multi-site randomized experiments. Proposition 1 establishes that under our sampling and random-assignment assumptions, one can form unbiased estimators of site-level effects, and of their variances, covariances, and third central moments, as required by the corrections. The subsequent propositions develop the six applications described above, including results for LATE heterogeneity that cannot be obtained by directly applying the correction to site-specific LATE estimators.⁴ These applications connect to work on mediation analysis (Imai, Keele, and Yamamoto, 2010; Daniels and Hughes, 1997; Dunn, Emsley, Liu, Landau, Green, White, and Pickles, 2015; Reardon and Raudenbush, 2013; VanderWeele, 2013; Kwon and Roth, 2026), the contribution of implementation to effect heterogeneity (Bloom, Hill, and Riccio, 2003; Hong, Deutsch, Kress, Trinidad, and Xu, 2024; Angrist and Meager, 2023), and LATE heterogeneity in multi-site RCTs (Walters, 2015; Adusumilli, Agostinelli, and Borghesan, 2024; Angrist and Meager, 2023). We discuss these connections in the corresponding subsections of Section 3.

The remainder of the paper is organized as follows. Section 2 presents the general framework and corrected-moment estimators. Section 3 develops their applications to multi-site RCTs. Section 4 applies the methods to public and private counseling programs for French job seekers. Appendix A presents the influence functions and inference results. Appendix B reports simulations. Appendix C contains the proofs. Appendix D presents a survey of multi-site RCTs published in economics.

2 General framework and estimators

In this section, we assume we observe unbiased noisy realizations of a univariate and a multivariate conditional expectation. In such settings, we propose estimators of the variance of the univariate conditional expectation, of its third central moment, and of the coefficient from the regression of the univariate on the multivariate conditional expectation. The following section will apply these results to multi-site RCTs.

⁴The estimators proposed in this paper are implemented in the `multisite` Stata package.

2.1 Setup and target parameters

Let (X, Y, S) be random variables, where Y and S are scalar and $X \in \mathbb{R}^K$. Throughout, vectors are understood to be column vectors. For any random variable or vector R such that the expectation below is well defined, let

$$\mu_R(s) = \mathbb{E}[R \mid S = s].$$

Let $W(S)$ be such that $W(S) > 0$ and $0 < \mathbb{E}[W(S)] < \infty$. $W(S)$ is a weighting variable that gives greater weight to some values of S than others. For instance, S may represent the identifier of a cluster of units, and $W(S)$ could represent the cluster's number of units. Let $w(S) = W(S)/\mathbb{E}[W(S)]$. For any random variables R_1 and R_2 , we let

$$\begin{aligned}\mathbb{E}_w[R_1] &= \mathbb{E}[w(S)R_1] \\ \text{Var}_w[R_1] &= \mathbb{E}[w(S)(R_1 - \mathbb{E}_w[R_1])^2] \\ \text{Cov}_w(R_1, R_2) &= \mathbb{E}[w(S)(R_1 - \mathbb{E}_w[R_1])(R_2 - \mathbb{E}_w[R_2])].\end{aligned}$$

Our three target parameters are

$$\sigma^2 := \text{Var}_w(\mu_Y(S)), \quad \beta := \text{Var}_w(\mu_X(S))^{-1} \text{Cov}_w(\mu_X(S), \mu_Y(S)), \quad \kappa_Y := \mathbb{E}_w[\{\mu_Y(S) - \mathbb{E}_w[\mu_Y(S)]\}^3].$$

σ^2 is the weighted variance of $\mu_Y(S)$. The vector β is the slope coefficient in the population OLS regression of $\mu_Y(S)$ on a constant and $\mu_X(S)$, weighted by $w(S)$. κ_Y is the weighted third central moment of $\mu_Y(S)$.

We observe $(S_1, \dots, S_n)$, n i.i.d. random variables with the same distribution as S . We also observe $(W(S_1), \dots, W(S_n))$, the weights of observations 1 to n . For $i = 1, \dots, n$, let

$$\hat{\mu}_X(S_i) \quad \text{and} \quad \hat{\mu}_Y(S_i)$$

denote estimators of $\mu_X(S_i)$ and $\mu_Y(S_i)$, respectively. Let

$$\hat{V}_{X,i} \in \mathbb{R}^{K \times K}, \quad \hat{V}_{Y,i} \in \mathbb{R}, \quad \hat{C}_{XY,i} \in \mathbb{R}^K, \quad \hat{C}_{YV,i} \in \mathbb{R}, \quad \hat{K}_{Y,i} \in \mathbb{R}$$

denote estimators of $\text{Var}(\hat{\mu}_X(S_i) \mid S_i)$, $\text{Var}(\hat{\mu}_Y(S_i) \mid S_i)$, $\text{Cov}(\hat{\mu}_X(S_i), \hat{\mu}_Y(S_i) \mid S_i)$, $\text{Cov}(\hat{\mu}_Y(S_i), \hat{V}_{Y,i} \mid S_i)$ and $\mathbb{E}[\{\hat{\mu}_Y(S_i) - \mu_Y(S_i)\}^3 \mid S_i]$, respectively.

Remark 1 (Predictors measured without error). *The framework allows some or all coordinates of $\mu_X(S_i)$ to be measured without error. For every such coordinate k , one simply has*

$$\hat{\mu}_{X,k}(S_i) = \mu_{X,k}(S_i).$$

The corresponding row and column of $\hat{V}_{X,i}$ and the corresponding coordinate of $\hat{C}_{XY,i}$ are then equal to zero.

Assumption 1 (Sampling and unbiasedness). *For every $i = 1, \dots, n$:*

(i) *We observe the vector*

$$(S_i, W(S_i), \hat{\mu}_X(S_i), \hat{\mu}_Y(S_i), \hat{V}_{X,i}, \hat{V}_{Y,i}, \hat{C}_{XY,i}, \hat{C}_{YV,i}, \hat{K}_{Y,i}).$$

Those vectors are i.i.d. across i .

(ii) *Almost surely, all the following conditional unbiasedness restrictions hold:*

$$\begin{aligned}\mathbb{E}[\hat{\mu}_X(S_i) | S_i] &= \mu_X(S_i), & \mathbb{E}[\hat{\mu}_Y(S_i) | S_i] &= \mu_Y(S_i), \\ \mathbb{E}[\hat{V}_{X,i} | S_i] &= \text{Var}(\hat{\mu}_X(S_i) | S_i), & \mathbb{E}[\hat{V}_{Y,i} | S_i] &= \text{Var}(\hat{\mu}_Y(S_i) | S_i), \\ \mathbb{E}[\hat{C}_{XY,i} | S_i] &= \text{Cov}(\hat{\mu}_X(S_i), \hat{\mu}_Y(S_i) | S_i), & \mathbb{E}[\hat{C}_{YV,i} | S_i] &= \text{Cov}(\hat{\mu}_Y(S_i), \hat{V}_{Y,i} | S_i), \\ \mathbb{E}[\hat{K}_{Y,i} | S_i] &= \mathbb{E}[\{\hat{\mu}_Y(S_i) - \mu_Y(S_i)\}^3 | S_i].\end{aligned}$$

Section 3.1.1 of Fuller (1987) considers a closely related setup in which a dependent variable and a vector of regressors are observed with measurement error, and where the variance of the measurement errors can be unbiasedly estimated. We further assume that the third centered moment of the estimation error and its covariance with its variance estimator can be unbiasedly estimated. These additional restrictions allow us to construct the corrected estimator of κ_Y below.

Lemma 1. *Under Assumption 1, provided the expectations below exist,*

$$\mathbb{E}_w[\{\hat{\mu}_Y(S_i) - \mathbb{E}_w[\mu_Y(S)]\}^2 - \hat{V}_{Y,i}] = \text{Var}_w(\mu_Y(S)), \quad (1)$$

$$\mathbb{E}_w[\{\hat{\mu}_X(S_i) - \mathbb{E}_w[\mu_X(S)]\}\{\hat{\mu}_X(S_i) - \mathbb{E}_w[\mu_X(S)]^\top - \hat{V}_{X,i}] = \text{Var}_w(\mu_X(S)), \quad (2)$$

$$\mathbb{E}_w[\{\hat{\mu}_X(S_i) - \mathbb{E}_w[\mu_X(S)]\}\{\hat{\mu}_Y(S_i) - \mathbb{E}_w[\mu_Y(S)]\} - \hat{C}_{XY,i}] = \text{Cov}_w(\mu_X(S), \mu_Y(S)), \quad (3)$$

$$\mathbb{E}_w[\{\hat{\mu}_Y(S_i) - \mathbb{E}_w[\mu_Y(S)]\}^3 - 3\{\hat{\mu}_Y(S_i) - \mathbb{E}_w[\mu_Y(S)]\}\hat{V}_{Y,i} + 3\hat{C}_{YV,i} - \hat{K}_{Y,i}] = \kappa_Y. \quad (4)$$

Equations (1)-(3) follow from the laws of total variance and covariance, combined with Assumption 1. Equation (4) additionally follows by expanding a cube.

2.2 Estimators

Let

$$\hat{w}(S_i) = \frac{W(S_i)}{\frac{1}{n} \sum_{j=1}^n W(S_j)}, \quad \bar{\mu}_X = \frac{1}{n} \sum_{i=1}^n \hat{w}(S_i) \hat{\mu}_X(S_i), \quad \bar{\mu}_Y = \frac{1}{n} \sum_{i=1}^n \hat{w}(S_i) \hat{\mu}_Y(S_i).$$

We estimate σ^2 by

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{w}(S_i) [\{\hat{\mu}_Y(S_i) - \bar{\mu}_Y\}^2 - \hat{V}_{Y,i}] = \frac{1}{n} \sum_{i=1}^n \hat{w}(S_i) \{\hat{\mu}_Y(S_i) - \bar{\mu}_Y\}^2 - \frac{1}{n} \sum_{i=1}^n \hat{w}(S_i) \hat{V}_{Y,i}. \quad (5)$$

$\hat{\sigma}^2$ is a corrected estimator of the variance of $\mu_Y(S)$: it subtracts the average estimated variance of $\hat{\mu}_Y(S_i)$ from the variance of $\hat{\mu}_Y(S_i)$ across the n observations.

Similarly, we consider the following corrected estimators of $\text{Var}_w(\mu_X(S))$ and $\text{Cov}_w(\mu_X(S), \mu_Y(S))$:

$$\widehat{\text{Var}}(\mu_X(S)) = \frac{1}{n} \sum_{i=1}^n \hat{w}(S_i) [\{\hat{\mu}_X(S_i) - \bar{\mu}_X\}\{\hat{\mu}_X(S_i) - \bar{\mu}_X\}^\top - \hat{V}_{X,i}], \quad (6)$$

$$\widehat{\text{Cov}}(\mu_X(S), \mu_Y(S)) = \frac{1}{n} \sum_{i=1}^n \hat{w}(S_i) [\{\hat{\mu}_X(S_i) - \bar{\mu}_X\}\{\hat{\mu}_Y(S_i) - \bar{\mu}_Y\} - \hat{C}_{XY,i}]. \quad (7)$$

Whenever $\widehat{\text{Var}}(\mu_X(S))$ is nonsingular, let

$$\widehat{\beta} = \widehat{\text{Var}}(\mu_X(S))^{-1} \widehat{\text{Cov}}(\mu_X(S), \mu_Y(S)). \quad (8)$$

Finally, following (4) we estimate κ_Y by

$$\widehat{\kappa}_Y = \frac{1}{n} \sum_{i=1}^n \widehat{w}(S_i) \left[\{\widehat{\mu}_Y(S_i) - \bar{\mu}_Y\}^3 - 3\{\widehat{\mu}_Y(S_i) - \bar{\mu}_Y\} \widehat{V}_{Y,i} + 3\widehat{C}_{YV,i} - \widehat{K}_{Y,i} \right]. \quad (9)$$

When $\widehat{\sigma}^2 > 0$, one can use $\widehat{\kappa}_Y / (\widehat{\sigma}^2)^{3/2}$ to estimate the skewness of $\mu_Y(S)$.

In Appendix A, we establish estimators' asymptotic normality and their influence functions, under an asymptotic sequence where n goes to infinity but the variances of $\widehat{\mu}_Y(S_i)$ and $\widehat{\mu}_X(S_i)$ does not go to zero. In Appendix B, we report simulations assessing the finite-sample validity of the confidence intervals (CIs) based on these influence functions, in DGPs where site effects can have light or heavy tails, and no or a lot of skewness. The 95%-level CI for β has coverage close to its nominal level in every design considered, with as few as 50 sites. The 95%-level CI for σ^2 can slightly undercover with 50 and 100 sites, but has coverage close to its nominal level with 200 sites. With a studentized site-level bootstrap, the CI for σ^2 has coverage close to its nominal level even with 50 or 100 sites. Finally, with 100 sites or more, the 95%-level CI for κ_Y has coverage close to its nominal level in all DGPs, except for the one with both heavy tails and a lot of skewness. The studentized site-level bootstrap does not improve coverage for that parameter, except in the DGP with both heavy tails and a lot of skewness.

3 Application to multi-site RCTs

In this section, we show that the results of the previous section can be applied to multi-site RCTs.

Assume that n sites $(S_1, \dots, S_n)$ are drawn with replacement from a super-population of sites. In each site, N_i units are drawn with replacement from the site's super-population of units. In each site, we allocate $N_{0,i}$ units to a control group and $N_{d,i}$ units to treatment arm d , for $d \in \{1, \dots, \mathcal{D}\}$, $\mathcal{D} \geq 1$, where $N_{d,i} \geq 3$ for every $d \in \{0, \dots, \mathcal{D}\}$, and $\sum_{d=0}^{\mathcal{D}} N_{d,i} = N_i$. For each unit $u \in \{1, \dots, N_i\}$, let $D_{u,i} \in \{0, \dots, \mathcal{D}\}$ denote the treatment arm it is assigned to. For every $i \in \{1, \dots, n\}$ and $u \in \{1, \dots, N_i\}$, we measure a vector of outcomes $\mathbf{O}_{u,i} \in \mathbb{R}^L$, $L \geq 1$, with coordinates $O_{u,i,1}, \dots, O_{u,i,L}$. For every $d \in \{0, \dots, \mathcal{D}\}$, let $\mathbf{O}_{u,i}(d)$ denote the potential-outcome vector of unit u in site i under treatment d . The observed outcome vector satisfies $\mathbf{O}_{u,i} = \mathbf{O}_{u,i}(D_{u,i})$. Under our assumptions on the sampling of sites and units within sites, we have:

- Assumption 2.** 1. The vectors $(S_i, N_i, (N_{d,i})_{d=0}^{\mathcal{D}}, ((\mathbf{O}_{u,i}(d))_{d=0}^{\mathcal{D}})_{u=1}^{N_i})$ are i.i.d. across i .
2. For all i , the vectors $(\mathbf{O}_{u,i}(d))_{d=0}^{\mathcal{D}}$ are i.i.d. across u conditional on S_i , and we let $(\mathbf{O}_i(d))_{d=0}^{\mathcal{D}}$ denote a vector with the same probability distribution as $(\mathbf{O}_{u,i}(d))_{d=0}^{\mathcal{D}}$ conditional on S_i .
3. For every $d \in \{0, \dots, \mathcal{D}\}$, $\mathbb{E}[\|\mathbf{O}_i(d)\|^3 \mid S_i] < \infty$ almost surely.

In multi-site RCTs, participating sites and units are not always probability samples from well-defined populations. In that case, Assumption 2 can be interpreted as a super-population modelling device: the observed sites and units are regarded as draws from hypothetical populations of sites and units that could have participated. This interpretation does not make the sample representative

of any particular external population. The i.i.d. sampling assumption within sites plays a key role: together with random assignment, it makes the usual variance estimator for a difference in means unbiased for its sampling variance around the site's super-population ATE. With fixed sampled units and randomness arising only from treatment assignment, this variance estimator is generally conservative, so subtracting it need not recover the cross-site variance of treatment effects. The assumption that sites are an i.i.d. sample is less crucial. Versions 1–3 of this paper (arXiv:2405.17254) instead treated the observed sites as a fixed finite population. For results common to those versions and the present one, moving to i.i.d. sampling leaves the estimators unchanged, but changes the interpretation of their variance estimators: under the finite-population framework, the estimators' asymptotic variances could in general only be conservatively estimated, whereas under Assumption 2 they can be consistently estimated.

Let

$$\mathcal{V}_i = \left(S_i, N_i, (N_{d,i})_{d=0}^{\mathcal{D}}, ((\mathbf{O}_{u,i}(d))_{d=0}^{\mathcal{D}})_{u=1}^{N_i} \right)$$

collect site i 's identifier, its number of units, its arm sizes, and its sampled potential outcomes. To allocate units to treatment arms, a completely randomized experiment is conducted in each site, independently across sites:

Assumption 3. N_i and the arm sizes $(N_{0,i}, \dots, N_{\mathcal{D},i})$ are measurable with respect to S_i . For every $i \in \{1, \dots, n\}$ and every assignment vector $\mathbf{d} = (d_1, \dots, d_{N_i}) \in \{0, \dots, \mathcal{D}\}^{N_i}$ such that $\sum_{u=1}^{N_i} \mathbf{1}\{d_u = d\} = N_{d,i}$ for every $d \in \{0, \dots, \mathcal{D}\}$,

$$\mathbb{P}\left((D_{u,i})_{u=1}^{N_i} = \mathbf{d} \mid (\mathcal{V}_{i'})_{i'=1}^n\right) = \binom{N_i}{N_{0,i}, \dots, N_{\mathcal{D},i}}^{-1} = \frac{\prod_{d=0}^{\mathcal{D}} N_{d,i}!}{N_i!} \quad \text{almost surely.}$$

Moreover, conditional on $(\mathcal{V}_i)_{i=1}^n$, the assignment vectors $(D_{u,i})_{u=1}^{N_i}$ are independent across i .

Assumption 3 states that within every site, a lottery assigns a fixed number of units to each arm, as in a stratified experiment with sites as strata. Importantly, the assumptions let sites differ arbitrarily in size, in the share of units assigned to each arm, and in the effects of the treatments. Moreover, the number of units per site can be small.

Let $\mu_O(S_i) = (\mathbb{E}[\mathbf{O}_i(d) \mid S_i])_{d=0}^{\mathcal{D}}$ denote the $L(\mathcal{D} + 1)$ vector of the average outcomes $(O_{i,1}, \dots, O_{i,L})$ under treatments $(0, \dots, \mathcal{D})$ in site i . The corresponding estimator is the $L(\mathcal{D} + 1)$ vector of arm-specific sample means

$$\hat{\mu}_O(S_i) = \left(\frac{1}{N_{d,i}} \sum_{u=1}^{N_i} \mathbf{1}\{D_{u,i} = d\} \mathbf{O}_{u,i} \right)_{d=0}^{\mathcal{D}}.$$

To estimate $\text{Var}(\hat{\mu}_O(S_i) \mid S_i)$, we use $\hat{V}_{O,i}$, an $L(\mathcal{D} + 1) \times L(\mathcal{D} + 1)$ matrix whose entries are

$$[\hat{V}_{O,i}]_{(d,\ell),(d',\ell')} = \mathbf{1}\{d = d'\} \frac{1}{N_{d,i}(N_{d,i} - 1)} \sum_{u=1}^{N_i} \mathbf{1}\{D_{u,i} = d\} (O_{u,i,\ell} - \bar{O}_{d,i,\ell}) (O_{u,i,\ell'} - \bar{O}_{d,i,\ell'}),$$

for every $(d, d') \in \{0, \dots, \mathcal{D}\}^2$ and $(\ell, \ell') \in \{1, \dots, L\}^2$, where

$$\bar{O}_{d,i,\ell} = \frac{1}{N_{d,i}} \sum_{u=1}^{N_i} \mathbf{1}\{D_{u,i} = d\} O_{u,i,\ell}.$$

Under i.i.d. sampling within sites and completely randomized assignment, sample means from distinct arms are independent conditional on S_i , after averaging over both sampling and assignment. Their conditional covariances are therefore zero, explaining the block-diagonal form of $\widehat{V}_{O,i}$.

Finally, to estimate centered third moments, for any non-random $L(\mathcal{D}+1)$ vector $c = (c_0^\top, \dots, c_{\mathcal{D}}^\top)^\top$, with $c_d \in \mathbb{R}^L$, let

$$\widehat{M}_{3,O,i}(c) = \sum_{d=0}^{\mathcal{D}} \frac{1}{N_{d,i}(N_{d,i}-1)(N_{d,i}-2)} \sum_{u=1}^{N_i} \mathbf{1}\{D_{u,i} = d\} \left[c_d^\top \{ \mathbf{O}_{u,i} - \overline{\mathbf{O}}_{d,i} \} \right]^3,$$

where $\overline{\mathbf{O}}_{d,i} = N_{d,i}^{-1} \sum_u \mathbf{1}\{D_{u,i} = d\} \mathbf{O}_{u,i}$.

Proposition 1. *Under Assumptions 2 and 3, for every i ,*

$$\mathbb{E}[\widehat{\mu}_O(S_i) \mid S_i] = \mu_O(S_i), \quad (10)$$

$$\mathbb{E}[\widehat{V}_{O,i} \mid S_i] = \text{Var}(\widehat{\mu}_O(S_i) \mid S_i), \quad (11)$$

$$\mathbb{E}[\widehat{M}_{3,O,i}(c) \mid S_i] = \mathbb{E}[\{c^\top (\widehat{\mu}_O(S_i) - \mu_O(S_i))\}^3 \mid S_i] \quad \text{for every non-random } c. \quad (12)$$

Let c and B be a non-random $L(\mathcal{D}+1)$ vector and a non-random $K \times L(\mathcal{D}+1)$ matrix, respectively, and let

$$Y_i = c^\top (\mathbf{O}_i(d))_{d=0}^{\mathcal{D}}, \quad X_i = B(\mathbf{O}_i(d))_{d=0}^{\mathcal{D}}.$$

Define

$$\begin{aligned} \widehat{\mu}_Y(S_i) &= c^\top \widehat{\mu}_O(S_i), & \widehat{\mu}_X(S_i) &= B \widehat{\mu}_O(S_i), \\ \widehat{V}_{Y,i} &= c^\top \widehat{V}_{O,i} c, & \widehat{V}_{X,i} &= B \widehat{V}_{O,i} B^\top, \\ \widehat{C}_{XY,i} &= B \widehat{V}_{O,i} c, & \widehat{C}_{YV,i} &= \widehat{K}_{Y,i} = \widehat{M}_{3,O,i}(c). \end{aligned}$$

Then,

$$(S_i, W(S_i), \widehat{\mu}_X(S_i), \widehat{\mu}_Y(S_i), \widehat{V}_{X,i}, \widehat{V}_{Y,i}, \widehat{C}_{XY,i}, \widehat{C}_{YV,i}, \widehat{K}_{Y,i})$$

satisfies Assumption 1.

Proposition 1 follows from the usual unbiasedness of sample means, sample covariances, and third centered moments. $\widehat{K}_{Y,i}$ and $\widehat{C}_{YV,i}$ are equal, because the centered third moment of a sample mean equals its covariance with its estimated variance under i.i.d. sampling.

For the applications below, write

$$\widehat{C}_{d,i,\ell\ell'} = [\widehat{V}_{O,i}]_{(d,\ell),(d,\ell')}, \quad \widehat{V}_{d,i,\ell} = \widehat{C}_{d,i,\ell\ell}.$$

These estimate the sampling covariance of the arm- d means of outcomes ℓ and ℓ' , and the sampling variance of the arm- d mean of outcome ℓ , respectively. With one outcome, abbreviate $\widehat{V}_{d,i,1}$ as $\widehat{V}_{d,i}$.

As an example, with one treatment arm and one outcome ($\mathcal{D} = 1, L = 1$),

$$\widehat{\mu}_O(S_i) = (\overline{O}_{0,i,1}, \overline{O}_{1,i,1})^\top$$

and $\widehat{V}_{O,i}$ is diagonal, with the estimated variances of the two averages on its diagonal. Letting $c = (-1, 1)^\top$ and $B = (1, 0)$ yields $\widehat{\mu}_Y(S_i)$ equal to the site's estimated ATE and $\widehat{\mu}_X(S_i)$ equal

to its control-group mean. Then $\widehat{V}_{Y,i} = \mathbf{c}^\top \widehat{V}_{O,i} \mathbf{c} = \widehat{V}_{0,i} + \widehat{V}_{1,i}$ recovers the standard variance formula for a difference in means, and $\widehat{C}_{XY,i} = B \widehat{V}_{O,i} \mathbf{c} = -\widehat{V}_{0,i}$ captures the mechanical negative covariance between the difference in means and the control group average. These variance and covariance corrections can also be obtained by specializing the leave-out covariance estimator in the random-coefficient model in Kline, Saggio, and Sølvesten (2020) to a binary regressor. Finally, with $\mathbf{c} = (-1, 1)^\top$

$$\begin{aligned} \widehat{C}_{YV,i} = \widehat{K}_{Y,i} = & \frac{1}{N_{1,i}(N_{1,i} - 1)(N_{1,i} - 2)} \sum_{u=1}^{N_i} \mathbf{1}\{D_{u,i} = 1\} (O_{u,i,1} - \overline{O}_{1,i,1})^3 \\ & - \frac{1}{N_{0,i}(N_{0,i} - 1)(N_{0,i} - 2)} \sum_{u=1}^{N_i} \mathbf{1}\{D_{u,i} = 0\} (O_{u,i,1} - \overline{O}_{0,i,1})^3. \end{aligned}$$

Proposition 1 shows that the corrected-moment estimators in the previous section are applicable to multi-site experiments, whenever $\mu_Y(S)$ and $\mu_X(S)$ are known linear transformations of arm-specific average potential outcomes. The site-level estimates and their sampling-moment corrections must be unbiased but need not be consistent as the number of sites grows.

Remark 2 (Possible choices for the weights). *One may let the weights $W(S_i)$ be equal to N_i , thus giving more weight to sites with more units. One could also let $W(S_i) = 1$, thus giving the same weight to each site. Those are two common choices of weights, but other choices are possible.*

Remark 3 (Site-level predictors measured without error). *As noted in Remark 1, X_i may also include site-specific variables that do not need to be estimated and are measured without error. Examples include the unemployment rate or GDP per capita in site i . For those coordinates, the corresponding variance and covariance corrections are zero.*

Remark 4 (Minimum number of units per arm). *The maintained requirement $N_{d,i} \geq 3$ is needed only to construct $\widehat{M}_{3,O,i}(c)$, and hence to estimate κ_Y . The variance and regression estimators $\widehat{\sigma}^2$ and $\widehat{\beta}$ require only $N_{d,i} \geq 2$ in every arm.*

We now highlight six potentially relevant applications of Theorems 1, 2, 3 and Proposition 1.

3.1 Estimating the variance and centered third moment of ATEs across sites

In this subsection, we momentarily assume that $\mathcal{D} = 1$ and $L = 1$: the RCT has one treatment arm, and we are interested in the treatment effect on one outcome. Let $\sigma_{\text{ATE}}^2 = \text{Var}_w(\mathbb{E}[O_{i,1}(1) - O_{i,1}(0) | S_i])$ denote the weighted variance across sites of $\mathbb{E}[O_{i,1}(1) - O_{i,1}(0) | S_i]$, the average effect of the treatment on $O_{i,1}$ in site i . Estimating this variance reveals whether the overall ATE summarizes effects that are similar across sites or instead masks differences across sites. Let

$$\kappa_{\text{ATE}} = \mathbb{E}_w \left[\left\{ \mathbb{E}[O_{i,1}(1) - O_{i,1}(0) | S_i] - \mathbb{E}_w[\mathbb{E}[O_{i,1}(1) - O_{i,1}(0) | S_i]] \right\}^3 \right]$$

denote the corresponding weighted centered third moment. Estimating κ_{ATE} reveals whether there is some asymmetry in the distribution of treatment effects across sites, with unusually large effects concentrated above or below the overall ATE.

To estimate σ_{ATE}^2 , one lets

$$Y_i = O_{i,1}(1) - O_{i,1}(0).$$

The corresponding estimator of $\mu_Y(S_i)$ and variance correction are

$$\hat{\mu}_Y(S_i) = \bar{O}_{1,i,1} - \bar{O}_{0,i,1}, \quad \hat{V}_{Y,i} = \hat{V}_{1,i} + \hat{V}_{0,i}.$$

Plugging those into (5) yields an estimator of σ_{ATE}^2 .

To estimate κ_{ATE} , Proposition 1 and Theorem 3 use the same difference-in-means estimator and variance correction, together with

$$\begin{aligned} \hat{C}_{YV,i} = \hat{K}_{Y,i} = & \frac{1}{N_{1,i}(N_{1,i} - 1)(N_{1,i} - 2)} \sum_{u=1}^{N_i} \mathbf{1}\{D_{u,i} = 1\} (O_{u,i,1} - \bar{O}_{1,i,1})^3 \\ & - \frac{1}{N_{0,i}(N_{0,i} - 1)(N_{0,i} - 2)} \sum_{u=1}^{N_i} \mathbf{1}\{D_{u,i} = 0\} (O_{u,i,1} - \bar{O}_{0,i,1})^3. \end{aligned}$$

Plugging these quantities into (9) yields an estimator of κ_{ATE} .

3.2 Regressing sites' ATEs on their average untreated outcome

In this subsection, we again assume that $\mathcal{D} = 1$ and $L = 1$. Let $\beta_{\text{ATE},0}$ denote the coefficient on $\mathbb{E}[O_{i,1}(0)|S_i]$ in a regression of $\mathbb{E}[O_{i,1}(1) - O_{i,1}(0)|S_i]$ on an intercept and $\mathbb{E}[O_{i,1}(0)|S_i]$, weighted by $W(S_i)$. $\mathbb{E}[O_{i,1}(0)|S_i]$ measures how well the site would fare in the absence of the intervention, so the regression reveals whether treatment effects tend to be larger or smaller in sites with worse untreated outcomes. This regression may also inform targeting outside the RCT if the relationship between treatment effects and untreated outcomes carries over to the new sites, and if outcomes among previous cohorts provide a useful proxy for new sites' untreated outcomes. Then, if $\beta_{\text{ATE},0}$ is, say, negative, one may prioritize non-RCT sites with the worst outcomes among previous cohorts.

To estimate $\beta_{\text{ATE},0}$, one lets

$$Y_i = O_{i,1}(1) - O_{i,1}(0), \quad X_i = O_{i,1}(0).$$

The corresponding estimators of $\mu_Y(S_i)$ and $\mu_X(S_i)$ and variance-covariance corrections are

$$\begin{aligned} \hat{\mu}_Y(S_i) &= \bar{O}_{1,i,1} - \bar{O}_{0,i,1}, & \hat{\mu}_X(S_i) &= \bar{O}_{0,i,1}, \\ \hat{V}_{Y,i} &= \hat{V}_{1,i} + \hat{V}_{0,i}, & \hat{V}_{X,i} &= \hat{V}_{0,i}, & \hat{C}_{XY,i} &= -\hat{V}_{0,i}. \end{aligned}$$

Plugging those into (8) yields an estimator of $\beta_{\text{ATE},0}$. Note that a naive regression of estimated effects on estimated untreated outcomes is generally biased. In particular, the control-group mean enters positively in the estimator of the untreated outcome and negatively in the difference-in-means estimator, making their estimation errors mechanically negatively correlated. $\hat{\beta}_{\text{ATE},0}$ removes this covariance, as well as the usual attenuation bias coming from measurement error in the estimation of $\mathbb{E}[O_{i,1}(0) | S_i]$.

3.3 Predictive and causal mediation analysis

In this subsection, we momentarily assume that $\mathcal{D} = 1$ and $L = 2$: the RCT has one treatment arm, and we are interested in the treatment effect on two outcomes. $O_{i,2}$ is the final outcome and $O_{i,1}$ is an intermediate outcome that might mediate the effect of the treatment on the final outcome.

3.3.1 Predictive mediation

Let β_{med} denote the coefficient on $\mathbb{E}[O_{i,1}(1) - O_{i,1}(0) \mid S_i]$ in a regression of $\mathbb{E}[O_{i,2}(1) - O_{i,2}(0) \mid S_i]$ on an intercept and $\mathbb{E}[O_{i,1}(1) - O_{i,1}(0) \mid S_i]$, weighted by $W(S_i)$. This yields a predictive mediation regression, which assesses whether sites in which the treatment has larger effects on the mediator also tend to be sites in which the treatment has larger effects on the final outcome. For example, in a job-search counseling RCT, one may ask whether sites with larger effects on job seekers' search effort also have larger effects on their job-finding rate. β_{med} can be helpful to generate hypotheses about the mechanisms underlying the treatment effect. However, additional assumptions are needed for β_{med} to identify the mediator's effect.

To estimate β_{med} , one lets

$$Y_i = O_{i,2}(1) - O_{i,2}(0), \quad X_i = O_{i,1}(1) - O_{i,1}(0).$$

The corresponding estimators of $\mu_Y(S_i)$ and $\mu_X(S_i)$ and variance-covariance corrections are

$$\begin{aligned} \hat{\mu}_Y(S_i) &= \bar{O}_{1,i,2} - \bar{O}_{0,i,2}, & \hat{\mu}_X(S_i) &= \bar{O}_{1,i,1} - \bar{O}_{0,i,1}, \\ \hat{V}_{Y,i} &= \hat{V}_{1,i,2} + \hat{V}_{0,i,2}, & \hat{V}_{X,i} &= \hat{V}_{1,i,1} + \hat{V}_{0,i,1}, \\ \hat{C}_{XY,i} &= \hat{C}_{1,i,12} + \hat{C}_{0,i,12}. \end{aligned}$$

Plugging those into (8) yields an estimator of β_{med} . Note that one could also estimate a predictive mediation regression with several mediators. Then, μ_X is a vector collecting site-specific effects on all mediators. Proposition 1 supplies the corresponding matrix of variance corrections and vector of covariance corrections.

3.3.2 Causal mediation

We now give conditions under which the coefficients from the predictive-mediation regression have a causal interpretation. For a generic unit in site i , let $O_{i,1}(d)$ denote the potential mediator under treatment d , and let $O_{i,2}(d, o_1)$ denote the potential final outcome when treatment is set to d and the mediator is set to o_1 . Thus, the potential final outcome under treatment d defined above satisfies $O_{i,2}(d) = O_{i,2}(d, O_{i,1}(d))$, and $\mu_Y(S_i) = \mathbb{E}[O_{i,2}(1, O_{i,1}(1)) - O_{i,2}(0, O_{i,1}(0)) \mid S_i]$.

Assumption 4 (Binary mediator and monotonicity). *$O_{i,1}$ is binary and satisfies the monotonicity condition*

$$O_{i,1}(1) \geq O_{i,1}(0) \quad \text{almost surely.}$$

Moreover,

$$\mathbb{P}\{O_{i,1}(1) > O_{i,1}(0) \mid S_i\} > 0 \quad \text{almost surely.}$$

For site i , define

$$\begin{aligned} \text{DE}(S_i) &= \mathbb{E}[O_{i,2}(1, O_{i,1}(0)) - O_{i,2}(0, O_{i,1}(0)) \mid S_i], \\ \text{IDE}(S_i) &= \mathbb{E}[O_{i,2}(1, 1) - O_{i,2}(1, 0) \mid O_{i,1}(1) > O_{i,1}(0), S_i]. \end{aligned}$$

$\text{DE}(S_i)$ is the site's average natural direct effect: it measures the effect of switching treatment from zero to one while holding the mediator at the value it would take under control. $\text{IDE}(S_i)$ is the average effect on the final outcome of switching the mediator from zero to one while holding

treatment fixed at one, among units in site i whose mediator is switched from zero to one by the treatment. By the usual natural direct–indirect effect decomposition (see, e.g., VanderWeele, 2013),

$$\begin{aligned}\mu_Y(S_i) &= \text{DE}(S_i) + \mathbb{E}[O_{i,2}(1, O_{i,1}(1)) - O_{i,2}(1, O_{i,1}(0)) \mid S_i] \\ &= \text{DE}(S_i) + \mathbb{E}[\mathbf{1}\{O_{i,1}(1) > O_{i,1}(0)\} \{O_{i,2}(1, 1) - O_{i,2}(1, 0)\} \mid S_i] \\ &= \text{DE}(S_i) + \mu_X(S_i) \text{IDE}(S_i).\end{aligned}\tag{13}$$

The first equality follows by adding and subtracting $O_{i,2}(1, O_{i,1}(0))$. The second equality follows from Assumption 4, and the third from the law of iterated expectations and the fact that monotonicity implies $\mathbb{P}\{O_{i,1}(1) > O_{i,1}(0) \mid S_i\} = \mathbb{E}[O_{i,1}(1) - O_{i,1}(0) \mid S_i] = \mu_X(S_i)$.

Assumption 5 (Cross-site independence for causal mediation). *Under the probability distribution that weights sites by $w(S)$, $\mu_X(S)$ is independent of $(\text{IDE}(S), \text{DE}(S))$.*

Assumption 5 requires that sites' effects on the mediator be unrelated to both their direct effects and their compliers' gains from the mediator.

Proposition 2 (Causal interpretation of the predictive-mediation regression). *Suppose that Assumptions 4 and 5 hold, and that $\text{Var}_w(\mu_X(S)) > 0$. Then the slope β_{med} and intercept α_{med} from the weighted population regression of $\mu_Y(S)$ on an intercept and $\mu_X(S)$ satisfy*

$$\beta_{\text{med}} = \mathbb{E}_w[\text{IDE}(S)], \quad \alpha_{\text{med}} = \mathbb{E}_w[\text{DE}(S)].$$

Under Assumptions 4 and 5, β_{med} identifies the w -weighted average across sites of the mediator's effect among each site's mediator compliers. The corresponding average indirect effect of treatment on the final outcome is therefore also identified: it equals the average treatment effect on the mediator multiplied by the mediation regression slope:

$$\mathbb{E}_w[\mathbb{E}[O_{i,2}(1, O_{i,1}(1)) - O_{i,2}(1, O_{i,1}(0)) \mid S_i]] = \mathbb{E}_w[\mu_X(S_i) \text{IDE}(S_i)] = \mathbb{E}_w[\mu_X(S_i)] \beta_{\text{med}}.$$

3.3.3 Testing homogeneity of direct and mediator effects

Consider the condition

$$\text{IDE}(S_i) = \text{IDE}, \quad \text{DE}(S_i) = \text{DE} \quad \text{almost surely.}\tag{14}$$

This condition requires both the direct effect and the mediator's effect among mediator compliers to be constant across sites. It is sufficient for the cross-site independence assumption in Assumption 5. We derive a testable implication of this homogeneity condition, maintaining Assumption 4. Without that maintained assumption, rejection can instead reflect a failure of either condition.

Proposition 3 (Testable implication of homogeneous direct and mediator effects). *Suppose that Assumption 4 and (14) hold, and that $\text{Var}_w(\mu_X(S)) > 0$. Then $\text{IDE} = \beta_{\text{med}}$,*

$$\mu_Y(S) - \mu_X(S) \beta_{\text{med}} = \text{DE}(S) + \mu_X(S) \{\text{IDE}(S) - \beta_{\text{med}}\} = \text{DE} \quad \text{almost surely,}\tag{15}$$

and therefore

$$\text{Var}_w(\mu_Y(S) - \mu_X(S) \beta_{\text{med}}) = 0.\tag{16}$$

Under (14) it follows from Proposition 2 that $\text{IDE} = \beta_{\text{med}}$. Then the first equality in (15) follows from (13), while the second follows from the fact that $\text{DE}(S) = \text{DE}$ and $\text{IDE}(S) = \text{IDE} = \beta_{\text{med}}$.

Remark 5 (Power of a test based on (16)). *Under the regularity conditions below, a test based on (16) has power tending to one against fixed violations of (14) that generate a strictly positive residual variance. It cannot consistently detect violations that preserve an exact linear relationship between $\mu_Y(S)$ and $\mu_X(S)$. Let us characterize such alternatives under the following models for $\text{DE}(S)$ and $\text{IDE}(S)$:*

$$\text{DE}(S) = \gamma_0 + \gamma_1 \mu_X(S), \quad \text{IDE}(S) = \theta_0 + \theta_1 \mu_X(S).$$

Suppose also that $\mu_X(S)$ has at least three support points. Then, (16) holds if and only if $\theta_1 = 0$.⁵ Therefore, (14) fails while (16) holds if and only if $\theta_1 = 0$ and $\gamma_1 \neq 0$. However, one may expect heterogeneity associated with the share of mediator compliers to arise more readily in the mediator's effect among compliers than in the treatment's direct effect. An additional assumption motivated by this view is

$$\theta_1 = 0 \implies \gamma_1 = 0.$$

This restriction rules out variation in the direct effect with the share of mediator compliers when the mediator effect does not vary with that share. Under the linear model above, the requirement of at least three support points for $\mu_X(S)$, and this additional restriction, (16) holds if and only if (14) holds. Thus, within this restricted model, homogeneity is fully testable.

Turning to the test's implementation, let

$$\begin{aligned} \hat{\mu}_\varepsilon(S_i) &= \hat{\mu}_Y(S_i) - \hat{\beta}_{\text{med}} \hat{\mu}_X(S_i), \\ \hat{V}_{\varepsilon,i} &= \hat{V}_{Y,i} + \hat{\beta}_{\text{med}}^2 \hat{V}_{X,i} - 2\hat{\beta}_{\text{med}} \hat{C}_{XY,i}. \end{aligned}$$

To test (14), we use

$$\hat{\sigma}_\varepsilon^2 := \frac{1}{n} \sum_{i=1}^n \hat{w}(S_i) \left[\{\hat{\mu}_\varepsilon(S_i) - \hat{\tau}_\varepsilon\}^2 - \hat{V}_{\varepsilon,i} \right], \quad \text{where } \hat{\tau}_\varepsilon := \frac{1}{n} \sum_{i=1}^n \hat{w}(S_i) \hat{\mu}_\varepsilon(S_i).$$

Appendix A.2 derives the influence function and asymptotic variance of $\hat{\sigma}_\varepsilon^2$ and describes how to use it to test homogeneity of direct and mediator effects. Under the maintained assumptions, rejection provides evidence against homogeneous direct and mediator effects, but does not necessarily reject Assumption 5. Conversely, nonrejection does not establish homogeneity or validate the causal interpretation of the regression coefficients.

3.3.4 Comparison to other approaches for mediation analysis in RCTs

Identifying the effect of a mediator on an outcome is notoriously difficult. Standard approaches compare outcomes across units with the same treatment assignment and covariates but different mediator values. Giving those comparisons a causal interpretation requires sequential ignorability: conditional on treatment and pretreatment covariates, the mediator must be as good as randomly

⁵Indeed, if $\theta_1 = 0$ then $\mu_Y(S) = \gamma_0 + (\gamma_1 + \theta_0)\mu_X(S)$, so $\beta_{\text{med}} = \gamma_1 + \theta_0$ and (16) holds. Conversely, if $\theta_1 \neq 0$, $\mu_Y(S)$ contains a nonzero quadratic term in $\mu_X(S)$ and therefore cannot be affine in $\mu_X(S)$ on three or more support points.

assigned (Imai, Keele, and Yamamoto, 2010). This condition is often hard to justify because the mediator is not randomized, even when the treatment is.

Our approach instead regresses site-level treatment effects on the outcome on site-level treatment effects on the mediator. Such regressions have a long history in meta-analyses (see, e.g., Daniels and Hughes, 1997). Prior work has shown that those regressions have a causal interpretation under the homogeneity condition in (14) (see, e.g., Dunn, Emsley, Liu, Landau, Green, White, and Pickles, 2015). Our Proposition 2 slightly extends those results by showing that this condition can be relaxed, though our Assumption 5 remains a strong assumption. To our knowledge, the procedure in Section 3.3.3 provides the first sampling-error-corrected test of a necessary implication of (14). This provides a way to assess a sufficient condition for the causal interpretation of those regressions.

Another approach treats the randomized treatment as an instrument for the mediator. In this context, the IV exclusion restriction requires that the treatment affects the outcome only through the mediator, referred to as the full mediation assumption. If i) the full mediation assumption holds, ii) the treatment is randomly assigned and iii) the treatment has a monotonic effect on the mediator, a 2SLS regression of the outcome on the mediator using the treatment as the instrument identifies the effect of the mediator for mediator compliers (Reardon and Raudenbush, 2013). Like Assumption 4 and (14), i), ii) and iii) are not fully testable but have a testable implication (Kwon and Roth, 2026). The full-mediation assumption and Assumption 5 are not nested. The former does not restrict the mediator’s effect on the outcome, but rules out any direct effect of treatment. Assumption 5, by contrast, permits direct effects, but requires that direct effects and mediator effects be independent of the treatment effect on the mediator.

3.4 Sites’ responsiveness or providers’ effectiveness?

In multi-site RCTs, treatment providers often vary across sites. Then, a natural question is whether heterogeneous effects across sites are driven by heterogeneity in providers’ effectiveness, or by sites’ heterogeneous responsiveness. In a multi-arm RCT, if in each site the treatments are delivered by different providers, as is the case in our empirical application, the variance and regression estimators proposed in this paper can help shed some light on this question.

In this subsection, we assume that $\mathcal{D} = 2$ and $L = 1$: we consider an RCT with two treatment arms and one outcome. For $d \in \{1, 2\}$, let $\Delta_d = O_1(d) - O_1(0)$. Then, let

$$\sigma_{\Delta}^2 = \text{Var}_w(\mu_{\Delta_2}(S_i) - \mu_{\Delta_1}(S_i))$$

denote the weighted variance across sites of the difference between the two treatments’ ATEs. Let also β_{21} denote the coefficient on $\mu_{\Delta_1}(S_i)$ in a weighted regression of $\mu_{\Delta_2}(S_i)$ on an intercept and $\mu_{\Delta_1}(S_i)$. A small σ_{Δ}^2 means that the difference between the two treatments’ effects varies little across sites, while a positive β_{21} means that the treatments tend to be more effective in the same sites. Both findings may indicate that treatment-effect heterogeneity is not entirely driven by heterogeneity in providers’ effectiveness.

To formalize this idea, assume that for all $d \in \{1, 2\}$,

$$\mu_{\Delta_d}(S_i) = f_d(P_d(S_i), U_d(S_i)), \quad (17)$$

where $P_d(S_i)$ is the effectiveness of the provider of treatment d in site S_i , $U_d(S_i)$ is a vector of characteristics of site S_i , and $f_d(\cdot)$ is a possibly non-separable production function mapping the provider’s

effectiveness and the site's characteristics into the site's ATE. Let $R_d(S_i) = f_d(\mathbb{E}[P_d(S_i)], U_d(S_i))$, where $\mathbb{E}[P_d(S_i)]$ is providers' average effectiveness: $R_d(S_i)$, the responsiveness of S_i to treatment d , is defined as the effect of d when implemented with average effectiveness in S_i . Let $\bar{R}(S_i) = (R_1(S_i) + R_2(S_i))/2$ denote the average responsiveness of S_i to the two treatments, let $\epsilon_d^S(S_i) = R_d(S_i) - \bar{R}(S_i)$ denote the difference between its responsiveness to d and its average responsiveness, so that $\epsilon_1^S(S_i) = -\epsilon_2^S(S_i) = (R_1(S_i) - R_2(S_i))/2$. Finally, let $\epsilon_d^P(S_i, P_d(S_i)) = f_d(P_d(S_i), U_d(S_i)) - R_d(S_i)$ denote the difference between its treatment- d ATE and what its treatment- d ATE would be with an average-effectiveness provider. It follows from these definitions that

$$\mu_{\Delta_d}(S_i) = \bar{R}(S_i) + \epsilon_d^S(S_i) + \epsilon_d^P(S_i, P_d(S_i)). \quad (18)$$

Proposition 4 (Does providers' effectiveness contribute to the variance of effects across sites?)

1. Assume that (17) holds.

(a) Then $\sigma_{\Delta}^2 = 0$ if and only if

$$\text{Var}_w(\epsilon_2^S(S_i) - \epsilon_1^S(S_i) + \epsilon_2^P(S_i, P_2(S_i)) - \epsilon_1^P(S_i, P_1(S_i))) = 0.$$

(b) If one further assumes that the w -weighted correlation between $\epsilon_2^S(S_i) - \epsilon_1^S(S_i)$ and $\epsilon_2^P(S_i, P_2(S_i)) - \epsilon_1^P(S_i, P_1(S_i))$ is strictly larger than -1 ,⁶ and that the weighted correlation between $\epsilon_2^P(S_i, P_2(S_i))$ and $\epsilon_1^P(S_i, P_1(S_i))$ is strictly lower than 1, then $\sigma_{\Delta}^2 = 0$ if and only if

$$\text{Var}_w(\epsilon_2^S(S_i) - \epsilon_1^S(S_i)) = 0, \quad \text{Var}_w(\epsilon_2^P(S_i, P_2(S_i))) = 0, \quad \text{Var}_w(\epsilon_1^P(S_i, P_1(S_i))) = 0.$$

2. If $\text{Var}_w(\mu_{\Delta_1}(S_i)) > 0$ and $\text{Var}_w(\mu_{\Delta_2}(S_i)) = \text{Var}_w(\mu_{\Delta_1}(S_i))$, then $\sigma_{\Delta}^2 = 0$ if and only if $\beta_{21} = 1$.

Point 1.a) of Proposition 4 follows directly from (18): differencing the two treatments' effects cancels $\bar{R}(S_i)$, so σ_{Δ}^2 only reflects sites' relative responsiveness to the two treatments and providers' contributions. Point 1.b) shows that, except for knife-edge cases, $\sigma_{\Delta}^2 = 0$ rules out providers' effectiveness as a source of treatment-effect heterogeneity across sites. The first knife-edge case arises when sites' differences in responsiveness to the two treatments and differences in providers' contributions are perfectly negatively correlated, so that the two sources of heterogeneity exactly offset each other, a fairly implausible scenario. The second arises when providers' contributions to the effects of the two treatments are perfectly positively correlated, also a fairly implausible scenario if in each site the treatments are delivered by different providers. Outside these cases, $\sigma_{\Delta}^2 = 0$ implies that providers do not contribute to effect heterogeneity across sites, and that sites' relative responsiveness to the two treatments is constant. Hence, all treatment-effect heterogeneity comes from $\bar{R}(S_i)$, sites' average responsiveness to the two treatments. Point 2 shows that under the testable equal-variance restriction $\text{Var}_w(\mu_{\Delta_2}(S_i)) = \text{Var}_w(\mu_{\Delta_1}(S_i))$, the null hypotheses $\sigma_{\Delta}^2 = 0$ and $\beta_{21} = 1$ are equivalent.

To estimate σ_{Δ}^2 from Proposition 1 and Theorem 1, let

$$Y_i = O_{i,1}(2) - O_{i,1}(1).$$

⁶With the convention that $0/0 = 0$, a convention we use here and in a few other places, when doing so avoids proving results separately in knife-edge cases where a ratio's numerator and denominator are both equal to zero.

The corresponding estimator of $\mu_Y(S_i)$ and variance correction are

$$\hat{\mu}_Y(S_i) = \bar{O}_{2,i,1} - \bar{O}_{1,i,1}, \quad \hat{V}_{Y,i} = \hat{V}_{2,i} + \hat{V}_{1,i}.$$

Plugging these quantities into (5) yields an estimator of σ_Δ^2 . The control-group mean cancels from the difference between the two estimated ATEs, so neither the estimator nor its variance correction uses observations from the control group.

To estimate β_{21} from Proposition 1 and Theorem 2, let

$$Y_i = O_{i,1}(2) - O_{i,1}(0), \quad X_i = O_{i,1}(1) - O_{i,1}(0).$$

The corresponding estimators of $\mu_Y(S_i)$ and $\mu_X(S_i)$ and variance–covariance corrections are

$$\begin{aligned} \hat{\mu}_Y(S_i) &= \bar{O}_{2,i,1} - \bar{O}_{0,i,1}, & \hat{\mu}_X(S_i) &= \bar{O}_{1,i,1} - \bar{O}_{0,i,1}, \\ \hat{V}_{Y,i} &= \hat{V}_{2,i} + \hat{V}_{0,i}, & \hat{V}_{X,i} &= \hat{V}_{1,i} + \hat{V}_{0,i}, \\ \hat{C}_{XY,i} &= \hat{V}_{0,i}. \end{aligned}$$

Plugging those into (8) yields an estimator of β_{21} .

Results in this section relate to work that seeks to separate implementers’ or organizations’ effectiveness from differences in the populations and contexts in which they operate. Bloom, Hill, and Riccio (2003) pool multi-site welfare-to-work experiments and relate local impacts to measured features of program implementation. Hong, Deutsch, Kress, Trinidad, and Xu (2024) instead define an organization’s relative effectiveness by comparing its performance with that of organizations facing similar ecological conditions, using control outcomes and observed participant composition to adjust for those conditions. Relatedly, Angrist and Meager (2023) explain cross-study heterogeneity in an educational intervention using take-up and whether teachers or volunteers delivered the program. Our approach is complementary. It does not require the researcher to observe and model all relevant implementation practices, or to assume that observed variables suffice to adjust for differences in local conditions. Instead, it uses a design in which two related treatments are delivered by different providers within the same sites. Comparing their site-specific effects holds the site fixed, and their covariance reveals whether the two treatments tend to succeed in the same environments. Under the decomposition in (18), the variance of the difference between the treatments’ effects can further indicate whether provider effectiveness contributes to cross-site heterogeneity. This advantage comes with a narrower scope: the comparison requires multiple treatments delivered in the same sites, and does not identify which provider practices account for differences in effectiveness.

3.5 Estimating LATEs’ heterogeneity in RCTs with imperfect compliance

3.5.1 Set-up

In this subsection, we assume that $\mathcal{D} = 1$ and $L = 2$, and allow for imperfect compliance with treatment assignment. Outcome $O_{i,1}$ is treatment take-up and outcome $O_{i,2}$ is the outcome of interest. Let $O_{i,1}(d)$ denote the treatment that a generic unit in site i would receive if assigned to arm d , and let $O_{i,2}(d, o_1)$ denote the potential value of the outcome of interest if the unit were assigned to arm d and its treatment were set to o_1 . Thus, the assignment-indexed potential outcome introduced above satisfies

$$O_{i,2}(d) = O_{i,2}(d, O_{i,1}(d)).$$

Let

$$\begin{aligned}\text{FS}_i &= O_{i,1}(1) - O_{i,1}(0), \\ \text{ITT}_i &= O_{i,2}(1) - O_{i,2}(0).\end{aligned}$$

Then, $\mu_{\text{FS}}(S_i)$ and $\mu_{\text{ITT}}(S_i)$ respectively denote site i 's first-stage effect and its intention-to-treat (ITT) effect on the outcome of interest.

We impose the usual exclusion restriction and monotonicity conditions.

Assumption 6 (Exclusion and monotonicity). *For every o_1 in the support of the received treatment,*

$$O_{i,2}(1, o_1) = O_{i,2}(0, o_1) \quad \text{almost surely.}$$

Moreover, for $d \in \{0, 1\}$ $O_{i,1}(d) \in \{0, 1\}$ almost surely,⁷ $O_{i,1}(1) \geq O_{i,1}(0)$ almost surely, and $\mu_{\text{FS}}(S_i) > 0$ almost surely.

The exclusion restriction states that assignment affects outcomes only through the treatment received, while monotonicity rules out defiers.

Define the LATE of site i on the outcome of interest as

$$\mu_{\text{LATE}}(S_i) = \frac{\mu_{\text{ITT}}(S_i)}{\mu_{\text{FS}}(S_i)}.$$

Under Assumption 6, $\mu_{\text{LATE}}(S_i)$ is the average effect of the treatment on the outcome among compliers in site i .

Let

$$\text{FS} = \mathbb{E}_w[\mu_{\text{FS}}(S)], \quad \text{ITT} = \mathbb{E}_w[\mu_{\text{ITT}}(S)]$$

denote the weighted average first-stage and ITT effects across all sites. The overall LATE is

$$\text{LATE} = \frac{\text{ITT}}{\text{FS}} = \mathbb{E}_w \left[\frac{\mu_{\text{FS}}(S)}{\text{FS}} \mu_{\text{LATE}}(S) \right]. \quad (19)$$

Thus, the overall LATE averages site-specific LATEs with weights proportional to $w(S)$ times sites' first stages.

For estimation, let $\hat{\mu}_{\text{FS}}(S_i)$ and $\hat{\mu}_{\text{ITT}}(S_i)$ be the usual difference-in-means estimators. Unbiased estimators of their variances and covariances can be obtained following Proposition 1. Apply (7) to define $\hat{C}_{\text{FS}, \text{ITT}}$, the corrected estimator of $\text{Cov}_w(\mu_{\text{FS}}(S), \mu_{\text{ITT}}(S))$. Also let

$$\widehat{\text{FS}} = \frac{1}{n} \sum_{i=1}^n \hat{w}(S_i) \hat{\mu}_{\text{FS}}(S_i), \quad \widehat{\text{ITT}} = \frac{1}{n} \sum_{i=1}^n \hat{w}(S_i) \hat{\mu}_{\text{ITT}}(S_i), \quad \widehat{\text{LATE}} = \frac{\widehat{\text{ITT}}}{\widehat{\text{FS}}}.$$

⁷This condition ensures that results apply to site-specific LATEs. Without it, (22) below instead applies to site-specific average causal responses (Angrist and Imbens, 1995). More generally, results in the spirit of (22) apply to site-specific parameters that can be written as ratios with positive denominators, provided the numerators and denominators admit unbiased estimators, whose variances and covariance can also be unbiasedly estimated. This may accommodate suitably defined estimands with a scalar nonbinary instrument, though researchers need to verify that their numerator and denominator estimators satisfy the stated requirements.

3.5.2 Estimating a variance of LATEs across sites

We consider two target parameters. The first is the variance of site-specific LATEs under the complier weights defining the overall LATE in (19):

$$\sigma_{\text{LATE}}^2 = \mathbb{E}_w \left[\frac{\mu_{\text{FS}}(S)}{\text{FS}} \{ \mu_{\text{LATE}}(S) - \text{LATE} \}^2 \right]. \quad (20)$$

The second uses weights proportional to $w(S)$ times sites' squared first stages:

$$\sigma_{\text{LATE,FS}^2}^2 = \mathbb{E}_w \left[\frac{\mu_{\text{FS}}(S)^2}{\mathbb{E}_w[\mu_{\text{FS}}(S)^2]} \{ \mu_{\text{LATE}}(S) - \text{LATE}_{\text{FS}^2} \}^2 \right], \quad (21)$$

where

$$\text{LATE}_{\text{FS}^2} = \frac{\mathbb{E}_w[\mu_{\text{FS}}(S)^2 \mu_{\text{LATE}}(S)]}{\mathbb{E}_w[\mu_{\text{FS}}(S)^2]}.$$

The squared-first-stage weights give greater importance to sites with stronger first stages. Such weights also arise in saturated 2SLS regressions.⁸

If first stages are constant across sites, both variances equal $\text{Var}_w(\mu_{\text{ITT}}(S))/\text{FS}^2$, which can be consistently estimated using the corrected ITT variance and the average first stage. In the remainder of this section, we assume that first stages are heterogeneous. With heterogeneous first stages, corrected estimators cannot be applied directly to site-specific LATE estimators: $\hat{\mu}_{\text{ITT}}(S_i)/\hat{\mu}_{\text{FS}}(S_i)$ is a ratio and is generally not unbiased, and its estimated variance can be very large when the estimated first stage is small. In an application to Head Start, Walters (2015) finds that this approach yields a negative corrected variance estimate.

Instead, we propose estimators that use only moments of ITTs and first stages.

Proposition 5 (Identifying $\sigma_{\text{LATE,FS}^2}^2$ and a lower bound for σ_{LATE}^2). *Suppose Assumption 6 holds and the moments below exist. Then*

$$\sigma_{\text{LATE,FS}^2}^2 = \frac{\mathbb{E}_w[\{ \mu_{\text{ITT}}(S) - \text{LATE}_{\text{FS}^2} \mu_{\text{FS}}(S) \}^2]}{\mathbb{E}_w[\mu_{\text{FS}}(S)^2]}. \quad (22)$$

If $\mathbb{E}_w[\mu_{\text{FS}}(S)^2] < \text{FS}$,

$$\begin{aligned} \sigma_{\text{LATE}}^2 &\geq \underline{\sigma}_{\text{LATE}}^2 := \frac{\mathbb{E}_w[\{ \mu_{\text{ITT}}(S) - \text{LATE} \mu_{\text{FS}}(S) \}^2]}{\text{FS}} \\ &\quad + \frac{\{ \mathbb{E}_w[\mu_{\text{FS}}(S) \mu_{\text{ITT}}(S)] - \text{LATE} \times \mathbb{E}_w[\mu_{\text{FS}}(S)^2] \}^2}{\text{FS} \{ \text{FS} - \mathbb{E}_w[\mu_{\text{FS}}(S)^2] \}} \\ &= \frac{\mathbb{E}_w[\mu_{\text{FS}}(S)^2]}{\text{FS}} \left\{ \sigma_{\text{LATE,FS}^2}^2 + \frac{\text{FS}}{\text{FS} - \mathbb{E}_w[\mu_{\text{FS}}(S)^2]} (\text{LATE}_{\text{FS}^2} - \text{LATE})^2 \right\}. \end{aligned} \quad (23)$$

Because first stages are strictly positive, $\sigma_{\text{LATE,FS}^2}^2$ is zero if and only if LATEs are constant across sites. Whenever the lower bound is defined, $\underline{\sigma}_{\text{LATE}}^2$ has the same property. (22) and (23) can therefore be used to test if differences in take-up alone explain ITT heterogeneity.

⁸Theorem 3 of Angrist and Imbens (1995) shows that such regressions weight covariate-specific LATEs by the conditional variance of the first-stage fitted value. With a binary instrument, this variance equals the squared conditional first stage times the conditional variance of the instrument. With equally probable treatment assignment in every covariate cell, the latter factor is constant.

(22) follows from the fact that

$$\mu_{\text{FS}}(S)^2 \{\mu_{\text{LATE}}(S) - \text{LATE}_{\text{FS}^2}\}^2 = \{\mu_{\text{ITT}}(S) - \text{LATE}_{\text{FS}^2} \mu_{\text{FS}}(S)\}^2.$$

For (23), binary take-up and monotonicity imply $0 < \mu_{\text{FS}}(S) \leq 1$. Since $\mathbb{E}_w[\mu_{\text{ITT}}(S) - \text{LATE} \mu_{\text{FS}}(S)] = 0$, the fact that

$$1 - \mu_{\text{FS}}(S) = \sqrt{(1 - \mu_{\text{FS}}(S))\mu_{\text{FS}}(S)} \sqrt{(1 - \mu_{\text{FS}}(S))/\mu_{\text{FS}}(S)}$$

and Cauchy-Schwarz imply

$$\begin{aligned} & \{\mathbb{E}_w[\mu_{\text{FS}}(S)\{\mu_{\text{ITT}}(S) - \text{LATE} \mu_{\text{FS}}(S)\}]\}^2 \\ &= \{\mathbb{E}_w[\{1 - \mu_{\text{FS}}(S)\}\{\mu_{\text{ITT}}(S) - \text{LATE} \mu_{\text{FS}}(S)\}]\}^2 \\ &\leq \mathbb{E}_w[\mu_{\text{FS}}(S)\{1 - \mu_{\text{FS}}(S)\}] \mathbb{E}_w\left[\frac{1 - \mu_{\text{FS}}(S)}{\mu_{\text{FS}}(S)} \{\mu_{\text{ITT}}(S) - \text{LATE} \mu_{\text{FS}}(S)\}^2\right] \\ &= \{\text{FS} - \mathbb{E}_w[\mu_{\text{FS}}(S)^2]\} \{\text{FS} \sigma_{\text{LATE}}^2 - \mathbb{E}_w[\{\mu_{\text{ITT}}(S) - \text{LATE} \mu_{\text{FS}}(S)\}^2]\}. \end{aligned}$$

Rearranging yields (23). The equality below (23) follows by writing

$$\mu_{\text{LATE}}(S) - \text{LATE} = \mu_{\text{LATE}}(S) - \text{LATE}_{\text{FS}^2} + \text{LATE}_{\text{FS}^2} - \text{LATE},$$

expanding the square, and using $\mathbb{E}_w[\mu_{\text{FS}}(S)^2 \{\mu_{\text{LATE}}(S) - \text{LATE}_{\text{FS}^2}\}] = 0$.

When first stages vary across sites, we conjecture that given the first and second moments of ITTs and first stages, without outcome-support restriction σ_{LATE}^2 is a sharp lower bound for σ_{LATE}^2 . We leave the study of this conjecture for future work. If $\mathbb{E}_w[\mu_{\text{FS}}(S)^2] = \text{FS}$, strict positivity of first stages implies full compliance at every site; the complier-weighted variance then equals the ITT variance and is identified directly.

Turning to estimation, let

$$m_2 = \mathbb{E}_w[\mu_{\text{FS}}(S)^2], \quad r = \mathbb{E}_w[\mu_{\text{FS}}(S)\mu_{\text{ITT}}(S)].$$

Then

$$\text{LATE}_{\text{FS}^2} = \frac{r}{m_2} = \frac{\text{Cov}_w(\mu_{\text{FS}}(S), \mu_{\text{ITT}}(S)) + \text{FS ITT}}{\text{Var}_w(\mu_{\text{FS}}(S)) + \text{FS}^2}. \quad (24)$$

For a scalar l , define

$$\begin{aligned} \hat{\nu}_i(l) &= \hat{\mu}_{\text{ITT}}(S_i) - l\hat{\mu}_{\text{FS}}(S_i), \\ \hat{V}_{\nu,i}(l) &= \hat{V}_{\text{ITT},i} + l^2\hat{V}_{\text{FS},i} - 2l\hat{C}_{\text{FS,ITT},i}, \\ g_i(l) &= \hat{\nu}_i(l)^2 - \hat{V}_{\nu,i}(l), \quad \hat{Q}(l) = \frac{1}{n} \sum_{i=1}^n \hat{w}(S_i)g_i(l). \end{aligned}$$

The first two quantities are the estimated site effect on $O_{i,2} - lO_{i,1}$ and its estimated sampling variance. Define

$$\begin{aligned} \hat{r} &= \frac{1}{n} \sum_{i=1}^n \hat{w}(S_i) \{\hat{\mu}_{\text{FS}}(S_i)\hat{\mu}_{\text{ITT}}(S_i) - \hat{C}_{\text{FS,ITT},i}\}, \\ \hat{m}_2 &= \frac{1}{n} \sum_{i=1}^n \hat{w}(S_i) \{\hat{\mu}_{\text{FS}}(S_i)^2 - \hat{V}_{\text{FS},i}\}, \\ \widehat{\text{LATE}}_{\text{FS}^2} &= \hat{r}/\hat{m}_2, \quad \hat{q} = \hat{Q}(\widehat{\text{LATE}}_{\text{FS}^2}). \end{aligned} \quad (25)$$

Our estimators are

$$\hat{\sigma}_{\text{LATE,FS}^2}^2 = \frac{\hat{q}}{\hat{m}_2} \quad (26)$$

and

$$\begin{aligned} \hat{\sigma}_{\text{LATE}}^2 &= \frac{\widehat{Q}(\widehat{\text{LATE}})}{\widehat{\text{FS}}} + \frac{(\hat{r} - \hat{m}_2 \widehat{\text{LATE}})^2}{\widehat{\text{FS}}(\widehat{\text{FS}} - \hat{m}_2)} \\ &= \frac{\hat{m}_2}{\widehat{\text{FS}}} \left\{ \hat{\sigma}_{\text{LATE,FS}^2}^2 + \frac{\widehat{\text{FS}}}{\widehat{\text{FS}} - \hat{m}_2} (\widehat{\text{LATE}}_{\text{FS}^2} - \widehat{\text{LATE}})^2 \right\}. \end{aligned} \quad (27)$$

These expressions are used whenever their denominators are nonzero. Theorem 4 in Appendix A.3 establishes asymptotic normality and gives the influence functions and standard-error estimators for both quantities. Estimating $\text{LATE}_{\text{FS}^2}$ has no first-order effect on $\hat{\sigma}_{\text{LATE,FS}^2}^2$, whereas estimating LATE generally contributes to the influence function of $\hat{\sigma}_{\text{LATE}}^2$.

Comparison with existing approaches to LATE heterogeneity. To estimate the variance of LATEs across sites, Walters (2015) estimates a parametric selection model with jointly normal center-specific potential-outcome and compliance parameters. Adusumilli, Agostinelli, and Borghesan (2024) use a grouped-random-effect model allowing compliance and treatment effects to be jointly heterogeneous. Our approach identifies a squared-first-stage-weighted variance and a lower bound for the complier-weighted variance without parametric restrictions on their joint distribution. Angrist and Meager (2023) propose frequentist and Bayesian estimators of LATE heterogeneity with few sites and many units per site, a different setting from the many-sites, few-units-per-site setting considered here. Similarly, Walters (2015) uses an overidentification test of LATE homogeneity, which tests whether sites' ITTs and first stages lie on a common line through the origin, up to sampling error. The test's conventional asymptotic justification applies with few sites and many units per site, but it may not be valid in the many-sites, few-units-per-site setting considered here.

3.5.3 Estimating the covariance between sites' LATEs and their first-stage effects

In this subsection, our target parameter is

$$\sigma_{\text{LATE,FS}} = \mathbb{E}_w \left[\frac{\mu_{\text{FS}}(S)}{\text{FS}} \{ \mu_{\text{LATE}}(S) - \text{LATE} \} \mu_{\text{FS}}(S) \right]. \quad (28)$$

$\sigma_{\text{LATE,FS}}$ is the covariance between site-specific LATEs and first stages under the same complier weights as those defining the overall LATE in (19). It reveals whether sites with larger shares of compliers also tend to have compliers who benefit more from treatment. A positive covariance is consistent with Roy selection across sites: sites where compliers benefit more tend to have larger shares of compliers.

Let

$$\beta_{\text{ITT,FS}} = \frac{\text{Cov}_w(\mu_{\text{FS}}(S), \mu_{\text{ITT}}(S))}{\text{Var}_w(\mu_{\text{FS}}(S))}.$$

Proposition 6. *Suppose Assumption 6 holds, the moments below exist, and $\text{Var}_w(\mu_{\text{FS}}(S)) > 0$. Then,*

$$\sigma_{\text{LATE,FS}} = \frac{\text{Var}_w(\mu_{\text{FS}}(S))}{\text{FS}} (\beta_{\text{ITT,FS}} - \text{LATE}). \quad (29)$$

Thus, $\sigma_{\text{LATE,FS}}$ is positive if and only if the slope from the regression of sites' ITTs on an intercept and their first stages exceeds the overall LATE. Equivalently, one can show that

$$\sigma_{\text{LATE,FS}} = \frac{\mathbb{E}_w[\mu_{\text{FS}}(S)^2]}{\text{FS}} (\text{LATE}_{\text{FS}^2} - \text{LATE}).$$

Thus, the covariance is proportional to the change in the mean LATE when moving from first-stage to squared-first-stage weights.

We estimate (28) by

$$\hat{\sigma}_{\text{LATE,FS}} = \frac{\widehat{C}_{\text{FS,ITT}} - \widehat{\text{LATE}}_{\text{FS}} \widehat{V}_{\text{FS}}}{\widehat{\text{FS}}}, \quad (30)$$

where $\widehat{V}_{\text{FS}}$ denotes the corrected variance estimator in (6) with $X = \text{FS}$. Appendix A.4 gives a heuristic derivation of the influence function of $\hat{\sigma}_{\text{LATE,FS}}$.

3.5.4 Relationships between LATEs across outcomes and treatment arms

In this section we study two targets: a weighted regression of one outcome's LATE on other outcomes' LATEs within a treatment arm, and a covariance–variance ratio relating LATEs across treatment arms. The latter uses different weights in its numerator and denominator and therefore is generally not a regression coefficient.

To define these targets, we consider a more general setting, namely an RCT with two treatment arms and $J + 2$ outcomes. Outcomes $O_{i,1}, \dots, O_{i,J}$ are the $J \geq 2$ outcomes of interest, while $O_{i,J+1}$ and $O_{i,J+2}$ indicate take-up of treatments 1 and 2, respectively. For $d \in \{1, 2\}$ and $k \in \{1, \dots, J\}$, let

$$\mu_{\text{FS}_d}(S) = \mathbb{E}[O_{i,J+d}(d) - O_{i,J+d}(0) \mid S_i = S], \quad \mu_{\text{ITT}_{d,k}}(S) = \mathbb{E}[O_{i,k}(d) - O_{i,k}(0) \mid S_i = S],$$

respectively denote the site-specific first-stages of treatment d , and its ITT effects on outcome k . Then, let

$$\mu_{\text{LATE}_{d,k}}(S) = \frac{\mu_{\text{ITT}_{d,k}}(S)}{\mu_{\text{FS}_d}(S)}$$

denote the site-specific LATEs of treatment d on outcome k . We assume the analogues of Assumption 6 for both treatment arms, and we assume that assignment to arm d does not affect take-up of treatment $d' \neq d$.⁹ For positive random variables $A(S)$, write

$$\mathbb{E}_A[Z(S)] = \frac{\mathbb{E}_w[A(S)Z(S)]}{\mathbb{E}_w[A(S)]},$$

and use Var_A and Cov_A for the corresponding variance and covariance.

Our first target predicts the LATE of arm d on outcome 1 using its LATEs on outcomes 2 through J . Let

$$\boldsymbol{\mu}_{\text{LATE},d,-1}(S) = (\mu_{\text{LATE}_{d,2}}(S), \dots, \mu_{\text{LATE}_{d,J}}(S))^{\top}.$$

For $d \in \{1, 2\}$, define

$$\beta^d = \text{Var}_{\text{FS}_d^2}(\boldsymbol{\mu}_{\text{LATE},d,-1}(S))^{-1} \text{Cov}_{\text{FS}_d^2}(\boldsymbol{\mu}_{\text{LATE},d,-1}(S), \mu_{\text{LATE}_{d,1}}(S)), \quad (31)$$

⁹For instance, this holds automatically with one-sided imperfect compliance, where one can take up treatment d only if assigned to it.

whenever the variance matrix is nonsingular. This is the coefficient on $\mu_{\text{LATE},d,-1}(S)$ from the regression of $\mu_{\text{LATE},d,1}(S)$ on an intercept and $\mu_{\text{LATE},d,-1}(S)$, under weights proportional to arm d 's squared first stage. β^d can for instance be useful to estimate predictive mediation regressions as in Section 3.3.1, but on LATEs instead of ITT effects.

Our second target, for each outcome k , is the covariance–variance ratio

$$\beta_k^{12} = \frac{\text{Cov}_{\text{FS}_1\text{FS}_2}(\mu_{\text{LATE}_{1,k}}(S), \mu_{\text{LATE}_{2,k}}(S))}{\text{Var}_{\text{FS}_1^2}(\mu_{\text{LATE}_{1,k}}(S))}. \quad (32)$$

β_k^{12} resembles the coefficient from a regression of arm 2's outcome k LATE on arm 1's outcome k LATE. It is not literally a weighted least-squares coefficient, however, because its numerator and denominator use different weights. To make the comparison precise, when the corresponding weighted regression is well defined, let $\tilde{\beta}_k^{12}$ denote the coefficient on $\mu_{\text{LATE}_{1,k}}(S)$ from the regression of $\mu_{\text{LATE}_{2,k}}(S)$ on an intercept and $\mu_{\text{LATE}_{1,k}}(S)$, with weights proportional to $w(S)\mu_{\text{FS}_1}(S)\mu_{\text{FS}_2}(S)$:

$$\tilde{\beta}_k^{12} = \frac{\text{Cov}_{\text{FS}_1\text{FS}_2}(\mu_{\text{LATE}_{1,k}}(S), \mu_{\text{LATE}_{2,k}}(S))}{\text{Var}_{\text{FS}_1\text{FS}_2}(\mu_{\text{LATE}_{1,k}}(S))} = \frac{\text{Var}_{\text{FS}_1^2}(\mu_{\text{LATE}_{1,k}}(S))}{\text{Var}_{\text{FS}_1\text{FS}_2}(\mu_{\text{LATE}_{1,k}}(S))} \beta_k^{12}.$$

$\tilde{\beta}_k^{12}$ and β_k^{12} have the same sign. Moreover, $\tilde{\beta}_k^{12} = \beta_k^{12}$ when the two arms have equal first stages within every site, or, more generally, when $\mu_{\text{FS}_2}(S) = c\mu_{\text{FS}_1}(S)$ almost surely for a constant $c > 0$. We focus on β_k^{12} , because it can be expressed using only second moments of site-specific ITTs and first stages, whereas $\tilde{\beta}_k^{12}$ generally cannot. Finally, note that the two arms may have different complier populations within each site. Consequently, the cross-arm association may reflect both differences in treatment effectiveness and differences in complier composition.

Proposition 7 (Identification of cross-outcome and cross-treatment LATE relationships). *Suppose the analogues of Assumption 6 hold for each treatment arm $d \in \{1, 2\}$ and each outcome $k \in \{1, \dots, J\}$, with binary take-up and $\mu_{\text{FS}_d}(S) > 0$ almost surely. Suppose also that assignment to one treatment arm does not change take-up of the other treatment:*

$$O_{i,J+d'}(d) = O_{i,J+d'}(0) \quad \text{almost surely for } d, d' \in \{1, 2\}, \ d \neq d'.$$

Assume $\mathbb{E}_w[\mu_{\text{ITT}_{d,k}}(S)^2] < \infty$ for every $d \in \{1, 2\}$ and $k \in \{1, \dots, J\}$.

1. For each $d \in \{1, 2\}$ such that $\text{Var}_{\text{FS}_d^2}(\mu_{\text{LATE},d,-1}(S))$ is nonsingular, and for $j \in \{1, \dots, J\}$, let

$$\text{LATE}_{d,j,\text{FS}_d^2} = \mathbb{E}_{\text{FS}_d^2}[\mu_{\text{LATE}_{d,j}}(S)].$$

Then β^d is the coefficient from the w -weighted regression without an intercept of

$$\mu_{\text{ITT}_{d,1}}(S) - \mu_{\text{FS}_d}(S)\text{LATE}_{d,1,\text{FS}_d^2}$$

on

$$\left(\mu_{\text{ITT}_{d,2}}(S) - \mu_{\text{FS}_d}(S)\text{LATE}_{d,2,\text{FS}_d^2}, \dots, \mu_{\text{ITT}_{d,J}}(S) - \mu_{\text{FS}_d}(S)\text{LATE}_{d,J,\text{FS}_d^2} \right)^\top.$$

2. For each $k \in \{1, \dots, J\}$ such that $\text{Var}_{\text{FS}_1^2}(\mu_{\text{LATE}_{1,k}}(S)) > 0$,

$$\beta_k^{12} = \frac{\frac{\mathbb{E}_w[\mu_{\text{ITT}_{1,k}}(S)\mu_{\text{ITT}_{2,k}}(S)]}{\mathbb{E}_w[\mu_{\text{FS}_1}(S)\mu_{\text{FS}_2}(S)]} - \frac{\mathbb{E}_w[\mu_{\text{FS}_2}(S)\mu_{\text{ITT}_{1,k}}(S)]\mathbb{E}_w[\mu_{\text{FS}_1}(S)\mu_{\text{ITT}_{2,k}}(S)]}{\mathbb{E}_w[\mu_{\text{FS}_1}(S)\mu_{\text{FS}_2}(S)]^2}}{\frac{\mathbb{E}_w[\mu_{\text{ITT}_{1,k}}(S)^2]}{\mathbb{E}_w[\mu_{\text{FS}_1}(S)^2]} - \frac{\mathbb{E}_w[\mu_{\text{FS}_1}(S)\mu_{\text{ITT}_{1,k}}(S)]^2}{\mathbb{E}_w[\mu_{\text{FS}_1}(S)^2]^2}}. \quad (33)$$

Thus, both targets depend only on second moments of site-specific ITTs and first stages. Under the sampling and moment conditions ensuring consistency of the corrected second-moment estimators, both targets can therefore be consistently estimated. Appendix A.5 gives estimators of β^d and β_k^{12} and heuristic derivations of their influence functions.

3.6 Extensions and avenues for future research

3.6.1 Estimating effect heterogeneity in other RCTs

Replacing sites by strata, the estimators above can also be used to estimate and predict treatment-effect heterogeneity across strata in any stratified RCT, irrespective of whether the strata correspond to geographic sites, provided every treatment arm contains at least two units in each stratum (resp. three units if one wants to estimate centered third moments), under analogues of our sampling assumptions. The estimators above are therefore not applicable to paired RCTs. If one is interested in estimating effect heterogeneity across strata, this may be an argument to use larger strata, such as strata of four in a two-arm experiment.

We conjecture that our results are also applicable to nonstratified RCTs. Suppose that instead of separate randomizations within sites, a single completely randomized experiment is conducted, with fixed overall treatment-arm sizes. Let $C = ((N_{d,i})_{d=0}^{\mathcal{D}})_{i=1}^n$ denote the realized site-by-arm count table. Conditional on C , the assignment vectors are independent across sites and, within each site i , assignment is uniform over all allocations having arm sizes $(N_{d,i})_{d=0}^{\mathcal{D}}$ (Miratrix, Sekhon, and Yu, 2013).¹⁰ Thus, conditional on C , the nonstratified experiment is equivalent to independent stratified experiments conducted within sites. This suggests modifying Assumption 2 by excluding the arm sizes from the vectors drawn independently across sites, because the $N_{d,i}$ are now generated by the single randomization and are dependent across sites, and replacing Assumption 3 by its conditional-on- C counterpart. For sites such that $N_{d,i} \geq 2$ for every d , the site-specific estimators remain unchanged after inserting the realized arm sizes, and their conditional unbiasedness and variance-correction properties continue to hold. Assuming $N_i \geq 2(\mathcal{D}+1)$ almost surely, one possible approach to retaining the original population target when aggregating across admissible sites is to replace $W(S_i)$ by

$$\frac{A_i W(S_i)}{p_i}, \quad A_i := \mathbf{1}\{N_{d,i} \geq 2 \text{ for every } d\},$$

where p_i is the probability that $A_i = 1$, conditional on the realized site sizes and the overall treatment-arm totals. This approach requires positive inclusion probabilities and suitable moment conditions on the inverse-probability-weighted contributions. Establishing consistency and asymptotic normality would require accounting for both within-site estimation error and the randomness of the count table C . In particular, a conditional central limit theorem would need to be complemented by an analysis of fluctuations in the estimator's conditional mean around the population target. Developing these results is an interesting extension for future research.

3.6.2 Application to other research designs

Moment-correction arguments may also apply to other research designs that provide unbiased site-specific effect estimators and unbiased estimators of their variances. For instance, consider a

¹⁰Their formal result considers binary treatment; the extension to multiple arms follows from the same multinomial counting argument.

two-period panel satisfying no anticipation and within-site parallel trends. If treated and control units are independent i.i.d. samples within each site, the difference between their average outcome changes is unbiased for the site’s superpopulation average treatment effect on the treated. With at least two observations in each arm, the estimator’s variance can be unbiasedly estimated. This provides a route to applying our methods to difference in difference estimation. Teachers’ value added offers another potential application, which would require specifying conditions under which teacher-specific estimators and their variance and covariance can be unbiasedly estimated.

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

4.1 Study design and data

Behaghel, Crépon, and Gurgand (2014) conduct an 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. The control group receives the standard services provided by the PES. The other two groups are assigned to intensive counseling programs delivered either by the PES or by a private provider. In this application, PES offices are the sites and job seekers are the randomization units. When analyzing the public program, we exclude 16 PES offices where less than two units were assigned to that arm or to the control group. For the private program, the corresponding sample restrictions lead us to exclude 12 PES offices.

Compliance with assignment is imperfect. Almost no job seeker assigned to the control group receives intensive counseling, but only 32% of those assigned to public counseling and 43% of those assigned to private counseling enroll. Our main outcome is an indicator for holding a job six months after randomization, one of the three main employment outcomes considered by Behaghel, Crépon, and Gurgand (2014). Results are similar for their other employment outcomes. We weight sites using the weights in the original study.

The study has both limitations and strengths for our purposes. The data do not contain potential mediators such as measures of job-search effort, precluding the mediation analyses discussed above. Moreover, because randomization takes place within local labor markets, the programs may generate displacement effects. The estimated effects are therefore partial-equilibrium effects. On the other hand, in every site, two similar programs are delivered by different providers, thus allowing us to assess if effect heterogeneity is driven by providers’ effectiveness or by sites’ responsiveness.

4.2 Estimating the variance of ITTs across sites

Table 1 shows that the estimated average effects on employment after six months are 2.4 percentage points (pp) for public counseling and 1.9 pp for private counseling. Those averages conceal substantial heterogeneity. The estimated variances are positive and statistically significant for both programs. The estimated standard deviations of site-specific effects are 9.2 pp for public counseling and 8.5 pp for private counseling, about four times their respective average effects.

The estimated centered third moments are positive but not statistically significant. Thus, there is no evidence that the distribution of site-specific effects is asymmetric. Dividing each estimator by

Table 1: Variance and centered third moment across sites of counseling effects on employment

	$\widehat{\text{ITT}}$	$\widehat{\sigma}_{\text{ITT}}^2$	$\sqrt{\widehat{\sigma}_{\text{ITT}}^2/\widehat{\text{ITT}}}$	$\widehat{\kappa}_{\text{ITT}}$	Units
	(1)	(2)	(3)	(4)	(5)
Public counseling	0.024 (0.011)	0.0084 (0.0037)	3.809	0.00063 (0.00137)	7,198
Private counseling	0.019 (0.008)	0.0073 (0.0022)	4.478	0.00008 (0.00069)	34,768

Notes: The outcome is an indicator for holding a job six months after randomization. Column (1) reports the weighted average ITT effect. Column (2) reports the corrected variance estimator in (5). Standard errors, based on the influence functions in Appendix A, appear in parentheses. Column (3) divides the estimated standard deviation of site-specific effects by the corresponding average effect. Column (4) reports the corrected centered third-moment estimator in (9), using sites with at least three treatment and three control observations: 194 sites and 7,154 units for public counseling, and 201 sites and 34,709 units for private counseling. Columns (1)–(3) and (5) use sites with at least two treatment and two control observations. Results use data from Behaghel, Crépon, and Gurgand (2014).

the estimated variance on its estimation sample raised to the power $3/2$ yields implied skewnesses of 0.87 for public counseling and 0.13 for private counseling. Because the outcome is binary, site-specific ITTs are bounded between -1 and 1 , ruling out the unbounded tails of the heavy-tail, high-skewness simulation design in which confidence intervals based on analytic standard errors undercovered (Appendix B).

4.3 Regressing sites' ITTs on their average untreated outcome

Column (1) of Table 2 reports measurement-error-corrected regressions of site-specific effects on sites' average untreated outcome. The coefficients are negative for both programs and especially large and precise for private counseling. For that program, the untreated outcome explains almost one half of the variance in site-specific effects. Intensive counseling thus tends to be more effective in sites where job seekers would otherwise have lower employment rates. Column (2) reports regressions of site-specific effects on the local unemployment rate, which we are able to recover for all but one site. We find no statistically significant relationship between local unemployment and either program's effect. This does not contradict the results in Column (1). The local unemployment rate is an imperfect proxy for the labor-market conditions facing the RCT's population, which consists of job seekers at high risk of long-term unemployment. Regressing the control-group employment rate on local unemployment produces significantly negative coefficients, but measurement-error corrected R^2 s are equal to only 0.134 and 0.100 for the public- and private-program samples, respectively. Finally, the naive regressions in Table 2 illustrate the importance of correcting for estimation error when the predictor is estimated. In Column (1), the naive coefficients differ from the corrected coefficients, and their conventional standard errors are much smaller. On the other hand, when the predictor is measured without error, as is the case in Column (2), both approaches yield the same coefficient and very similar standard errors.

The association between sites' ITT effects and untreated outcomes may just be due to heterogeneous job seeker characteristic across sites, which correlate with both the untreated outcome and the program's effect. To investigate this possibility, we follow Behaghel, Crépon, and Gurgand (2014) and estimate, in the control group, a job-seeker-level logistic regression of employment on 43 characteristics measuring education, previous employment and unemployment histories, and reser-

Table 2: Predictors of site-specific counseling effects

	Untreated outcome (1)	Unemployment rate (2)
<i>Panel A: Public counseling</i>		
$\hat{\beta}$	−0.480 (0.307)	−0.034 (0.265)
$\hat{R}^2$	0.124	0.0002
Naive coefficient	−0.850 (0.081)	−0.034 (0.266)
Sites	200	199
<i>Panel B: Private counseling</i>		
$\hat{\beta}$	−0.804 (0.126)	0.095 (0.242)
$\hat{R}^2$	0.496	0.002
Naive coefficient	−0.939 (0.036)	0.095 (0.242)
Sites	204	203

Notes: Each column reports a separate weighted regression of site-specific ITT effects on the indicated site characteristic. The untreated outcome is the site’s control-group employment rate after six months. The corrected coefficient and R^2 use (8) and (45). Standard errors, based on the influence functions in Appendix A, appear in parentheses. Naive regressions do not correct for measurement error. The corrected and naive coefficients coincide in Column (2) because the unemployment rate is measured without error. Results use data from Behaghel, Crépon, and Gurgand (2014).

vation wages. We average the fitted values within each site to obtain its predicted employment rate without treatment. Finally we decompose the site’s actual employment rate without treatment into this prediction and a residual. The predicted employment rate without treatment does not predict treatment effects. By contrast, the residual is strongly negatively related to effects, as shown in Table 3. This suggests that the relationship between untreated outcomes and treatment effects is not driven by heterogeneous job seeker characteristic across sites.

The negative association between sites’ ITT effects and untreated outcomes may be useful for targeting. Although a site’s untreated outcome is unavailable before treatment, the employment rate of an earlier cohort of similarly eligible job seekers could serve as a proxy.

4.4 Sites’ responsiveness or providers’ effectiveness?

If heterogeneity were driven entirely by differences in provider effectiveness, the effects of the two programs need not be related because public and private counseling are delivered by different providers within a site. Conversely, a positive association would indicate that the programs tend to work well in the same places and would therefore point toward a role for sites’ responsiveness.

Following Proposition 4, we first estimate σ_{Δ}^2 , the variance across sites of the difference between the private and public programs’ effects, using the 200 sites in which all three arms have at least two units. The estimate is 0.0072, with a standard error of 0.0025, so we reject the null that $\sigma_{\Delta}^2 = 0$. Under the conditions of Point 1.b) of Proposition 4, this implies that providers’ effectiveness, and / or sites’ relative responsiveness to the two programs, contribute to effect heterogeneity. Were the two programs’ effects uncorrelated across sites, σ_{Δ}^2 would be approximately equal to the sum of

Table 3: Job-seeker composition and site-specific counseling effects

	Predicted untreated outcome (1)	Residual untreated outcome (2)
<i>Panel A: Public counseling</i>		
$\hat{\beta}$	0.032 (0.304)	-0.679 (0.356)
$\hat{R}^2$		0.198
Sites	200	200
<i>Panel B: Private counseling</i>		
$\hat{\beta}$	0.029 (0.125)	-0.978 (0.113)
$\hat{R}^2$		0.681
Sites	204	204

Notes: Each panel reports one multivariate regression of site-specific ITT effects on the site’s predicted untreated employment rate and the residual component of its control-group employment rate. Predictions are obtained from a job-seeker-level logistic regression using 43 baseline characteristics. Coefficients correct for measurement error in the site-level variables. Standard errors, based on the influence functions in Appendix A, appear in parentheses. Results use data from Behaghel, Crépon, and Gurgand (2014).

their variances in Table 1, about 0.016. The estimate is less than half that, suggesting there is a positive covariance between the two programs’ site-specific effects. Indeed, regressing the public program’s site-specific effect on the private program’s effect, correcting for estimation error in both variables, yields a coefficient of 0.563 (standard error 0.215) and an R^2 of 0.277.

Overall, the difference between the programs’ effects varies significantly across sites. This variation can reflect differences in provider effectiveness, differences in sites’ relative responsiveness to the two programs, or both. The positive association between the programs’ effects is consistent with a role for shared site responsiveness, although correlated provider effectiveness could also generate this association. These results do not separately identify the contributions of providers and sites.

4.5 Estimating and predicting LATEs’ heterogeneity

We first document substantial heterogeneity in first-stage effects across sites. Table 4 shows that assignment to public and private counseling increases enrollment in the corresponding program by 31.2 and 40.2 pp on average, respectively. The estimated standard deviations of first-stage effects across sites are approximately 10 and 11.8 pp, respectively, amounting to 32% and 29% of the corresponding average first stages.

We next assess whether sites with more compliers also have compliers with systematically different gains. As shown in (29), the sign of the covariance between sites’ first stage and LATE effects is the sign of $\beta_{\text{ITT,FS}} - \text{LATE}$. Table 5 reports estimates of this difference. They are negative and insignificant for both programs, so we cannot reject the null that site-specific LATEs and first stages are uncorrelated.

Table 6 shows that the estimates of the average LATEs are equal to 7.7 pp for public counseling and 4.8 pp for private counseling. The lower bounds for the complier-weighted variances of LATEs are both positive and statistically significant. They imply estimated standard deviations of site-specific LATEs of at least 16.4 and 13.4 pp for public and private counseling, respectively, about

Table 4: Variance of first-stage effects across sites

	$\widehat{\text{FS}}$	$\widehat{\sigma}_{\text{FS}}^2$	$\sqrt{\widehat{\sigma}_{\text{FS}}^2/\widehat{\text{FS}}}$	Units
	(1)	(2)	(3)	(4)
Public counseling	0.312 (0.009)	0.010 (0.003)	0.320	7,198
Private counseling	0.402 (0.004)	0.014 (0.002)	0.293	34,768

Notes: The outcome is enrollment in the assigned intensive counseling program. Column (1) reports the weighted average first stage. Column (2) reports its corrected variance across sites. Standard errors, based on the influence functions in Appendix A, appear in parentheses. Column (3) reports the estimated standard deviation relative to the average first stage. Results use data from Behaghel, Crépon, and Gurgand (2014).

Table 5: Association between site-specific first stages and LATEs

	$\widehat{\beta}_{\text{ITT,FS}} - \widehat{\text{LATE}}$	Standard error	Units
	(1)	(2)	(3)
Public counseling	-0.086	0.224	7,198
Private counseling	-0.081	0.095	34,768

Notes: Column (1) reports the statistic whose sign identifies the sign of the complier-weighted covariance between site-specific LATEs and first stages, as shown in (29). Column (2) reports its standard error. Results use data from Behaghel, Crépon, and Gurgand (2014).

2.1 and 2.8 times the respective overall LATE. Under squared-first-stage weights, which give more importance to sites with stronger first stages, the estimated average LATEs are slightly lower, at 6.9 and 4.1 pp, respectively. Still, the similarity of LATEs under first-stage and squared-first-stage weights indicates that changing the weights does not materially shift the center of the LATE distribution in this application. With these weights, the estimated LATE variances are again positive and statistically significant, and the estimated standard deviations of site-specific LATEs are approximately 28.1 pp for public counseling and 20.5 pp for private counseling, about four and five times the respective weighted average effects.

Table 6: Variance of LATEs across sites

	$\widehat{\text{LATE}}$	$\widehat{\sigma}_{\text{LATE}}^2$	$\widehat{\text{LATE}}_{\text{FS}^2}$	$\widehat{\sigma}_{\text{LATE,FS}^2}^2$	Units
	(1)	(2)	(3)	(4)	(5)
Public counseling	0.077 (0.044)	0.027 (0.012)	0.069 (0.048)	0.079 (0.034)	7,198
Private counseling	0.048 (0.025)	0.018 (0.005)	0.041 (0.024)	0.042 (0.013)	34,768

Notes: Column (1) reports the overall LATE under the complier weights in (19). Column (2) reports the estimated lower bound for the complier-weighted LATE variance in (27). Column (3) reports $\widehat{\text{LATE}}_{\text{FS}^2} = \widehat{r}/\widehat{m}_2$ in (25). Column (4) reports the squared-first-stage-weighted variance estimator in (26). Standard errors appear in parentheses, using (58) for Column (1), (57) for Column (2), (50) for Column (3), and (56) for Column (4). Results use data from Behaghel, Crépon, and Gurgand (2014).

Finally, following Section 3.5.4 we relate the two programs' site-specific LATEs. Letting the private program be arm 1 and the public program be arm 2, β^{12} in (32) is the analogue, for LATEs, of the regression of the public program's site-specific effect on the private program's effect reported in Section 4.4. Using the 200 sites where all three arms have at least two units, the estimator in

(66) is equal to 0.728, with a standard error based on (69) equal to 0.274 . The positive correlation between the programs’ effects therefore persists when considering LATEs instead of ITTs.

5 Conclusion

We develop methods to investigate treatment-effect heterogeneity across sites in multi-site RCTs, even when each site contains too few observations to estimate its effect precisely. Combining established and new measurement-error corrections, we estimate the variance and centered third moment of site effects, as well as coefficients from regressions relating site effects to their untreated outcomes or to their effects on other outcomes, including potential mediators. We construct confidence intervals that are asymptotically valid as the number of sites grows, even when the number of units per site remains bounded. With imperfect compliance, we also identify a squared-first-stage-weighted variance of site-specific LATEs and a lower bound for their complier-weighted variance, allowing researchers to assess whether differences in take-up alone explain ITT heterogeneity.

Our application to public and private counseling programs for French job seekers illustrates why looking beyond average effects matters. Average employment gains of approximately 2 percentage points conceal cross-site standard deviations of ITT effects of 8–9 percentage points. We also find substantial LATE heterogeneity, so differences in take-up cannot fully explain the variation in ITT effects. Estimated effects are larger in sites with lower untreated employment rates. Finally, the public and private programs tend to be more effective in the same sites despite being implemented by different providers. This pattern is consistent with a role for site responsiveness, although correlated provider effectiveness could also explain it.

References

- AARONSON, D., L. BARROW, AND W. SANDER (2007): “Teachers and student achievement in the Chicago public high schools,” *Journal of labor Economics*, 25(1), 95–135.
- ADUSUMILLI, K., F. AGOSTINELLI, AND E. BORGHESEAN (2024): “Heterogeneity and Endogenous Compliance: Implications for Scaling Class Size Interventions,” Discussion paper, National Bureau of Economic Research.
- ANGRIST, J. D., AND G. W. IMBENS (1995): “Two-stage least squares estimation of average causal effects in models with variable treatment intensity,” *Journal of the American statistical Association*, 90(430), 431–442.
- 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.
- BEN-MOSHE, D., AND D. GENESOVE (2026): “Unbiased Estimation of Central Moments in Unbalanced Two- and Three-Level Models,” *Economics Letters*, 265, 113013.

- BLOOM, H. S., C. J. HILL, AND J. A. RICCIO (2003): “Linking Program Implementation and Effectiveness: Lessons from a Pooled Sample of Welfare-to-Work Experiments,” *Journal of Policy Analysis and Management*, 22(4), 551–575.
- CHEN, J., J. GU, AND S. KWON (2025): “Empirical Bayes Shrinkage (Mostly) Does Not Correct the Measurement Error in Regression,” Working paper, arXiv:2503.19095.
- DANIELS, M. J., AND M. D. HUGHES (1997): “Meta-analysis for the Evaluation of Potential Surrogate Markers,” *Statistics in Medicine*, 16(17), 1965–1982.
- DEATON, A. (1985): “Panel data from time series of cross-sections,” *Journal of econometrics*, 30(1-2), 109–126.
- DERSIMONIAN, R., AND N. LAIRD (1986): “Meta-analysis in clinical trials,” *Controlled clinical trials*, 7(3), 177–188.
- DUNN, G., R. EMSLEY, H. LIU, S. LANDAU, J. GREEN, I. WHITE, AND A. PICKLES (2015): “Evaluation and Validation of Social and Psychological Markers in Randomised Trials of Complex Interventions in Mental Health: A Methodological Research Programme,” Health Technology Assessment 19.93, NIHR Journals Library, Southampton, UK.
- FULLER, W. A. (1987): “Measurement error models,” *Wiley Series in Probability and Mathematical Statistics*.
- HEMMING, K., J. P. HUGHES, J. E. MCKENZIE, AND A. B. FORBES (2021): “Extending the I-squared statistic to describe treatment effect heterogeneity in cluster, multi-centre randomized trials and individual patient data meta-analysis,” *Statistical Methods in Medical Research*, 30(2), 376–395.
- HONG, G., J. DEUTSCH, P. KRESS, J. E. TRINIDAD, AND Z. XU (2024): “Evaluating Organizational Effectiveness: A New Strategy to Leverage Multisite Randomized Trials for Valid Assessment,” Working paper, arXiv:2407.18360.
- IMAI, K., L. KEELE, AND T. YAMAMOTO (2010): “Identification, Inference and Sensitivity Analysis for Causal Mediation Effects,” *Statistical Science*, 25(1), 51–71.
- KAHAN, B. C., AND M. O. HARHAY (2015): “Many multicenter trials had few events per center, requiring analysis via random-effects models or GEEs,” *Journal of Clinical Epidemiology*, 68(12), 1504–1511.
- KLINE, P., E. K. ROSE, AND C. R. WALTERS (2022): “Systemic discrimination among large US employers,” *The Quarterly Journal of Economics*, 137(4), 1963–2036.
- KLINE, P., R. SAGGIO, AND M. SØLVSTEN (2020): “Leave-out estimation of variance components,” *Econometrica*, 88(5), 1859–1898.
- KRUEGER, A. B., AND L. H. SUMMERS (1988): “Efficiency wages and the inter-industry wage structure,” *Econometrica: Journal of the Econometric Society*, pp. 259–293.
- KWON, S., AND J. ROTH (2026): “Testing Mechanisms,” Working paper.
- LINGSMA, H. F., B. ROOZENBEEK, P. PEREL, I. ROBERTS, A. I. MAAS, AND E. W. STEYERBERG (2011): “Between-centre differences and treatment effects in randomized controlled trials: A case study in traumatic brain injury,” *Trials*, 12, 201.

- MIRATRIX, L. W., J. S. SEKHON, AND B. YU (2013): “Adjusting Treatment Effect Estimates by Post-Stratification in Randomized Experiments,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(2), 369–396.
- REARDON, S. F., AND S. W. RAUDENBUSH (2013): “Under What Assumptions Do Site-by-Treatment Instruments Identify Average Causal Effects?,” *Sociological Methods & Research*, 42(2), 143–163.
- RIMOLDINI, L. (2014): “Skewness and Kurtosis Unbiased by Gaussian Uncertainties,” *Astronomy and Computing*, 5, 1–8.
- ROSE, E. K., J. T. SCHELLENBERG, AND Y. SHEM-TOV (2022): “The effects of teacher quality on adult criminal justice contact,” Discussion paper, National Bureau of Economic Research.
- VANDERWEELE, T. J. (2013): “A Three-Way Decomposition of a Total Effect into Direct, Indirect, and Interactive Effects,” *Epidemiology*, 24(2), 224–232.
- 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.

A Estimators’ influence function and inference

A.1 Variance, regression, and centered-third-moment estimators

Let

$$\psi_{\sigma^2,i} = w(S_i) \left[\{\hat{\mu}_Y(S_i) - \mathbb{E}_w[\mu_Y(S)]\}^2 - \hat{V}_{Y,i} - \sigma^2 \right], \quad (34)$$

$$\begin{aligned} \psi_{\beta,i} = \text{Var}_w(\mu_X(S))^{-1} w(S_i) & \left[\{\hat{\mu}_X(S_i) - \mathbb{E}_w[\mu_X(S)]\} \{\hat{\mu}_Y(S_i) - \mathbb{E}_w[\mu_Y(S)] - \{\hat{\mu}_X(S_i) - \mathbb{E}_w[\mu_X(S)]\}^\top \beta\} \right. \\ & \left. - \hat{C}_{XY,i} + \hat{V}_{X,i} \beta \right] \end{aligned} \quad (35)$$

$$\begin{aligned} \psi_{\kappa_Y,i} = w(S_i) & \left[\{\hat{\mu}_Y(S_i) - \mathbb{E}_w[\mu_Y(S)]\}^3 - 3\{\hat{\mu}_Y(S_i) - \mathbb{E}_w[\mu_Y(S)]\} \hat{V}_{Y,i} + 3\hat{C}_{YV,i} - \hat{K}_{Y,i} - \kappa_Y \right. \\ & \left. - 3\sigma^2 \{\hat{\mu}_Y(S_i) - \mathbb{E}_w[\mu_Y(S)]\} \right]. \end{aligned} \quad (36)$$

Lemma 1 and the definition of β imply $\mathbb{E}[\psi_{\sigma^2,i}] = 0$, $\mathbb{E}[\psi_{\beta,i}] = 0$, and $\mathbb{E}[\psi_{\kappa_Y,i}] = 0$.

Assumption 7 (Conditions for estimating σ^2). (i) $\mathbb{E}[w(S_i)^2 \{1 + \hat{\mu}_Y(S_i)^2\}] < \infty$.

(ii) $\mathbb{E}[\psi_{\sigma^2,i}^2] < \infty$.

Assumption 8 (Conditions for estimating β). (i) $\mathbb{E}[w(S_i)^2 \{1 + \|\hat{\mu}_X(S_i)\|^2 + \hat{\mu}_Y(S_i)^2\}] < \infty$ and $\mathbb{E}[w(S_i) \|\hat{\mu}_X(S_i)\|^2] < \infty$.

(ii) For every $(j, k) \in \{1, \dots, K\}^2$,

$$\mathbb{E} \left[|w(S_i) \hat{V}_{X,i,jk}| \right] < \infty,$$

where $\hat{V}_{X,i,jk}$ denotes the (j, k) entry of the matrix $\hat{V}_{X,i}$.

(iii) $\text{Var}_w(\mu_X(S))$ is positive definite.

(iv) $\mathbb{E}[\|\psi_{\beta,i}\|^2] < \infty$.

Assumption 9 (Conditions for estimating κ_Y). (i) $\mathbb{E}[w(S_i)^2\{1 + \hat{\mu}_Y(S_i)^2\}] < \infty$.

(ii) $\mathbb{E}[w(S_i)\{\hat{\mu}_Y(S_i)^2 - \hat{V}_{Y,i}\}] < \infty$.

(iii) $\mathbb{E}[\psi_{\kappa_Y,i}^2] < \infty$.

Assumptions 7(ii), 8(iv), and Assumption 9(iii) are imposed directly on the influence functions. Stronger primitive sufficient conditions could instead be imposed on the estimated conditional means and on the estimated variances and covariance.

Let

$$\hat{\psi}_{\sigma^2,i} = \hat{w}(S_i) \left[\{\hat{\mu}_Y(S_i) - \bar{\mu}_Y\}^2 - \hat{V}_{Y,i} - \hat{\sigma}^2 \right], \quad (37)$$

$$\begin{aligned} \hat{\psi}_{\beta,i} = & \widehat{\text{Var}}(\mu_X(S))^{-1} \hat{w}(S_i) \left[\{\hat{\mu}_X(S_i) - \bar{\mu}_X\} \{\hat{\mu}_Y(S_i) - \bar{\mu}_Y - \{\hat{\mu}_X(S_i) - \bar{\mu}_X\}^\top \hat{\beta}\} \right. \\ & \left. - \hat{C}_{XY,i} + \hat{V}_{X,i} \hat{\beta} \right] \end{aligned} \quad (38)$$

$$\begin{aligned} \hat{\psi}_{\kappa_Y,i} = & \hat{w}(S_i) \left[\{\hat{\mu}_Y(S_i) - \bar{\mu}_Y\}^3 - 3\{\hat{\mu}_Y(S_i) - \bar{\mu}_Y\} \hat{V}_{Y,i} + 3\hat{C}_{YV,i} - \hat{K}_{Y,i} - \hat{\kappa}_Y \right. \\ & \left. - 3\hat{\sigma}^2 \{\hat{\mu}_Y(S_i) - \bar{\mu}_Y\} \right]. \end{aligned} \quad (39)$$

Let

$$\hat{V}_{\sigma^2} = \frac{1}{n} \sum_{i=1}^n \hat{\psi}_{\sigma^2,i}^2, \quad (40)$$

$$\hat{V}_{\beta} = \frac{1}{n} \sum_{i=1}^n \hat{\psi}_{\beta,i} \hat{\psi}_{\beta,i}^\top, \quad (41)$$

$$\hat{V}_{\kappa_Y} = \frac{1}{n} \sum_{i=1}^n \hat{\psi}_{\kappa_Y,i}^2. \quad (42)$$

Assumption 10 (Additional conditions for estimating V_{β}).

$$\mathbb{E}[w(S_i)^2 \|\hat{\mu}_X(S_i)\|^4] < \infty, \quad \mathbb{E}[w(S_i)^2 \hat{V}_{X,i,jk}^2] < \infty \quad \text{for every } (j,k) \in \{1, \dots, K\}^2.$$

The following theorems establish conditions under which $\hat{\sigma}^2$, $\hat{\beta}$, and $\hat{\kappa}_Y$ are $\sqrt{n}$ -consistent and asymptotically normal. They derive the estimators' influence functions and show that the asymptotic variances of $\hat{\sigma}^2$ and $\hat{\beta}$ can be consistently estimated. To preserve space, we do not prove that $\hat{V}_{\kappa_Y}$ is consistent for the asymptotic variance of $\hat{\kappa}_Y$, but this result holds as well.

Theorem 1. Suppose Assumptions 1 and 7 hold. Then

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{\sigma^2,i} + o_{\mathbb{P}}(1). \quad (43)$$

Consequently,

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{d} N(0, V_{\sigma^2}), \quad V_{\sigma^2} := \mathbb{E}[\psi_{\sigma^2,i}^2].$$

Moreover,

$$\hat{V}_{\sigma^2} \xrightarrow{\mathbb{P}} V_{\sigma^2}.$$

Remark 6 (Possibly negative variance estimator). (5) shows that $\hat{\sigma}^2$ can be negative in finite samples, even though $\sigma^2 \geq 0$. Truncating the estimator at zero changes its asymptotic distribution if $\sigma^2 = 0$ and is therefore not imposed here.

Theorem 2. Suppose Assumptions 1 and 8 hold. Then $\widehat{\text{Var}}(\mu_X(S))$ is nonsingular with probability approaching one and

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{\beta,i} + o_{\mathbb{P}}(1). \quad (44)$$

Consequently,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V_{\beta}), \quad V_{\beta} := \mathbb{E}[\psi_{\beta,i} \psi_{\beta,i}^{\top}].$$

Moreover, if Assumption 10 further holds,

$$\hat{V}_{\beta} \xrightarrow{\mathbb{P}} V_{\beta}.$$

If $\sigma^2 > 0$, the R^2 of the weighted population regression of $\mu_Y(S)$ on a constant and $\mu_X(S)$ is

$$R^2 = \frac{\text{Var}_w(\mu_X(S)^{\top} \beta)}{\text{Var}_w(\mu_Y(S))} = \frac{\beta^{\top} \text{Var}_w(\mu_X(S)) \beta}{\sigma^2}.$$

Whenever $\hat{\beta}$ and the ratio below are well defined, we estimate R^2 by the plug-in estimator

$$\hat{R}^2 = \frac{\hat{\beta}^{\top} \widehat{\text{Var}}(\mu_X(S)) \hat{\beta}}{\hat{\sigma}^2}. \quad (45)$$

Because the corrected variance estimators need not be positive semidefinite in finite samples, $\hat{R}^2$ need not belong to $[0, 1]$.

Theorem 3. Suppose Assumptions 1 and 9 hold. Then

$$\sqrt{n}(\hat{\kappa}_Y - \kappa_Y) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{\kappa_Y,i} + o_{\mathbb{P}}(1). \quad (46)$$

Consequently,

$$\sqrt{n}(\hat{\kappa}_Y - \kappa_Y) \xrightarrow{d} N(0, V_{\kappa_Y}), \quad V_{\kappa_Y} := \mathbb{E}[\psi_{\kappa_Y,i}^2].$$

A.2 Inference on the mediation residual variance

The first-step estimation of β_{med} does not contribute to the asymptotic variance of $\hat{\sigma}_{\varepsilon}^2$. To see this, write

$$\sigma_{\varepsilon}^2 = \text{Var}_w(\mu_Y(S) - \beta_{\text{med}} \mu_X(S)),$$

and, for any scalar b , let $\hat{\sigma}_{\varepsilon}^2(b)$ denote the estimator in Section 3.3.3 with $\hat{\beta}_{\text{med}}$ replaced by b . By construction,

$$\begin{aligned} \hat{\sigma}_{\varepsilon}^2(b) &= \widehat{\text{Var}}(\mu_Y(S)) - 2b \widehat{\text{Cov}}(\mu_X(S), \mu_Y(S)) \\ &\quad + b^2 \widehat{\text{Var}}(\mu_X(S)). \end{aligned}$$

For compactness, let

$$\widehat{V}_2 = \widehat{\text{Var}}(\mu_Y(S)), \quad \widehat{V}_1 = \widehat{\text{Var}}(\mu_X(S)), \quad \widehat{C}_{12} = \widehat{\text{Cov}}(\mu_X(S), \mu_Y(S)).$$

With this notation, $\widehat{\beta}_{\text{med}} = \widehat{C}_{12}/\widehat{V}_1$. Therefore, for any scalar b ,

$$\begin{aligned} \widehat{\sigma}_\varepsilon^2(b) &= \widehat{V}_2 - 2b\widehat{C}_{12} + b^2\widehat{V}_1 \\ &= \widehat{V}_2 + \widehat{V}_1 \left\{ b^2 - 2b\frac{\widehat{C}_{12}}{\widehat{V}_1} \right\} \\ &= \widehat{V}_2 + \widehat{V}_1 \left\{ \left(b - \frac{\widehat{C}_{12}}{\widehat{V}_1} \right)^2 - \left(\frac{\widehat{C}_{12}}{\widehat{V}_1} \right)^2 \right\} \\ &= \widehat{V}_2 - \frac{\widehat{C}_{12}^2}{\widehat{V}_1} + \widehat{V}_1(b - \widehat{\beta}_{\text{med}})^2. \end{aligned}$$

Since $\widehat{\sigma}_\varepsilon^2 = \widehat{\sigma}_\varepsilon^2(\widehat{\beta}_{\text{med}})$, evaluating the last display first at $b = \widehat{\beta}_{\text{med}}$ and then at $b = \beta_{\text{med}}$ yields

$$\widehat{\sigma}_\varepsilon^2 - \widehat{\sigma}_\varepsilon^2(\beta_{\text{med}}) = -\widehat{V}_1(\widehat{\beta}_{\text{med}} - \beta_{\text{med}})^2 = o_{\mathbb{P}}(n^{-1/2}).$$

Here, the final equality follows because $\widehat{\beta}_{\text{med}}$ is $\sqrt{n}$ -consistent and the corrected regressor variance converges. Therefore, the influence function of $\widehat{\sigma}_\varepsilon^2$ is

$$\begin{aligned} \psi_{\sigma_\varepsilon^2, i} &= w(S_i) \left[\left\{ \widehat{\mu}_Y(S_i) - \beta_{\text{med}}\widehat{\mu}_X(S_i) - \mathbb{E}_w[\mu_Y(S) - \beta_{\text{med}}\mu_X(S)] \right\}^2 \right. \\ &\quad \left. - \widehat{V}_{Y, i} - \beta_{\text{med}}^2\widehat{V}_{X, i} + 2\beta_{\text{med}}\widehat{C}_{XY, i} - \sigma_\varepsilon^2 \right], \end{aligned}$$

which can be estimated by

$$\widehat{\psi}_{\sigma_\varepsilon^2, i} = \widehat{w}(S_i) \left[\left\{ \widehat{\mu}_\varepsilon(S_i) - \widehat{\tau}_\varepsilon \right\}^2 - \widehat{V}_{\varepsilon, i} - \widehat{\sigma}_\varepsilon^2 \right].$$

Under the conditions of Theorems 1 and 2 and the corresponding moment conditions for the transformed outcome

$$Y_i - \beta_{\text{med}}X_i = O_{i,2}(1) - O_{i,2}(0) - \beta_{\text{med}}\{O_{i,1}(1) - O_{i,1}(0)\},$$

one can show that

$$\sqrt{n}(\widehat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{\sigma_\varepsilon^2, i} + o_{\mathbb{P}}(1),$$

and

$$\widehat{V}_{\sigma_\varepsilon^2} = \frac{1}{n} \sum_{i=1}^n \widehat{\psi}_{\sigma_\varepsilon^2, i}^2$$

is a consistent estimator of the asymptotic variance of $\widehat{\sigma}_\varepsilon^2$.

One can therefore test the implication of (14) in (16) by comparing $\sqrt{n}\widehat{\sigma}_\varepsilon^2/\widehat{V}_{\sigma_\varepsilon^2}^{1/2}$ to standard normal critical values.

A.3 Inference on $\underline{\sigma}_{\text{LATE}}^2$ and $\sigma_{\text{LATE,FS}^2}^2$

Use the estimators and notation from Section 3.5.2, and let

$$Q(l) = \mathbb{E}_w[\{\mu_{\text{ITT}}(S) - l\mu_{\text{FS}}(S)\}^2], \quad q = Q(\text{LATE}_{\text{FS}^2}), \quad h = Q(\text{LATE}).$$

For the lower bound, write $\delta = \text{FS} - m_2 > 0$ and $d = r - m_2 \text{LATE}$, so that $\underline{\sigma}_{\text{LATE}}^2 = (h + d^2/\delta)/\text{FS}$. Define

$$\psi_{q,i} = w(S_i)\{g_i(\text{LATE}_{\text{FS}^2}) - q\}, \quad (47)$$

$$\psi_{\text{FS}^2,i} = w(S_i)\{\hat{\mu}_{\text{FS}}(S_i)^2 - \hat{V}_{\text{FS},i} - m_2\}, \quad (48)$$

$$\psi_{\sigma_{\text{LATE,FS}^2}^2,i} = \frac{\psi_{q,i} - \sigma_{\text{LATE,FS}^2}^2 \psi_{\text{FS}^2,i}}{m_2}. \quad (49)$$

As $\text{LATE}_{\text{FS}^2}$ minimizes $Q(l)$, its estimation has no first-order effect on $\hat{q}$. With

$$\psi_{r,i} = w(S_i)\{\hat{\mu}_{\text{FS}}(S_i)\hat{\mu}_{\text{ITT}}(S_i) - \hat{C}_{\text{FS,ITT},i} - r\},$$

the influence function of its estimator is

$$\psi_{\text{LATE}_{\text{FS}^2},i} = \frac{\psi_{r,i} - \text{LATE}_{\text{FS}^2} \psi_{\text{FS}^2,i}}{m_2}. \quad (50)$$

For the lower bound, define the influence functions of the average first stage and overall LATE:

$$\psi_{\text{FS},i} = w(S_i)\{\hat{\mu}_{\text{FS}}(S_i) - \text{FS}\}, \quad \psi_{\text{LATE},i} = \frac{w(S_i)\{\hat{\mu}_{\text{ITT}}(S_i) - \text{LATE}\hat{\mu}_{\text{FS}}(S_i)\}}{\text{FS}}. \quad (51)$$

The derivative $Q'(\text{LATE}) = 2(m_2 \text{LATE} - r)$ need not vanish. Consequently, define

$$\psi_{h,i} = w(S_i)\{g_i(\text{LATE}) - h\} + 2(m_2 \text{LATE} - r)\psi_{\text{LATE},i}, \quad (52)$$

$$\psi_{d,i} = \psi_{r,i} - \text{LATE} \psi_{\text{FS}^2,i} - m_2 \psi_{\text{LATE},i},$$

$$\psi_{\delta,i} = \psi_{\text{FS},i} - \psi_{\text{FS}^2,i},$$

$$\psi_{\underline{\sigma}_{\text{LATE}}^2,i} = \frac{\psi_{h,i} + (2d/\delta)\psi_{d,i} - (d^2/\delta^2)\psi_{\delta,i} - \underline{\sigma}_{\text{LATE}}^2 \psi_{\text{FS},i}}{\text{FS}}. \quad (53)$$

The extra term in (52) accounts for estimating the centering constant LATE.

Assumption 11 (Conditions for estimating $\sigma_{\text{LATE,FS}^2}^2$). *The random variables $\psi_{r,i}$, $\psi_{q,i}$, and $\psi_{\text{FS}^2,i}$ have finite second moments.*

Assumption 12 (Additional conditions for estimating the lower bound). *The first-stage moments satisfy $\delta = \text{FS} - m_2 > 0$, and*

$$\mathbb{E}[w(S_i)^2\{1 + \hat{\mu}_{\text{ITT}}(S_i)^2\}] < \infty.$$

Theorem 4. *Suppose Assumptions 2, 3, 6, and 11 hold.*

1. *Then*

$$\sqrt{n}(\hat{\sigma}_{\text{LATE,FS}^2}^2 - \sigma_{\text{LATE,FS}^2}^2) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{\sigma_{\text{LATE,FS}^2}^2,i} + o_{\mathbb{P}}(1). \quad (54)$$

Consequently,

$$\sqrt{n}(\hat{\sigma}_{\text{LATE,FS}^2}^2 - \sigma_{\text{LATE,FS}^2}^2) \xrightarrow{d} N(0, V_{\sigma_{\text{LATE,FS}^2}^2}^2), \quad V_{\sigma_{\text{LATE,FS}^2}^2}^2 := \mathbb{E}[\psi_{\sigma_{\text{LATE,FS}^2}^2,i}^2].$$

2. Suppose Assumption 12 also holds. Then

$$\sqrt{n}(\widehat{\sigma}_{\text{LATE}}^2 - \underline{\sigma}_{\text{LATE}}^2) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{\underline{\sigma}_{\text{LATE}}^2, i} + o_{\mathbb{P}}(1). \quad (55)$$

Consequently,

$$\sqrt{n}(\widehat{\sigma}_{\text{LATE}}^2 - \underline{\sigma}_{\text{LATE}}^2) \xrightarrow{d} N(0, V_{\underline{\sigma}_{\text{LATE}}^2}), \quad V_{\underline{\sigma}_{\text{LATE}}^2} := \mathbb{E}[\psi_{\underline{\sigma}_{\text{LATE}}^2, i}^2].$$

For implementation, set

$$\begin{aligned} \widehat{\psi}_{q,i} &= \widehat{w}(S_i) \{g_i(\widehat{\text{LATE}}_{\text{FS}^2}) - \widehat{q}\}, \\ \widehat{\psi}_{\text{FS}^2, i} &= \widehat{w}(S_i) \{\widehat{\mu}_{\text{FS}}(S_i)^2 - \widehat{V}_{\text{FS}, i} - \widehat{m}_2\}, \\ \widehat{\psi}_{\sigma_{\text{LATE}, \text{FS}^2}^2, i} &= \frac{\widehat{\psi}_{q,i} - \widehat{\sigma}_{\text{LATE}, \text{FS}^2}^2 \widehat{\psi}_{\text{FS}^2, i}}{\widehat{m}_2}, \\ \widehat{\psi}_{\text{FS}, i} &= \widehat{w}(S_i) \{\widehat{\mu}_{\text{FS}}(S_i) - \widehat{\text{FS}}\}, \\ \widehat{\psi}_{\text{LATE}, i} &= \frac{\widehat{w}(S_i) \{\widehat{\mu}_{\text{ITT}}(S_i) - \widehat{\text{LATE}} \widehat{\mu}_{\text{FS}}(S_i)\}}{\widehat{\text{FS}}}, \\ \widehat{h} &= \widehat{Q}(\widehat{\text{LATE}}), \\ \widehat{\psi}_{h,i} &= \widehat{w}(S_i) \{g_i(\widehat{\text{LATE}}) - \widehat{h}\} + 2(\widehat{m}_2 \widehat{\text{LATE}} - \widehat{r}) \widehat{\psi}_{\text{LATE}, i}, \\ \widehat{d} &= \widehat{r} - \widehat{m}_2 \widehat{\text{LATE}}, \quad \widehat{\delta} = \widehat{\text{FS}} - \widehat{m}_2, \\ \widehat{\psi}_{r,i} &= \widehat{w}(S_i) \{\widehat{\mu}_{\text{FS}}(S_i) \widehat{\mu}_{\text{ITT}}(S_i) - \widehat{C}_{\text{FS}, \text{ITT}, i} - \widehat{r}\}, \\ \widehat{\psi}_{d,i} &= \widehat{\psi}_{r,i} - \widehat{\text{LATE}} \widehat{\psi}_{\text{FS}^2, i} - \widehat{m}_2 \widehat{\psi}_{\text{LATE}, i}, \\ \widehat{\psi}_{\delta, i} &= \widehat{\psi}_{\text{FS}, i} - \widehat{\psi}_{\text{FS}^2, i}, \\ \widehat{\psi}_{\sigma_{\text{LATE}}^2, i} &= \frac{\widehat{\psi}_{h,i} + (2\widehat{d}/\widehat{\delta}) \widehat{\psi}_{d,i} - (\widehat{d}^2/\widehat{\delta}^2) \widehat{\psi}_{\delta, i} - \widehat{\sigma}_{\text{LATE}}^2 \widehat{\psi}_{\text{FS}, i}}{\widehat{\text{FS}}}. \end{aligned}$$

The asymptotic-variance estimators are

$$\widehat{V}_{\sigma_{\text{LATE}, \text{FS}^2}^2} = \frac{1}{n} \sum_{i=1}^n \widehat{\psi}_{\sigma_{\text{LATE}, \text{FS}^2}^2, i}^2, \quad (56)$$

$$\widehat{V}_{\sigma_{\text{LATE}}^2} = \frac{1}{n} \sum_{i=1}^n \widehat{\psi}_{\sigma_{\text{LATE}}^2, i}^2. \quad (57)$$

A.4 Inference on the covariance between sites' LATEs and first-stage effects

We give a heuristic calculation of $\widehat{\sigma}_{\text{LATE}, \text{FS}}$'s influence function, using $\psi_{\text{FS}, i}$ and $\psi_{\text{LATE}, i}$ defined in (51). The influence functions of $\widehat{C}_{\text{FS}, \text{ITT}}$ and $\widehat{V}_{\text{FS}}$ are, respectively,

$$\begin{aligned} \psi_{C_{\text{FS}, \text{ITT}}, i} &= w(S_i) \left[\{\widehat{\mu}_{\text{FS}}(S_i) - \widehat{\text{FS}}\} \{\widehat{\mu}_{\text{ITT}}(S_i) - \widehat{\text{ITT}}\} - \widehat{C}_{\text{FS}, \text{ITT}, i} - \text{Cov}_w(\mu_{\text{FS}}(S), \mu_{\text{ITT}}(S)) \right], \\ \psi_{V_{\text{FS}}, i} &= w(S_i) \left[\{\widehat{\mu}_{\text{FS}}(S_i) - \widehat{\text{FS}}\}^2 - \widehat{V}_{\text{FS}, i} - \text{Var}_w(\mu_{\text{FS}}(S)) \right]. \end{aligned}$$

A ratio expansion gives

$$\sqrt{n}(\widehat{\text{LATE}} - \text{LATE}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{\text{LATE},i} + o_{\mathbb{P}}(1). \quad (58)$$

Applying the delta method to (30) gives

$$\psi_{\text{LATE,FS},i} = \frac{1}{\text{FS}} (\psi_{C_{\text{FS,ITT}},i} - \text{LATE} \psi_{V_{\text{FS}},i} - \text{Var}_w(\mu_{\text{FS}}(S)) \psi_{\text{LATE},i} - \sigma_{\text{LATE,FS}} \psi_{\text{FS},i}). \quad (59)$$

This suggests estimating the asymptotic variance by replacing all population quantities in (59) with their sample analogues and averaging the squared estimated influence functions across sites.

A.5 Estimation of, and inference on, β^d and β_k^{12}

For any two site effects A and B considered below, let

$$\begin{aligned} \mathcal{M}_{A,B} &= \mathbb{E}_w[\mu_A(S)\mu_B(S)], \\ \widehat{\mathcal{M}}_{A,B} &= \frac{1}{n} \sum_{i=1}^n \widehat{w}(S_i) \{\widehat{\mu}_A(S_i)\widehat{\mu}_B(S_i) - \widehat{C}_{A,B,i}\}, \end{aligned} \quad (60)$$

where $\widehat{C}_{A,B,i}$ estimates the conditional covariance of the two site-effect estimators. The influence function of $\widehat{\mathcal{M}}_{A,B}$ is

$$\psi_{A,B,i} = w(S_i) \{\widehat{\mu}_A(S_i)\widehat{\mu}_B(S_i) - \widehat{C}_{A,B,i} - \mathcal{M}_{A,B}\}. \quad (61)$$

The formulas below are heuristic and require the same types of moment and nonsingularity conditions as Theorem 2.

A.5.1 LATEs for several outcomes under the same treatment arm

For $j \in \{1, \dots, J\}$, estimate the squared-first-stage-weighted mean LATE by

$$\widehat{\text{LATE}}_{d,j,\text{FS}_d^2} = \frac{\widehat{\mathcal{M}}_{\text{FS}_d, \text{ITT}_{d,j}}}{\widehat{\mathcal{M}}_{\text{FS}_d, \text{FS}_d}}. \quad (62)$$

For each site, form

$$\widehat{U}_{d,j}(S_i) = \widehat{\mu}_{\text{ITT}_{d,j}}(S_i) - \widehat{\mu}_{\text{FS}_d}(S_i) \widehat{\text{LATE}}_{d,j,\text{FS}_d^2}.$$

The conditional covariance correction for $\widehat{U}_{d,j}(S_i)$ and $\widehat{U}_{d,\ell}(S_i)$ is

$$\begin{aligned} \widehat{C}_{U_{d,j}, U_{d,\ell}, i} &= \widehat{C}_{\text{ITT}_{d,j}, \text{ITT}_{d,\ell}, i} - \widehat{\text{LATE}}_{d,\ell, \text{FS}_d^2} \widehat{C}_{\text{ITT}_{d,j}, \text{FS}_d, i} \\ &\quad - \widehat{\text{LATE}}_{d,j, \text{FS}_d^2} \widehat{C}_{\text{FS}_d, \text{ITT}_{d,\ell}, i} + \widehat{\text{LATE}}_{d,j, \text{FS}_d^2} \widehat{\text{LATE}}_{d,\ell, \text{FS}_d^2} \widehat{V}_{\text{FS}_d, i}. \end{aligned} \quad (63)$$

Stacking outcomes $2, \dots, J$ in $\widehat{U}_{d,-1}(S_i)$, the estimator is

$$\widehat{\beta}^d = \left[\frac{1}{n} \sum_{i=1}^n \widehat{w}(S_i) \{ \widehat{U}_{d,-1}(S_i) \widehat{U}_{d,-1}(S_i)^\top - \widehat{V}_{U_{d,-1}, i} \} \right]^{-1} \left[\frac{1}{n} \sum_{i=1}^n \widehat{w}(S_i) \{ \widehat{U}_{d,-1}(S_i) \widehat{U}_{d,1}(S_i) - \widehat{C}_{U_{d,-1}, U_{d,1}, i} \} \right]. \quad (64)$$

Let $U_{d,j}(S) = \mu_{\text{ITT}_{d,j}}(S) - \mu_{\text{FS}_d}(S)\text{LATE}_{d,j,\text{FS}_d^2}$, and

$$\tilde{U}_{d,j,i} = \hat{\mu}_{\text{ITT}_{d,j}}(S_i) - \hat{\mu}_{\text{FS}_d}(S_i)\text{LATE}_{d,j,\text{FS}_d^2}.$$

For $j, \ell \in \{1, \dots, J\}$, let $\tilde{C}_{U_{d,j}, U_{d,\ell}, i}$ denote (63) with the estimated weighted LATEs replaced by their population counterparts. Stack $\tilde{U}_{d,2,i}, \dots, \tilde{U}_{d,J,i}$ in $\tilde{U}_{d,-1,i}$. Let the (j, ℓ) entry of $\tilde{V}_{U_{d,-1}, i}$ be $\tilde{C}_{U_{d,j}, U_{d,\ell}, i}$ and the j th entry of $\tilde{C}_{U_{d,-1}, U_{d,1}, i}$ be $\tilde{C}_{U_{d,j}, U_{d,1}, i}$, for $j, \ell \in \{2, \dots, J\}$. The influence function of (64) is

$$\begin{aligned} \psi_{\beta^d, i} = & \mathbb{E}_w[U_{d,-1}(S)U_{d,-1}(S)^\top]^{-1}w(S_i)\left[\tilde{U}_{d,-1,i}\{\tilde{U}_{d,1,i} - \tilde{U}_{d,-1,i}^\top\beta^d\} \right. \\ & \left. - \tilde{C}_{U_{d,-1}, U_{d,1}, i} + \tilde{V}_{U_{d,-1}, i}\beta^d\right]. \end{aligned} \quad (65)$$

Estimation of the weighted mean LATEs in (62) has no first-order effect. Indeed, their definition implies the orthogonality condition

$$\mathbb{E}_w[\mu_{\text{FS}_d}(S)U_{d,j}(S)] = 0.$$

Then, one can show that the derivatives of $\mathbb{E}_w[U_{d,j}(S)U_{d,\ell}(S)]$ with respect to $\text{LATE}_{d,j,\text{FS}_d^2}$ and $\text{LATE}_{d,\ell,\text{FS}_d^2}$ are zero.

A.5.2 LATEs for the same outcome under two treatment arms

The plug-in estimator of (33) is

$$\hat{\beta}_k^{12} = \frac{\hat{C}_k^{12}}{\hat{V}_k^1}, \quad (66)$$

where

$$\begin{aligned} \hat{C}_k^{12} &= \frac{\hat{\mathcal{M}}_{\text{ITT}_{1,k}, \text{ITT}_{2,k}}}{\hat{\mathcal{M}}_{\text{FS}_1, \text{FS}_2}} - \frac{\hat{\mathcal{M}}_{\text{FS}_2, \text{ITT}_{1,k}}\hat{\mathcal{M}}_{\text{FS}_1, \text{ITT}_{2,k}}}{\hat{\mathcal{M}}_{\text{FS}_1, \text{FS}_2}^2}, \\ \hat{V}_k^1 &= \frac{\hat{\mathcal{M}}_{\text{ITT}_{1,k}, \text{ITT}_{1,k}}}{\hat{\mathcal{M}}_{\text{FS}_1, \text{FS}_1}} - \frac{\hat{\mathcal{M}}_{\text{FS}_1, \text{ITT}_{1,k}}^2}{\hat{\mathcal{M}}_{\text{FS}_1, \text{FS}_1}^2}. \end{aligned}$$

Let C_k^{12} and V_k^1 denote the probability limits of the preceding two displays. Applying the quotient and product rules to their corrected moments gives

$$\begin{aligned} \psi_{C_k^{12}, i} &= \frac{\psi_{\text{ITT}_{1,k}, \text{ITT}_{2,k}, i}}{\mathcal{M}_{\text{FS}_1, \text{FS}_2}} - \frac{\mathcal{M}_{\text{ITT}_{1,k}, \text{ITT}_{2,k}}\psi_{\text{FS}_1, \text{FS}_2, i}}{\mathcal{M}_{\text{FS}_1, \text{FS}_2}^2} \\ &\quad - \frac{\mathcal{M}_{\text{FS}_1, \text{ITT}_{2,k}}\psi_{\text{FS}_2, \text{ITT}_{1,k}, i} + \mathcal{M}_{\text{FS}_2, \text{ITT}_{1,k}}\psi_{\text{FS}_1, \text{ITT}_{2,k}, i}}{\mathcal{M}_{\text{FS}_1, \text{FS}_2}^2} \\ &\quad + \frac{2\mathcal{M}_{\text{FS}_2, \text{ITT}_{1,k}}\mathcal{M}_{\text{FS}_1, \text{ITT}_{2,k}}\psi_{\text{FS}_1, \text{FS}_2, i}}{\mathcal{M}_{\text{FS}_1, \text{FS}_2}^3}, \end{aligned} \quad (67)$$

$$\begin{aligned} \psi_{V_k^1, i} &= \frac{\psi_{\text{ITT}_{1,k}, \text{ITT}_{1,k}, i}}{\mathcal{M}_{\text{FS}_1, \text{FS}_1}} - \frac{\mathcal{M}_{\text{ITT}_{1,k}, \text{ITT}_{1,k}}\psi_{\text{FS}_1, \text{FS}_1, i}}{\mathcal{M}_{\text{FS}_1, \text{FS}_1}^2} \\ &\quad - \frac{2\mathcal{M}_{\text{FS}_1, \text{ITT}_{1,k}}\psi_{\text{FS}_1, \text{ITT}_{1,k}, i}}{\mathcal{M}_{\text{FS}_1, \text{FS}_1}^2} + \frac{2\mathcal{M}_{\text{FS}_1, \text{ITT}_{1,k}}^2\psi_{\text{FS}_1, \text{FS}_1, i}}{\mathcal{M}_{\text{FS}_1, \text{FS}_1}^3}. \end{aligned} \quad (68)$$

Therefore,

$$\psi_{\beta_k^{12},i} = \frac{\psi_{C_k^{12},i} - \beta_k^{12}\psi_{V_k^1,i}}{V_k^1}. \quad (69)$$

B Simulations

We simulate multi-site RCTs with n sites. Site i has an untreated mean $\alpha_i \sim N(0, 0.2^2)$ and an average treatment effect $\tau_i = 0.1 - 0.5\alpha_i + e_i$, where e_i is independent of α_i , has mean zero and standard deviation 0.3, and is drawn from one of four distributions: normal; Laplace, which is symmetric with heavier tails than the normal; a Beta(2,5) distribution, which is right-skewed with bounded support; and an exponential distribution, which is right-skewed with a long tail. Their standardized skewnesses are 0, 0, 0.6, and 2. Site i has $N_{1,i}$ treated and $N_{0,i}$ control units, drawn independently and uniformly from $\{8, \dots, 25\}$, and receives weight $W(S_i) = N_{1,i} + N_{0,i}$. Unit outcomes are $O_{u,i}(0) = \alpha_i + 0.5(E_{u,i} - 1)$ and $O_{u,i}(1) = O_{u,i}(0) + \tau_i + 0.3(E'_{u,i} - 1)$, where $E_{u,i}$ and $E'_{u,i}$ are independent standard exponential variables, so that unit-level errors are skewed. The targets are $\sigma^2 = \text{Var}_w(\tau_i) = 0.1$, $\beta = -0.5$, the coefficient from the regression of τ_i on α_i as in Section 3.2, and $\kappa = \mathbb{E}_w[(\tau_i - \mathbb{E}_w[\tau_i])^3]$, which equals 0.3^3 times the standardized skewness of e_i .

For each design we draw 1,000 samples and compute $\hat{\sigma}^2$, $\hat{\beta}$, and $\hat{\kappa}_Y$ with the estimators of Section 2, applied to the site-level estimators and corrections of Proposition 1. Table 7 reports the coverage rates of nominal 95% confidence intervals based on the influence functions (IFs) in Theorems 1 to 3. For σ^2 and κ_Y , it also reports the coverage of intervals from a studentized site-level bootstrap with 999 replications: sites are resampled with replacement, the estimator and its influence-function standard error are recomputed in each bootstrap sample, and the interval is formed with critical values equal to the 2.5th and 97.5th percentiles of the bootstrapped t -statistics. Panels A to C use 50, 100, and 200 sites and all four distributions. Panel D uses larger numbers of sites for the exponential distribution.

Table 7: Coverage rates of nominal 95% confidence intervals

Distribution of site effects	σ^2 (Theorem 1)		β (Theorem 2)		κ_Y (Theorem 3)	
	IF	Bootstrap	IF	IF	Bootstrap	
	(1)	(2)	(3)	(4)	(5)	
<i>Panel A: 50 sites</i>						
Normal	0.904	0.947	0.919	0.966	0.904	
Laplace	0.840	0.928	0.920	0.954	0.896	
Beta(2,5)	0.895	0.935	0.916	0.879	0.857	
Exponential	0.839	0.908	0.929	0.621	0.748	
<i>Panel B: 100 sites</i>						
Normal	0.909	0.930	0.943	0.952	0.906	
Laplace	0.896	0.944	0.954	0.966	0.909	
Beta(2,5)	0.917	0.929	0.935	0.914	0.896	
Exponential	0.887	0.925	0.949	0.727	0.831	
<i>Panel C: 200 sites</i>						
Normal	0.944	0.953	0.941	0.954	0.911	
Laplace	0.908	0.935	0.938	0.965	0.898	
Beta(2,5)	0.938	0.954	0.946	0.923	0.904	
Exponential	0.892	0.929	0.938	0.769	0.862	
<i>Panel D: more sites, exponential distribution</i>						
Exponential, 500 sites	0.926	0.949	0.939	0.839	0.908	
Exponential, 1,000 sites	0.943	0.950	0.948	0.889	0.926	
Exponential, 5,000 sites	0.944	0.953	0.943	0.928	0.947	

Notes: IF denotes influence function. Each cell reports the share of 1,000 simulated samples in which a nominal 95% confidence interval contains the true parameter. Columns (1), (3), and (4) use intervals of the form estimate ± 1.96 standard error, with standard errors based on the influence functions in Theorems 1 to 3. Columns (2) and (5) use a studentized site-level bootstrap with 999 replications. The data-generating process is described in the text; the standardized skewness of e_i is 0 for the normal and Laplace distributions, 0.6 for the Beta(2,5), and 2 for the exponential.

C Proofs

C.1 Proof of Lemma 1

Let

$$e_{X,i} = \hat{\mu}_X(S_i) - \mu_X(S_i), \quad e_{Y,i} = \hat{\mu}_Y(S_i) - \mu_Y(S_i).$$

Because $w(S_i)$ is measurable with respect to S_i , weighted expectations obey $\mathbb{E}_w[R_i] = \mathbb{E}_w\{\mathbb{E}[R_i | S_i]\}$ whenever the relevant expectations exist. Assumption 1(ii) implies that $\mathbb{E}[e_{X,i} | S_i] = 0$ and $\mathbb{E}[e_{Y,i} | S_i] = 0$. Therefore,

$$\begin{aligned}
& \mathbb{E}_w \left[\{\hat{\mu}_Y(S_i) - \mathbb{E}_w[\mu_Y(S)]\}^2 - \hat{V}_{Y,i} \right] \\
&= \mathbb{E}_w \left[\{\mu_Y(S_i) - \mathbb{E}_w[\mu_Y(S)]\}^2 + \text{Var}(\hat{\mu}_Y(S_i) | S_i) - \hat{V}_{Y,i} \right] \\
&= \text{Var}_w(\mu_Y(S)),
\end{aligned} \tag{70}$$

where the last equality uses the conditional unbiasedness of $\widehat{V}_{Y,i}$. The matrix-valued analogue proves (2). Similarly,

$$\begin{aligned} & \mathbb{E}_w \left[\{ \widehat{\mu}_X(S_i) - \mathbb{E}_w[\mu_X(S)] \} \{ \widehat{\mu}_Y(S_i) - \mathbb{E}_w[\mu_Y(S)] \} - \widehat{C}_{XY,i} \right] \\ &= \mathbb{E}_w \left[\{ \mu_X(S_i) - \mathbb{E}_w[\mu_X(S)] \} \{ \mu_Y(S_i) - \mathbb{E}_w[\mu_Y(S)] \} + \text{Cov}(\widehat{\mu}_X(S_i), \widehat{\mu}_Y(S_i) \mid S_i) - \widehat{C}_{XY,i} \right] \\ &= \text{Cov}_w(\mu_X(S), \mu_Y(S)), \end{aligned} \tag{71}$$

where the last equality uses the conditional unbiasedness of $\widehat{C}_{XY,i}$. This proves (1)–(3).

To prove (4), let

$$e_i = \widehat{\mu}_Y(S_i) - \mu_Y(S_i), \quad d_i = \mu_Y(S_i) - \mathbb{E}_w[\mu_Y(S)].$$

Conditional on S_i , $\mathbb{E}[e_i \mid S_i] = 0$, so

$$\mathbb{E}[(e_i + d_i)^3 \mid S_i] = d_i^3 + 3d_i \text{Var}(\widehat{\mu}_Y(S_i) \mid S_i) + \mathbb{E}[e_i^3 \mid S_i].$$

Moreover,

$$\mathbb{E}[(e_i + d_i)\widehat{V}_{Y,i} \mid S_i] = d_i \text{Var}(\widehat{\mu}_Y(S_i) \mid S_i) + \text{Cov}(\widehat{\mu}_Y(S_i), \widehat{V}_{Y,i} \mid S_i).$$

For brevity, write

$$v_i = \text{Var}(\widehat{\mu}_Y(S_i) \mid S_i), \quad c_i = \text{Cov}(\widehat{\mu}_Y(S_i), \widehat{V}_{Y,i} \mid S_i), \quad k_i = \mathbb{E}[e_i^3 \mid S_i].$$

The conditional unbiasedness restrictions then imply

$$\begin{aligned} & \mathbb{E} \left[(e_i + d_i)^3 - 3(e_i + d_i)\widehat{V}_{Y,i} + 3\widehat{C}_{YV,i} - \widehat{K}_{Y,i} \mid S_i \right] \\ &= d_i^3 + 3d_i v_i + k_i - 3(d_i v_i + c_i) + 3c_i - k_i \\ &= d_i^3. \end{aligned}$$

Taking a weighted expectation of the previous display, the result follows.

Lemma 2 (Conditional product representation). *For each d , order the indices assigned to arm d as $U_{d,1} < \dots < U_{d,N_{d,i}}$. Conditional on S_i , the collections*

$$(\mathbf{O}_{U_{d,1},i}(d), \dots, \mathbf{O}_{U_{d,N_{d,i}},i}(d)), \quad d = 0, \dots, \mathcal{D},$$

are mutually independent, and the collection for arm d consists of $N_{d,i}$ i.i.d. draws from the conditional marginal distribution of $\mathbf{O}_i(d)$ given S_i .

Proof. Write $\mathbf{D}_i = (D_{u,i})_{u=1}^{N_i}$ and let $\mathbf{P}_i = ((\mathbf{O}_{u,i}(d))_{d=0}^{\mathcal{D}})_{u=1}^{N_i}$ collect the sampled potential outcomes. Fix a feasible assignment vector $\mathbf{d}$, and let $u_{d,1} < \dots < u_{d,N_{d,i}}$ be the indices it assigns to arm d . For arbitrary Borel sets $A_{d,j} \subseteq \mathbb{R}^L$, define

$$E_{\mathbf{d}} := \bigcap_{d=0}^{\mathcal{D}} \bigcap_{j=1}^{N_{d,i}} \{ \mathbf{O}_{u_{d,j},i}(d) \in A_{d,j} \}, \quad q_i := \left(\binom{N_i}{N_{0,i}, \dots, N_{\mathcal{D},i}} \right)^{-1}.$$

Since the arm sizes are measurable with respect to S_i , we have

$$\begin{aligned}
& \mathbb{P} \left(\bigcap_{d=0}^{\mathcal{D}} \bigcap_{j=1}^{N_{d,i}} \{ \mathbf{O}_{U_{d,j},i}(d) \in A_{d,j} \} \mid \mathbf{D}_i = \mathbf{d}, S_i \right) \\
&= \mathbb{P} \left(\bigcap_{d=0}^{\mathcal{D}} \bigcap_{j=1}^{N_{d,i}} \{ \mathbf{O}_{u_{d,j},i}(d) \in A_{d,j} \} \mid \mathbf{D}_i = \mathbf{d}, S_i \right) \\
&= \frac{\mathbb{P}(E_{\mathbf{d}}, \mathbf{D}_i = \mathbf{d} \mid S_i)}{\mathbb{P}(\mathbf{D}_i = \mathbf{d} \mid S_i)} \\
&= \frac{\mathbb{E}[\mathbf{1}\{E_{\mathbf{d}}\} \mathbb{P}(\mathbf{D}_i = \mathbf{d} \mid \mathbf{P}_i, S_i) \mid S_i]}{\mathbb{E}[\mathbb{P}(\mathbf{D}_i = \mathbf{d} \mid \mathbf{P}_i, S_i) \mid S_i]} \\
&= \frac{q_i \mathbb{P}(E_{\mathbf{d}} \mid S_i)}{q_i} = \prod_{d=0}^{\mathcal{D}} \prod_{j=1}^{N_{d,i}} \mathbb{P}(\mathbf{O}_i(d) \in A_{d,j} \mid S_i).
\end{aligned}$$

The first equality uses $U_{d,j} = u_{d,j}$ on the event $\mathbf{D}_i = \mathbf{d}$. The second is the definition of conditional probability, and the third follows by iterated expectations, since $E_{\mathbf{d}}$ is measurable with respect to $(\mathbf{P}_i, S_i)$. The fourth uses Assumption 3, which makes the conditional assignment probability equal to q_i . Finally, all indices $u_{d,j}$ are distinct, so Point 2 of Assumption 2 gives the product in the last equality. This product does not depend on $\mathbf{d}$. Averaging over the assignment conditional on S_i therefore gives the same factorization conditional on S_i alone, proving both mutual independence across arms and i.i.d. sampling within each arm. $\square$

Lemma 3 (Normalized-weight expansion). *Let $(u_i, Z_i)_{i=1}^n$ be i.i.d., where $u_i > 0$, $Z_i \in \mathbb{R}^d$ for some fixed d , and $\mathbb{E}[u_i] = 1$. Define*

$$\bar{u} = \frac{1}{n} \sum_{i=1}^n u_i, \quad \hat{\theta} = \frac{1}{n} \sum_{i=1}^n \frac{u_i}{\bar{u}} Z_i, \quad \theta = \mathbb{E}[u_i Z_i].$$

If $\mathbb{E}[\|u_i(Z_i - \theta)\|^2] < \infty$, then

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n u_i(Z_i - \theta) + o_{\mathbb{P}}(1). \quad (72)$$

Proof. Because $n^{-1} \sum_i (u_i/\bar{u}) = 1$, the following identity is exact:

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{1}{\bar{u}} \frac{1}{\sqrt{n}} \sum_{i=1}^n u_i(Z_i - \theta).$$

The weak law of large numbers gives $\bar{u} \xrightarrow{\mathbb{P}} 1$, while the finite-second-moment condition and the multivariate central limit theorem give $n^{-1/2} \sum_i u_i(Z_i - \theta) = O_{\mathbb{P}}(1)$. Hence replacing $1/\bar{u}$ by one changes the right-hand side by $o_{\mathbb{P}}(1)$, proving (72). $\square$

C.2 Proof of Theorem 1

For brevity, let

$$m_Y = \mathbb{E}_w[\mu_Y(S)], \quad u_i = w(S_i), \quad \bar{u} = \frac{1}{n} \sum_{i=1}^n u_i.$$

Then, $\widehat{w}(S_i) = u_i/\bar{u}$. By Assumption 1(ii) and the law of iterated expectations,

$$\mathbb{E}[u_i \widehat{\mu}_Y(S_i)] = \mathbb{E}[u_i \mathbb{E}\{\widehat{\mu}_Y(S_i) \mid S_i\}] = \mathbb{E}[u_i \mu_Y(S_i)] = m_Y.$$

For every i ,

$$\widehat{\mu}_Y(S_i) - \widehat{\mu}_Y = \{\widehat{\mu}_Y(S_i) - m_Y\} - \{\widehat{\mu}_Y - m_Y\}.$$

Therefore,

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \widehat{w}(S_i) \{\widehat{\mu}_Y(S_i) - \widehat{\mu}_Y\}^2 \\ &= \frac{1}{n} \sum_{i=1}^n \widehat{w}(S_i) \left[\{\widehat{\mu}_Y(S_i) - m_Y\}^2 - 2\{\widehat{\mu}_Y(S_i) - m_Y\}\{\widehat{\mu}_Y - m_Y\} + \{\widehat{\mu}_Y - m_Y\}^2 \right] \\ &= \frac{1}{n} \sum_{i=1}^n \widehat{w}(S_i) \{\widehat{\mu}_Y(S_i) - m_Y\}^2 - 2\{\widehat{\mu}_Y - m_Y\} \frac{1}{n} \sum_{i=1}^n \widehat{w}(S_i) \{\widehat{\mu}_Y(S_i) - m_Y\} + \{\widehat{\mu}_Y - m_Y\}^2 \\ &= \frac{1}{n} \sum_{i=1}^n \widehat{w}(S_i) \{\widehat{\mu}_Y(S_i) - m_Y\}^2 - d_Y^2, \end{aligned}$$

where $d_Y = \widehat{\mu}_Y - m_Y$. Subtracting $n^{-1} \sum_{i=1}^n \widehat{w}(S_i) \widehat{V}_{Y,i}$ from both sides and using the definition of $\widehat{\sigma}^2$ gives

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \widehat{w}(S_i) \left[\{\widehat{\mu}_Y(S_i) - m_Y\}^2 - \widehat{V}_{Y,i} \right] - d_Y^2. \quad (73)$$

Apply Lemma 3 first with $Z_i = \widehat{\mu}_Y(S_i)$. Because $\mathbb{E}[u_i \widehat{\mu}_Y(S_i)] = m_Y$, its moment condition is $\mathbb{E}[u_i^2 \{\widehat{\mu}_Y(S_i) - m_Y\}^2] < \infty$, which follows from Assumption 7(i). Therefore, $d_Y = \widehat{\mu}_Y - m_Y = O_{\mathbb{P}}(n^{-1/2})$. Next apply the lemma with

$$Z_i = \{\widehat{\mu}_Y(S_i) - m_Y\}^2 - \widehat{V}_{Y,i}.$$

Lemma 1 gives $\mathbb{E}[u_i Z_i] = \sigma^2$, and Assumption 7(ii) is the required second-moment condition. Hence

$$\sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^n \widehat{w}(S_i) Z_i - \sigma^2 \right\} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{\sigma^2, i} + o_{\mathbb{P}}(1).$$

Combining this display with (73) and $\sqrt{n} d_Y^2 = o_{\mathbb{P}}(1)$ proves (43). The stated normal limit then follows from (43), Assumption 1(i), Assumption 7(ii), and the i.i.d. central limit theorem and Slutsky's theorem.

To prove consistency of $\widehat{V}_{\sigma^2}$, write

$$d_{\sigma} = \widehat{\sigma}^2 - \sigma^2, \quad A_i = \widehat{\mu}_Y(S_i) - m_Y.$$

Using (37) and (34),

$$\begin{aligned}
\widehat{\psi}_{\sigma^2,i} - \psi_{\sigma^2,i} &= \frac{u_i}{\bar{u}} \left[\{\widehat{\mu}_Y(S_i) - \widehat{\mu}_Y\}^2 - \widehat{V}_{Y,i} - \sigma^2 \right] \\
&\quad - \left[u_i \left[\{\widehat{\mu}_Y(S_i) - m_Y\}^2 - \widehat{V}_{Y,i} - \sigma^2 \right] \right] \\
&= \frac{u_i}{\bar{u}} \left[(A_i - d_Y)^2 - \widehat{V}_{Y,i} - (\sigma^2 + d_\sigma) \right] - u_i \left[A_i^2 - \widehat{V}_{Y,i} - \sigma^2 \right] \\
&= \frac{u_i}{\bar{u}} \left[A_i^2 - 2d_Y A_i + d_Y^2 - \widehat{V}_{Y,i} - \sigma^2 - d_\sigma \right] - u_i \left[A_i^2 - \widehat{V}_{Y,i} - \sigma^2 \right] \\
&= \left(\frac{u_i}{\bar{u}} - u_i \right) \left[A_i^2 - \widehat{V}_{Y,i} - \sigma^2 \right] + \frac{u_i}{\bar{u}} \left[-2d_Y A_i + d_Y^2 - d_\sigma \right] \\
&= \left(\frac{1}{\bar{u}} - 1 \right) \psi_{\sigma^2,i} + \frac{u_i}{\bar{u}} \left[-2d_Y A_i + d_Y^2 - d_\sigma \right].
\end{aligned} \tag{74}$$

Let $r_i = \widehat{\psi}_{\sigma^2,i} - \psi_{\sigma^2,i}$. From $(a_1 + a_2 + a_3 + a_4)^2 \leq 4 \sum_{j=1}^4 a_j^2$ and (74),

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n r_i^2 &\leq 4 \left(\frac{1}{\bar{u}} - 1 \right)^2 \frac{1}{n} \sum_{i=1}^n \psi_{\sigma^2,i}^2 + \frac{16d_Y^2}{\bar{u}^2} \frac{1}{n} \sum_{i=1}^n u_i^2 A_i^2 \\
&\quad + \frac{4d_Y^4}{\bar{u}^2} \frac{1}{n} \sum_{i=1}^n u_i^2 + \frac{4d_\sigma^2}{\bar{u}^2} \frac{1}{n} \sum_{i=1}^n u_i^2.
\end{aligned}$$

Assumption 7(i) and the weak law of large numbers imply

$$\frac{1}{n} \sum_{i=1}^n u_i^2 A_i^2 \xrightarrow{\mathbb{P}} \mathbb{E}[u_i^2 A_i^2] < \infty, \quad \frac{1}{n} \sum_{i=1}^n u_i^2 \xrightarrow{\mathbb{P}} \mathbb{E}[u_i^2] < \infty.$$

Moreover, $d_Y = o_{\mathbb{P}}(1)$ by Assumption 7(i) and the weak law of large numbers and $d_\sigma = o_{\mathbb{P}}(1)$ by (43). Also, $\bar{u} \xrightarrow{\mathbb{P}} 1$ and $n^{-1} \sum_{i=1}^n \psi_{\sigma^2,i}^2 = O_{\mathbb{P}}(1)$ by the weak law of large numbers and Assumption 7(i) and Assumption 7(ii), respectively. Consequently, every term on the right-hand side of the preceding bound is $o_{\mathbb{P}}(1)$, and hence

$$\frac{1}{n} \sum_{i=1}^n \{\widehat{\psi}_{\sigma^2,i} - \psi_{\sigma^2,i}\}^2 = o_{\mathbb{P}}(1). \tag{75}$$

Next,

$$\begin{aligned}
&\left| \frac{1}{n} \sum_{i=1}^n \widehat{\psi}_{\sigma^2,i}^2 - \frac{1}{n} \sum_{i=1}^n \psi_{\sigma^2,i}^2 \right| \\
&= \left| \frac{1}{n} \sum_{i=1}^n r_i \{2\psi_{\sigma^2,i} + r_i\} \right| \\
&\leq 2 \left(\frac{1}{n} \sum_{i=1}^n r_i^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n \psi_{\sigma^2,i}^2 \right)^{1/2} + \frac{1}{n} \sum_{i=1}^n r_i^2,
\end{aligned}$$

where the inequality follows from the triangle inequality and Cauchy-Schwarz. By Assumption 7(ii) and the weak law of large numbers,

$$\frac{1}{n} \sum_{i=1}^n \psi_{\sigma^2,i}^2 \xrightarrow{\mathbb{P}} \mathbb{E}[\psi_{\sigma^2,i}^2] < \infty.$$

Together with (75), this proves that

$$\frac{1}{n} \sum_{i=1}^n \widehat{\psi}_{\sigma^2,i}^2 - \frac{1}{n} \sum_{i=1}^n \psi_{\sigma^2,i}^2 = o_{\mathbb{P}}(1).$$

Therefore, $\widehat{V}_{\sigma^2} \xrightarrow{\mathbb{P}} V_{\sigma^2}$.

C.3 Proof of Theorem 2

For brevity, let $u_i = w(S_i)$ and $\bar{u} = n^{-1} \sum_{i=1}^n u_i$. Then, $\widehat{w}(S_i) = u_i/\bar{u}$. Let

$$\begin{aligned} \widetilde{\text{Var}}(\mu_X(S)) &= \frac{1}{n} \sum_{i=1}^n u_i \left[\{\widehat{\mu}_X(S_i) - \mathbb{E}_w[\mu_X(S)]\} \{\widehat{\mu}_X(S_i) - \mathbb{E}_w[\mu_X(S)]\}^{\top} - \widehat{V}_{X,i} \right], \\ \widetilde{\text{Cov}}(\mu_X(S), \mu_Y(S)) &= \frac{1}{n} \sum_{i=1}^n u_i \left[\{\widehat{\mu}_X(S_i) - \mathbb{E}_w[\mu_X(S)]\} \{\widehat{\mu}_Y(S_i) - \mathbb{E}_w[\mu_Y(S)]\} - \widehat{C}_{XY,i} \right]. \end{aligned}$$

Write

$$\begin{aligned} d_X &= \bar{\mu}_X - \mathbb{E}_w[\mu_X(S)], \quad d_Y = \bar{\mu}_Y - \mathbb{E}_w[\mu_Y(S)], \\ R_X &= \frac{1}{n} \sum_{i=1}^n u_i \{\widehat{\mu}_X(S_i) - \mathbb{E}_w[\mu_X(S)]\}, \quad R_Y = \frac{1}{n} \sum_{i=1}^n u_i \{\widehat{\mu}_Y(S_i) - \mathbb{E}_w[\mu_Y(S)]\}. \end{aligned}$$

Then $d_X = R_X/\bar{u}$ and $d_Y = R_Y/\bar{u}$. Expanding the centered products, one has

$$\widehat{\text{Var}}(\mu_X(S)) = \frac{1}{\bar{u}} \widetilde{\text{Var}}(\mu_X(S)) - d_X d_X^{\top}, \quad (76)$$

$$\widehat{\text{Cov}}(\mu_X(S), \mu_Y(S)) = \frac{1}{\bar{u}} \widetilde{\text{Cov}}(\mu_X(S), \mu_Y(S)) - d_X d_Y. \quad (77)$$

Applying Lemma 3 with the stacked vector $Z_i = (\widehat{\mu}_X(S_i)^{\top}, \widehat{\mu}_Y(S_i)^{\top})^{\top}$, Assumption 1(ii) gives

$$\mathbb{E}[u_i Z_i] = m_Z, \quad m_Z := (\mathbb{E}_w[\mu_X(S)]^{\top}, \mathbb{E}_w[\mu_Y(S)]^{\top})^{\top},$$

while Assumption 8(i) implies the lemma's moment condition

$$\mathbb{E}[\|u_i(Z_i - m_Z)\|^2] < \infty.$$

Therefore, the lemma gives

$$d_X = O_{\mathbb{P}}(n^{-1/2}), \quad d_Y = O_{\mathbb{P}}(n^{-1/2}).$$

We first establish consistency of $\widehat{\text{Var}}(\mu_X(S))$. Let

$$Z_i^{XX} = \text{vec} \left(\{\widehat{\mu}_X(S_i) - \mathbb{E}_w[\mu_X(S)]\} \{\widehat{\mu}_X(S_i) - \mathbb{E}_w[\mu_X(S)]\}^{\top} - \widehat{V}_{X,i} \right),$$

where $\text{vec}(A)$ stacks the columns of A into a column vector. Fix $(j, k) \in \{1, \dots, K\}^2$ and write $m_{X,j} = \mathbb{E}_w[\mu_{X,j}(S)]$. The (j, k) entry of $\widetilde{\text{Var}}(\mu_X(S))$ is the sample average of

$$Z_{i,jk} = u_i \left[\{\widehat{\mu}_{X,j}(S_i) - m_{X,j}\} \{\widehat{\mu}_{X,k}(S_i) - m_{X,k}\} - \widehat{V}_{X,i,jk} \right].$$

This random variable is integrable. Indeed,

$$\begin{aligned} |Z_{i,jk}| &\leq u_i \left[|\widehat{\mu}_{X,j}(S_i) - m_{X,j}| |\widehat{\mu}_{X,k}(S_i) - m_{X,k}| + |\widehat{V}_{X,i,jk}| \right] \\ &\leq \frac{u_i}{2} \{\widehat{\mu}_{X,j}(S_i) - m_{X,j}\}^2 + \frac{u_i}{2} \{\widehat{\mu}_{X,k}(S_i) - m_{X,k}\}^2 + u_i |\widehat{V}_{X,i,jk}|, \end{aligned}$$

where the second inequality uses $|ab| \leq (a^2 + b^2)/2$. The expectation of the right-hand side is finite by Assumption 8(i) and Assumption 8(ii). Therefore, the weak law of large numbers and Lemma 1 imply, for every (j, k) ,

$$\widetilde{\text{Var}}(\mu_X(S))_{jk} \xrightarrow{\mathbb{P}} \text{Var}_w(\mu_X(S))_{jk}.$$

Equivalently,

$$\text{vec}\{\widetilde{\text{Var}}(\mu_X(S))\} \xrightarrow{\mathbb{P}} \text{vec}\{\text{Var}_w(\mu_X(S))\}.$$

Because K is fixed, this implies convergence of the whole matrix. Moreover, $\bar{u}^{-1} = 1 + o_{\mathbb{P}}(1)$ and each entry of $d_X d_X^\top$ is $O_{\mathbb{P}}(n^{-1})$. Hence

$$\widehat{\text{Var}}(\mu_X(S)) \xrightarrow{\mathbb{P}} \text{Var}_w(\mu_X(S)). \quad (78)$$

Assumption 8(iii), continuity of the determinant, and (78) imply that $\widehat{\text{Var}}(\mu_X(S))$ is nonsingular with probability approaching one. Continuity of matrix inversion at nonsingular matrices then gives

$$\widehat{\text{Var}}(\mu_X(S))^{-1} \xrightarrow{\mathbb{P}} \text{Var}_w(\mu_X(S))^{-1}.$$

Using the definition of β , we have the exact identity

$$\widehat{\beta} - \beta = \widehat{\text{Var}}(\mu_X(S))^{-1} \left[\widehat{\text{Cov}}(\mu_X(S), \mu_Y(S)) - \widehat{\text{Var}}(\mu_X(S))\beta \right]. \quad (79)$$

Combining (76), (77), and (79), and using the $O_{\mathbb{P}}(n^{-1/2})$ rates for d_X and d_Y , gives

$$\begin{aligned} \sqrt{n}(\widehat{\beta} - \beta) &= \widehat{\text{Var}}(\mu_X(S))^{-1} \frac{1}{\bar{u}} \frac{1}{\sqrt{n}} \sum_{i=1}^n u_i \left[\{\widehat{\mu}_X(S_i) - \mathbb{E}_w[\mu_X(S)]\} \right. \\ &\quad \times \left. \{\widehat{\mu}_Y(S_i) - \mathbb{E}_w[\mu_Y(S)] - \{\widehat{\mu}_X(S_i) - \mathbb{E}_w[\mu_X(S)]\}^\top \beta\} \right. \\ &\quad \left. - \widehat{C}_{XY,i} + \widehat{V}_{X,i}\beta \right] + o_{\mathbb{P}}(1). \end{aligned} \quad (80)$$

Let H_i denote the bracketed vector in (80). Lemma 1 and the definition of β give $\mathbb{E}[u_i H_i] = 0$. Moreover, Assumption 8(iv) ensures the moment condition of Lemma 3. Applying that lemma with $Z_i = H_i$ gives

$$\frac{1}{\bar{u}} \frac{1}{\sqrt{n}} \sum_{i=1}^n u_i H_i = \frac{1}{\sqrt{n}} \sum_{i=1}^n u_i H_i + o_{\mathbb{P}}(1).$$

We also have

$$\widehat{\text{Var}}(\mu_X(S))^{-1} = \text{Var}_w(\mu_X(S))^{-1} + o_{\mathbb{P}}(1),$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{\beta,i} = O_{\mathbb{P}}(1),$$

where the last display follows from the multivariate central limit theorem and Assumption 8(iv). It follows from (80), the preceding displays, and Slutsky's theorem that

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{\beta,i} + o_{\mathbb{P}}(1),$$

establishing (44). The stated normal limit then follows from (44), Assumption 1(i), Assumption 8(iv), the multivariate i.i.d. central limit theorem and Slutsky's theorem.

Now suppose that Assumption 10 also holds. We prove consistency of $\hat{V}_{\beta}$. For brevity, let

$$m_X = \mathbb{E}_w[\mu_X(S)], \quad m_Y = \mathbb{E}_w[\mu_Y(S)], \quad M = \text{Var}_w(\mu_X(S)),$$

and define

$$A_i = \hat{\mu}_X(S_i) - m_X, \quad B_i = \hat{\mu}_Y(S_i) - m_Y, \quad d_{\beta} = \hat{\beta} - \beta, \quad \widehat{M} = \widehat{\text{Var}}(\mu_X(S)).$$

Write

$$\begin{aligned} q_i &= u_i \left\{ A_i(B_i - A_i^{\top} \beta) - \widehat{C}_{XY,i} + \widehat{V}_{X,i} \beta \right\}, \\ \widehat{q}_i &= \frac{u_i}{\widehat{u}} \left[(A_i - d_X) \{ B_i - d_Y - (A_i - d_X)^{\top} \widehat{\beta} \} - \widehat{C}_{XY,i} + \widehat{V}_{X,i} \widehat{\beta} \right]. \end{aligned}$$

Then $\psi_{\beta,i} = M^{-1} q_i$ and $\widehat{\psi}_{\beta,i} = \widehat{M}^{-1} \widehat{q}_i$. Expanding and collecting terms gives

$$\begin{aligned} \widehat{q}_i - q_i &= \left(\frac{1}{\widehat{u}} - 1 \right) q_i + \frac{u_i}{\widehat{u}} \left[-d_X B_i - d_Y A_i + d_X d_Y + A_i d_X^{\top} \beta + d_X A_i^{\top} \beta - d_X d_X^{\top} \beta \right. \\ &\quad \left. + \{ \widehat{V}_{X,i} - (A_i - d_X)(A_i - d_X)^{\top} \} d_{\beta} \right]. \end{aligned} \quad (81)$$

We show componentwise that the empirical mean squared difference between $\widehat{q}_i$ and q_i converges to zero. Fix $j \in \{1, \dots, K\}$. The j th component of (81) equals

$$\begin{aligned} \widehat{q}_{i,j} - q_{i,j} &= \left(\frac{1}{\widehat{u}} - 1 \right) q_{i,j} + \frac{u_i}{\widehat{u}} \left[-d_{X,j} B_i - d_Y A_{i,j} + d_{X,j} d_Y + A_{i,j} d_X^{\top} \beta + d_{X,j} A_i^{\top} \beta - d_{X,j} d_X^{\top} \beta \right. \\ &\quad \left. + \sum_{k=1}^K \left\{ \widehat{V}_{X,i,jk} - (A_{i,j} - d_{X,j})(A_{i,k} - d_{X,k}) \right\} d_{\beta,k} \right]. \end{aligned}$$

Since K is fixed, repeated use of $(a_1 + \dots + a_J)^2 \leq J \sum_{\ell=1}^J a_{\ell}^2$ gives a finite constant C , independent of n , such that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (\widehat{q}_{i,j} - q_{i,j})^2 &\leq C \left[\left(\frac{1}{\widehat{u}} - 1 \right)^2 \frac{1}{n} \sum_{i=1}^n q_{i,j}^2 \right. \\ &\quad + \frac{1}{\widehat{u}^2} \left[d_{X,j}^2 \frac{1}{n} \sum_{i=1}^n u_i^2 B_i^2 + d_Y^2 \frac{1}{n} \sum_{i=1}^n u_i^2 A_{i,j}^2 + d_{X,j}^2 d_Y^2 \frac{1}{n} \sum_{i=1}^n u_i^2 \right] \\ &\quad + \frac{1}{\widehat{u}^2} \left[(d_X^{\top} \beta)^2 \frac{1}{n} \sum_{i=1}^n u_i^2 A_{i,j}^2 + d_{X,j}^2 \frac{1}{n} \sum_{i=1}^n u_i^2 (A_i^{\top} \beta)^2 + d_{X,j}^2 (d_X^{\top} \beta)^2 \frac{1}{n} \sum_{i=1}^n u_i^2 \right] \\ &\quad \left. + \frac{1}{\widehat{u}^2} \sum_{k=1}^K d_{\beta,k}^2 \frac{1}{n} \sum_{i=1}^n u_i^2 \left\{ \widehat{V}_{X,i,jk} - (A_{i,j} - d_{X,j})(A_{i,k} - d_{X,k}) \right\}^2 \right]. \end{aligned}$$

Assumption 8(i) and the weak law of large numbers imply

$$\frac{1}{n} \sum_{i=1}^n u_i^2 B_i^2 = O_{\mathbb{P}}(1), \quad \frac{1}{n} \sum_{i=1}^n u_i^2 A_{i,j}^2 = O_{\mathbb{P}}(1), \quad \frac{1}{n} \sum_{i=1}^n u_i^2 = O_{\mathbb{P}}(1) \quad \frac{1}{n} \sum_{i=1}^n u_i^2 (A_i^\top \beta)^2 = O_{\mathbb{P}}(1).$$

For every j, k , the inequality $(a - b)^2 \leq 2a^2 + 2b^2$ yields

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n u_i^2 \left\{ \widehat{V}_{X,i,jk} - (A_{i,j} - d_{X,j})(A_{i,k} - d_{X,k}) \right\}^2 \\ & \leq 2 \frac{1}{n} \sum_{i=1}^n u_i^2 \widehat{V}_{X,i,jk}^2 + 2 \frac{1}{n} \sum_{i=1}^n u_i^2 (A_{i,j} - d_{X,j})^2 (A_{i,k} - d_{X,k})^2. \end{aligned}$$

The first term is $O_{\mathbb{P}}(1)$ by Assumption 10 and the weak law of large numbers. For the second term, using $(x - y)^2 \leq 2x^2 + 2y^2$ twice and $2uv \leq u^2 + v^2$ gives

$$\begin{aligned} & (A_{i,j} - d_{X,j})^2 (A_{i,k} - d_{X,k})^2 \\ & \leq 4A_{i,j}^2 A_{i,k}^2 + 4A_{i,j}^2 d_{X,k}^2 + 4d_{X,j}^2 A_{i,k}^2 + 4d_{X,j}^2 d_{X,k}^2 \\ & \leq 2A_{i,j}^4 + 2A_{i,k}^4 + 4A_{i,j}^2 d_{X,k}^2 + 4d_{X,j}^2 A_{i,k}^2 + 4d_{X,j}^2 d_{X,k}^2. \end{aligned}$$

Multiplying this inequality by u_i^2 , Assumption 10, the weak law of large numbers, and $d_X = O_{\mathbb{P}}(n^{-1/2})$ therefore imply that the second term is also $O_{\mathbb{P}}(1)$. Consequently, for every j, k ,

$$\frac{1}{n} \sum_{i=1}^n u_i^2 \left\{ \widehat{V}_{X,i,jk} - (A_{i,j} - d_{X,j})(A_{i,k} - d_{X,k}) \right\}^2 = O_{\mathbb{P}}(1).$$

Since $q_i = M\psi_{\beta,i}$, Assumption 8(iv) and the weak law of large numbers imply $n^{-1} \sum_{i=1}^n q_{i,j}^2 = O_{\mathbb{P}}(1)$. Moreover, $\bar{u} \xrightarrow{\mathbb{P}} 1$. Also, $d_X = O_{\mathbb{P}}(n^{-1/2})$, $d_Y = O_{\mathbb{P}}(n^{-1/2})$, and (44) implies $d_\beta = O_{\mathbb{P}}(n^{-1/2})$. Hence, for every j ,

$$\frac{1}{n} \sum_{i=1}^n (\widehat{q}_{i,j} - q_{i,j})^2 = o_{\mathbb{P}}(1). \quad (82)$$

Next fix $j \in \{1, \dots, K\}$. Since

$$\widehat{\psi}_{\beta,i} - \psi_{\beta,i} = \widehat{M}^{-1}(\widehat{q}_i - q_i) + (\widehat{M}^{-1} - M^{-1})q_i,$$

the j th component is a finite sum of scalar products. Entrywise convergence of $\widehat{M}^{-1}$ to M^{-1} , (82), and repeated use of $(a_1 + \dots + a_K)^2 \leq K \sum_{k=1}^K a_k^2$ imply

$$\frac{1}{n} \sum_{i=1}^n (\widehat{\psi}_{\beta,i,j} - \psi_{\beta,i,j})^2 = o_{\mathbb{P}}(1). \quad (83)$$

Finally, fix $(j, k) \in \{1, \dots, K\}^2$ and let $r_{i,j} = \widehat{\psi}_{\beta,i,j} - \psi_{\beta,i,j}$. Then

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \widehat{\psi}_{\beta,i,j} \widehat{\psi}_{\beta,i,k} - \frac{1}{n} \sum_{i=1}^n \psi_{\beta,i,j} \psi_{\beta,i,k} \\ & = \frac{1}{n} \sum_{i=1}^n r_{i,j} \psi_{\beta,i,k} + \frac{1}{n} \sum_{i=1}^n \psi_{\beta,i,j} r_{i,k} + \frac{1}{n} \sum_{i=1}^n r_{i,j} r_{i,k} \end{aligned} \quad (84)$$

$$= o_{\mathbb{P}}(1). \quad (85)$$

The second equality follows from the fact that by Cauchy–Schwarz, the absolute value of the first term is bounded by

$$\left\{ \frac{1}{n} \sum_{i=1}^n r_{i,j}^2 \right\}^{1/2} \left\{ \frac{1}{n} \sum_{i=1}^n \psi_{\beta,i,k}^2 \right\}^{1/2}.$$

By (83), $\frac{1}{n} \sum_{i=1}^n r_{i,j}^2 = o_{\mathbb{P}}(1)$, while Assumption 8(iv) and the weak law of large numbers imply

$$\frac{1}{n} \sum_{i=1}^n \psi_{\beta,i,j}^2 = O_{\mathbb{P}}(1) \quad \text{for every } j.$$

Thus, the first term in (84) is an $o_{\mathbb{P}}(1)$, and one can use similar steps to show that the two remaining terms are also $o_{\mathbb{P}}(1)$. Moreover, for each j, k , the weak law of large numbers gives

$$\frac{1}{n} \sum_{i=1}^n \psi_{\beta,i,j} \psi_{\beta,i,k} \xrightarrow{\mathbb{P}} \mathbb{E}[\psi_{\beta,i,j} \psi_{\beta,i,k}],$$

because $\mathbb{E}[|\psi_{\beta,i,j} \psi_{\beta,i,k}|] < \infty$ by Cauchy–Schwarz and Assumption 8(iv). Combined with (85), this proves that

$$\widehat{V}_{\beta} \xrightarrow{\mathbb{P}} V_{\beta}.$$

C.4 Proof of Theorem 3

Let

$$\begin{aligned} m_Y &= \mathbb{E}_w[\mu_Y(S)], & A_i &= \widehat{\mu}_Y(S_i) - m_Y, & d_Y &= \widehat{\mu}_Y - m_Y, \\ u_i &= w(S_i), & \bar{u} &= n^{-1} \sum_{i=1}^n u_i, & H_i &= A_i^3 - 3A_i \widehat{V}_{Y,i} + 3\widehat{C}_{YV,i} - \widehat{K}_{Y,i}. \end{aligned}$$

To derive the exact representation, first expand each centered term in (9):

$$\widehat{\kappa}_Y = \frac{1}{n} \sum_{i=1}^n \widehat{w}(S_i) \left[H_i - 3d_Y \{A_i^2 - \widehat{V}_{Y,i}\} + 3d_Y^2 A_i - d_Y^3 \right].$$

By the definition of d_Y ,

$$\frac{1}{n} \sum_{i=1}^n \widehat{w}(S_i) A_i = d_Y,$$

while (73) implies

$$\frac{1}{n} \sum_{i=1}^n \widehat{w}(S_i) (A_i^2 - \widehat{V}_{Y,i}) = \widehat{\sigma}^2 + d_Y^2.$$

Also, $n^{-1} \sum_i \widehat{w}(S_i) = 1$. Consequently,

$$\widehat{\kappa}_Y = \frac{1}{n} \sum_{i=1}^n \widehat{w}(S_i) H_i - 3d_Y(\widehat{\sigma}^2 + d_Y^2) + 3d_Y^3 - d_Y^3,$$

which simplifies to

$$\widehat{\kappa}_Y = \frac{1}{n} \sum_{i=1}^n \widehat{w}(S_i) H_i - 3d_Y \widehat{\sigma}^2 - d_Y^3. \tag{86}$$

We next derive the linearization. We start by rewriting (36) with the notation used in the proof:

$$\begin{aligned}\psi_{\kappa_Y, i} &= u_i \left[A_i^3 - 3A_i \widehat{V}_{Y, i} + 3\widehat{C}_{YV, i} - \widehat{K}_{Y, i} - \kappa_Y - 3\sigma^2 A_i \right] \\ &= u_i \{H_i - \kappa_Y - 3\sigma^2 A_i\}.\end{aligned}$$

Since $d_Y = \frac{n^{-1} \sum_i u_i A_i}{\bar{u}}$,

$$\frac{1}{\bar{u}} \frac{1}{n} \sum_{i=1}^n \psi_{\kappa_Y, i} = \frac{1}{\bar{u}} \frac{1}{n} \sum_{i=1}^n u_i (H_i - \kappa_Y) - 3\sigma^2 d_Y.$$

Subtracting κ_Y from (86) and using this equality,

$$\begin{aligned}\widehat{\kappa}_Y - \kappa_Y &= \frac{1}{\bar{u}} \frac{1}{n} \sum_{i=1}^n \psi_{\kappa_Y, i} + 3\sigma^2 d_Y - 3d_Y \widehat{\sigma}^2 - d_Y^3 \\ &= \frac{1}{\bar{u}} \frac{1}{n} \sum_{i=1}^n \psi_{\kappa_Y, i} - 3d_Y (\widehat{\sigma}^2 - \sigma^2) - d_Y^3.\end{aligned}$$

Multiplication by $\sqrt{n}$ yields

$$\sqrt{n}(\widehat{\kappa}_Y - \kappa_Y) = \frac{1}{\bar{u}} \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{\kappa_Y, i} - 3\sqrt{n} d_Y (\widehat{\sigma}^2 - \sigma^2) - \sqrt{n} d_Y^3. \quad (87)$$

Assumption 9(i) implies $\bar{u} - 1 = O_{\mathbb{P}}(n^{-1/2})$ and $d_Y = O_{\mathbb{P}}(n^{-1/2})$. Lemma 1, Assumptions 9(i)–(ii), the weak law of large numbers, and (73) imply $\widehat{\sigma}^2 - \sigma^2 = o_{\mathbb{P}}(1)$. Assumption 9(iii) and the central limit theorem give $n^{-1/2} \sum_i \psi_{\kappa_Y, i} = O_{\mathbb{P}}(1)$. Thus (87) proves (46); its stated normal limit follows from the i.i.d. central limit theorem and Slutsky's theorem.

C.5 Proof of Proposition 1

By Lemma 2, conditional on S_i , each arm contains an i.i.d. sample from its arm-specific conditional marginal distribution, and the samples are independent across arms. The usual properties of an i.i.d. sample therefore apply separately within each arm. In particular, its sample mean is unbiased and its sample covariance matrix is unbiased for the population covariance matrix. Dividing the latter by $N_{d, i}$ and stacking the arms gives

$$\begin{aligned}\mathbb{E}[\widehat{\mu}_O(S_i) \mid S_i] &= \mu_O(S_i), \\ \mathbb{E}[\widehat{V}_{O, i} \mid S_i] &= \text{Var}(\widehat{\mu}_O(S_i) \mid S_i),\end{aligned}$$

where the off-diagonal arm blocks of $\widehat{V}_{O, i}$ and $\text{Var}(\widehat{\mu}_O(S_i) \mid S_i)$ are zero by cross-arm independence. This proves (10) and (11).

Then, consider an i.i.d. scalar sample $Z_1, \dots, Z_m$, let $\mu = \mathbb{E}[Z_j]$, and define $\kappa_3 = \mathbb{E}[(Z_j - \mu)^3]$. One has

$$\mathbb{E}[(\bar{Z} - \mu)^3] = \frac{\kappa_3}{m^2}, \quad (88)$$

$$\mathbb{E} \left[\frac{1}{m} \sum_{j=1}^m (Z_j - \bar{Z})^3 \right] = \frac{(m-1)(m-2)}{m^2} \kappa_3. \quad (89)$$

The first identity follows by expanding the cube of the sample mean: all summands with unequal observation indices have expectation zero, leaving $m\kappa_3/m^3$. The second follows by expanding the cube and collecting the terms with coincident indices. Now apply (89) within arm d , conditionally on S_i , to the scalar variable $Z_j = c_d^\top \mathbf{O}_{j,i}(d)$, with $m = N_{d,i}$. By Lemma 2, these variables are i.i.d. within arms and the samples are independent across arms. Therefore,

$$\mathbb{E}[\widehat{M}_{3,O,i}(c) \mid S_i] = \sum_{d=0}^{\mathcal{D}} \frac{1}{N_{d,i}^2} \mathbb{E}[\{c_d^\top (\mathbf{O}_i(d) - \mathbb{E}[\mathbf{O}_i(d) \mid S_i])\}^3 \mid S_i].$$

Applying (88), and using conditional independence across arms from Lemma 2, it follows that the right-hand side of the last display equals

$$\mathbb{E}[\{c^\top (\widehat{\mu}_O(S_i) - \mu_O(S_i))\}^3 \mid S_i],$$

which proves (12) for every non-random c .

It remains to verify the statements for the linear transformations. The first two identities of the proposition, premultiplied and postmultiplied by B and/or c , give the conditional-unbiasedness restrictions for $\widehat{\mu}_X(S_i)$, $\widehat{\mu}_Y(S_i)$, $\widehat{V}_{X,i}$, $\widehat{V}_{Y,i}$, and $\widehat{C}_{XY,i}$. Moreover, (12) and the definition of $\widehat{K}_{Y,i}$ yield

$$\mathbb{E}[\widehat{K}_{Y,i} \mid S_i] = \mathbb{E}[\{\widehat{\mu}_Y(S_i) - \mu_Y(S_i)\}^3 \mid S_i].$$

Finally, for an i.i.d. scalar sample $Z_1, \dots, Z_m$,

$$\text{Cov}\left(\bar{Z}, \frac{1}{m(m-1)} \sum_{j=1}^m (Z_j - \bar{Z})^2\right) = \frac{\kappa_3}{m^2}.$$

Applying the previous display to $Z_j = c_d^\top \mathbf{O}_{j,i}(d)$ shows that

$$\text{Cov}(\widehat{\mu}_Y(S_i), \widehat{V}_{Y,i} \mid S_i) = \mathbb{E}[\widehat{M}_{3,O,i}(c) \mid S_i] = \mathbb{E}[\widehat{C}_{YV,i} \mid S_i].$$

Thus all the conditional-unbiasedness restrictions in Assumption 1 hold. The i.i.d. requirement follows because all the displayed statistics are fixed functions of the i.i.d. site vectors and of randomizations conducted independently across sites.

C.6 Proof of Proposition 2

For brevity, write $A = \mu_X(S)$, $I = \text{IDE}(S)$, and $G = \text{DE}(S)$. By (13), $\mu_Y(S) = AI + G$. Assumption 5 therefore implies

$$\begin{aligned} \text{Cov}_w(A, \mu_Y(S)) &= \text{Cov}_w(A, AI + G) \\ &= \mathbb{E}_w[I] \{\mathbb{E}_w[A^2] - \mathbb{E}_w[A]^2\} \\ &= \mathbb{E}_w[I] \text{Var}_w(A). \end{aligned}$$

Dividing by $\text{Var}_w(A)$ proves that $\beta_{\text{med}} = \mathbb{E}_w[I]$. The regression intercept is

$$\begin{aligned} \alpha_{\text{med}} &= \mathbb{E}_w[\mu_Y(S)] - \beta_{\text{med}} \mathbb{E}_w[A] \\ &= \mathbb{E}_w[A] \mathbb{E}_w[I] + \mathbb{E}_w[G] - \mathbb{E}_w[I] \mathbb{E}_w[A] = \mathbb{E}_w[G], \end{aligned}$$

which proves the result.

C.7 Proof of Proposition 4

By (18),

$$\mu_{\Delta_2}(S_i) - \mu_{\Delta_1}(S_i) = \epsilon_2^S(S_i) - \epsilon_1^S(S_i) + \epsilon_2^P(S_i, P_2(S_i)) - \epsilon_1^P(S_i, P_1(S_i)),$$

which proves Part 1.a). For Part 1.b), denote by A_i and B_i the first and the second difference on the right-hand side of the previous display, respectively. If $\text{Var}_w(A_i + B_i) = 0$ and A_i and B_i are not perfectly negatively correlated, then $\text{Var}_w(A_i) = \text{Var}_w(B_i) = 0$. Moreover, if $\text{Var}_w(B_i) = 0$ and $\epsilon_2^P(S_i, P_2(S_i))$ and $\epsilon_1^P(S_i, P_1(S_i))$ are not perfectly positively correlated, then $\text{Var}_w(\epsilon_2^P(S_i, P_2(S_i))) = \text{Var}_w(\epsilon_1^P(S_i, P_1(S_i))) = 0$. This proves necessity; sufficiency follows immediately by adding variables with zero weighted variance.

For Part 2, let $v = \text{Var}_w(\mu_{\Delta_1}(S_i)) = \text{Var}_w(\mu_{\Delta_2}(S_i)) > 0$. Then

$$\sigma_{\Delta}^2 = 2v - 2 \text{Cov}_w(\mu_{\Delta_1}(S_i), \mu_{\Delta_2}(S_i)), \quad \beta_{21} = \frac{\text{Cov}_w(\mu_{\Delta_1}(S_i), \mu_{\Delta_2}(S_i))}{v}.$$

Consequently, $\sigma_{\Delta}^2 = 0$ if and only if $\text{Cov}_w(\mu_{\Delta_1}(S_i), \mu_{\Delta_2}(S_i)) = v$, which holds if and only if $\beta_{21} = 1$.

C.8 Proof of Proposition 5

For brevity, let $F = \mu_{\text{FS}}(S)$, $T = \mu_{\text{LATE}}(S)$, $l_1 = \text{LATE}$, and $l_2 = \text{LATE}_{\text{FS}^2}$. Then $\mu_{\text{ITT}}(S) = FT$, $\text{FS} = \mathbb{E}_w[F]$, and $m_2 = \mathbb{E}_w[F^2]$.

For Point 1,

$$\mathbb{E}_w[\{\mu_{\text{ITT}}(S) - l_2 F\}^2] = \mathbb{E}_w[F^2(T - l_2)^2] = m_2 \sigma_{\text{LATE}, \text{FS}^2}^2.$$

Dividing by $m_2 > 0$ proves the equality.

For Point 2, Assumption 6 implies $0 < F \leq 1$ almost surely. Let $X = F(T - l_1)$, $h = \mathbb{E}_w[X^2]$, $d = \mathbb{E}_w[FX] = r - m_2 l_1$, and $\delta = \text{FS} - m_2 > 0$. Since $\mathbb{E}_w[X] = 0$, $\mathbb{E}_w[(1 - F)X] = -d$. Also,

$$\text{FS} \sigma_{\text{LATE}}^2 = \mathbb{E}_w[X^2/F] = h + \mathbb{E}_w[(1 - F)X^2/F].$$

If the target variance is infinite, the bound is immediate. Otherwise, Cauchy–Schwarz gives

$$d^2 = \left\{ \mathbb{E}_w \left[\sqrt{\frac{1-F}{F}} X \sqrt{F(1-F)} \right] \right\}^2 \leq \mathbb{E}_w[(1-F)X^2/F] \delta.$$

Consequently, $\sigma_{\text{LATE}}^2 \geq (h + d^2/\delta)/\text{FS}$. Moreover, the definition of l_2 gives $\mathbb{E}_w[F^2(T - l_2)] = 0$, so

$$\mathbb{E}_w[F^2(T - l_1)^2] = \mathbb{E}_w[F^2(T - l_2)^2] + m_2(l_2 - l_1)^2.$$

Combining this identity with $d = m_2(l_2 - l_1)$ proves the final equality in (23).

C.9 Proof of Theorem 4

Point 1. We first establish consistency and root- n convergence of $\widehat{\text{LATE}}_{\text{FS}^2}$. Conditional unbiasedness gives

$$\mathbb{E}_w[\widehat{\mu}_{\text{FS}}(S_i) \widehat{\mu}_{\text{ITT}}(S_i) - \widehat{C}_{\text{FS}, \text{ITT}, i}] = \mathbb{E}_w[\mu_{\text{FS}}(S) \mu_{\text{ITT}}(S)]$$

and

$$\mathbb{E}_w[\widehat{\mu}_{\text{FS}}(S_i)^2 - \widehat{V}_{\text{FS},i}] = \mathbb{E}_w[\mu_{\text{FS}}(S)^2].$$

The normalized-weight expansion and the second-moment conditions therefore imply that $\widehat{r}$ and $\widehat{m}_2$ are root- n consistent for these two expectations. Since $\mathbb{E}_w[\mu_{\text{FS}}(S)^2] > 0$, a ratio expansion yields

$$\widehat{\text{LATE}}_{\text{FS}^2} - \text{LATE}_{\text{FS}^2} = O_{\mathbb{P}}(n^{-1/2}). \quad (90)$$

We next expand the numerator of $\widehat{\sigma}_{\text{LATE},\text{FS}^2}^2$. By the unbiasedness of the variance and covariance corrections,

$$\mathbb{E}[g_i(l) \mid S_i] = \{\mu_{\text{ITT}}(S_i) - l\mu_{\text{FS}}(S_i)\}^2.$$

At $l = \text{LATE}_{\text{FS}^2}$, this implies $\mathbb{E}_w[g_i(\text{LATE}_{\text{FS}^2})] = q$. Assumption 11 allows us to apply Lemma 3 with $Z_i = g_i(\text{LATE}_{\text{FS}^2})$:

$$\sqrt{n}\{\widehat{Q}(\text{LATE}_{\text{FS}^2}) - q\} = \frac{1}{\sqrt{n}} \sum_{i=1}^n w(S_i)\{g_i(\text{LATE}_{\text{FS}^2}) - q\} + o_{\mathbb{P}}(1). \quad (91)$$

Differentiating $g_i(l)$ gives

$$\frac{\partial g_i(l)}{\partial l} = -2\widehat{\mu}_{\text{FS}}(S_i)\{\widehat{\mu}_{\text{ITT}}(S_i) - l\widehat{\mu}_{\text{FS}}(S_i)\} - 2l\widehat{V}_{\text{FS},i} + 2\widehat{C}_{\text{FS},\text{ITT},i}.$$

The first and third finite-second-moment conditions in Assumption 11 and the preceding display imply

$$\mathbb{E}[|w(S_i)g'_i(\text{LATE}_{\text{FS}^2})|] < \infty.$$

The conditional unbiasedness identities in Proposition 1 therefore imply

$$\begin{aligned} \mathbb{E}_w\left[\frac{\partial g_i(\text{LATE}_{\text{FS}^2})}{\partial l}\right] &= -2\mathbb{E}_w[\mu_{\text{FS}}(S)\{\mu_{\text{ITT}}(S) - \text{LATE}_{\text{FS}^2}\mu_{\text{FS}}(S)\}] \\ &= 0, \end{aligned} \quad (92)$$

Therefore, the weak law of large numbers and (92) give

$$\widehat{Q}'(\text{LATE}_{\text{FS}^2}) \xrightarrow{\mathbb{P}} 0. \quad (93)$$

Also, direct differentiation shows that

$$\widehat{Q}''(l) = 2\frac{1}{n} \sum_{i=1}^n \widehat{w}(S_i) \left\{ \widehat{\mu}_{\text{FS}}(S_i)^2 - \widehat{V}_{\text{FS},i} \right\},$$

which does not depend on l . The weak law of large numbers, Assumption 11, and unbiasedness of $\widehat{V}_{\text{FS},i}$ imply that this expression converges in probability to $2\mathbb{E}_w[\mu_{\text{FS}}(S)^2]$ and hence is $O_{\mathbb{P}}(1)$. The exact second-order Taylor identity for the quadratic function $\widehat{Q}(l)$ is

$$\begin{aligned} \widehat{q} &= \widehat{Q}(\text{LATE}_{\text{FS}^2}) + \widehat{Q}'(\text{LATE}_{\text{FS}^2})(\widehat{\text{LATE}}_{\text{FS}^2} - \text{LATE}_{\text{FS}^2}) \\ &\quad + \frac{1}{2}\widehat{Q}''(\text{LATE}_{\text{FS}^2})(\widehat{\text{LATE}}_{\text{FS}^2} - \text{LATE}_{\text{FS}^2})^2. \end{aligned}$$

By (93) and (90), $\widehat{Q}'(\text{LATE}_{\text{FS}^2})\sqrt{n}(\widehat{\text{LATE}}_{\text{FS}^2} - \text{LATE}_{\text{FS}^2}) = o_{\mathbb{P}}(1)$. Similarly,

$$\frac{1}{2}\widehat{Q}''(\text{LATE}_{\text{FS}^2})\sqrt{n}(\widehat{\text{LATE}}_{\text{FS}^2} - \text{LATE}_{\text{FS}^2})^2 = \frac{1}{2}\widehat{Q}''(\text{LATE}_{\text{FS}^2})n^{-1/2}(\sqrt{n}(\widehat{\text{LATE}}_{\text{FS}^2} - \text{LATE}_{\text{FS}^2}))^2 = o_{\mathbb{P}}(1),$$

because $\widehat{Q}''(\text{LATE}_{\text{FS}^2}) = O_{\mathbb{P}}(1)$ and in view of (90). Therefore,

$$\begin{aligned}\sqrt{n}(\widehat{q} - q) &= \sqrt{n}\{\widehat{Q}(\text{LATE}_{\text{FS}^2}) - q\} + o_{\mathbb{P}}(1) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{q,i} + o_{\mathbb{P}}(1),\end{aligned}\tag{94}$$

where the last equality follows from (91) and the definition in (47).

For the denominator, unbiasedness of $\widehat{V}_{\text{FS},i}$ implies

$$\mathbb{E}_w[\widehat{\mu}_{\text{FS}}(S_i)^2 - \widehat{V}_{\text{FS},i}] = \mathbb{E}_w[\mu_{\text{FS}}(S)^2].$$

Assumption 11 allows us to apply Lemma 3 with $Z_i = \widehat{\mu}_{\text{FS}}(S_i)^2 - \widehat{V}_{\text{FS},i}$. Therefore,

$$\begin{aligned}\sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^n \widehat{w}(S_i) [\widehat{\mu}_{\text{FS}}(S_i)^2 - \widehat{V}_{\text{FS},i}] - \mathbb{E}_w[\mu_{\text{FS}}(S)^2] \right\} \\ = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{\text{FS}^2,i} + o_{\mathbb{P}}(1).\end{aligned}\tag{95}$$

By Proposition 5, $q = \mathbb{E}_w[\mu_{\text{FS}}(S)^2]\sigma_{\text{LATE,FS}^2}^2$. Assumption 6 implies $\mathbb{E}_w[\mu_{\text{FS}}(S)^2] > 0$, while (95) implies that the denominator of (26) converges in probability to this strictly positive quantity. It is therefore bounded away from zero with probability approaching one. Expanding the ratio in (26) and using (94) and (95), we obtain

$$\begin{aligned}\sqrt{n} \left(\widehat{\sigma}_{\text{LATE,FS}^2}^2 - \sigma_{\text{LATE,FS}^2}^2 \right) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\psi_{q,i} - \sigma_{\text{LATE,FS}^2}^2 \psi_{\text{FS}^2,i}}{\mathbb{E}_w[\mu_{\text{FS}}(S)^2]} + o_{\mathbb{P}}(1) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{\sigma_{\text{LATE,FS}^2}^2,i} + o_{\mathbb{P}}(1),\end{aligned}$$

which proves (54).

Each influence function in the last display has mean zero. The observations are i.i.d. across sites by Assumptions 2 and 3, and $\mathbb{E}[\psi_{\sigma_{\text{LATE,FS}^2}^2,i}^2] < \infty$ by Assumption 11. The central limit theorem and Slutsky's lemma therefore imply

$$\sqrt{n} \left(\widehat{\sigma}_{\text{LATE,FS}^2}^2 - \sigma_{\text{LATE,FS}^2}^2 \right) \xrightarrow{d} N(0, V_{\sigma_{\text{LATE,FS}^2}^2}),$$

which proves Point 1.

Point 2. For this point impose Assumption 12 as well. Write $F_i^* = \widehat{\mu}_{\text{FS}}(S_i)$, $I_i^* = \widehat{\mu}_{\text{ITT}}(S_i)$, and

$$a_i = (I_i^*)^2 - \widehat{V}_{\text{ITT},i}, \quad b_i = F_i^* I_i^* - \widehat{C}_{\text{FS,ITT},i}, \quad c_i = (F_i^*)^2 - \widehat{V}_{\text{FS},i}.$$

Thus $g_i(l) = a_i - 2lb_i + l^2 c_i$. The additional assumption implies $\mathbb{E}[w(S_i)^2] < \infty$ and $\mathbb{E}[w(S_i)^2(I_i^*)^2] < \infty$. Binary take-up gives $|F_i^*| \leq 1$. Assumption 11, together with $\mathbb{E}[w(S_i)^2] < \infty$, implies that $w(S_i)b_i$, $w(S_i)c_i$, and $w(S_i)g_i(\text{LATE}_{\text{FS}^2})$ have finite second moments. Since

$$a_i = g_i(\text{LATE}_{\text{FS}^2}) + 2\text{LATE}_{\text{FS}^2}b_i - \text{LATE}_{\text{FS}^2}^2 c_i,$$

$w(S_i)a_i$ also has a finite second moment. Consequently, for $Z_i = (1, F_i^*, I_i^*, a_i, b_i, c_i)^\top$,

$$\mathbb{E}[\|w(S_i)Z_i\|^2] < \infty. \quad (96)$$

The normalized-weight expansion applied jointly to these variables gives root- n consistency and joint asymptotic linearity of their weighted averages. In particular,

$$\sqrt{n}(\widehat{\text{LATE}} - \text{LATE}) = \frac{1}{\sqrt{n}} \sum_i \psi_{\text{LATE},i} + o_{\mathbb{P}}(1).$$

For a fixed $l = \text{LATE}$ it also gives

$$\sqrt{n}\{\widehat{Q}(l) - Q(l)\} = \frac{1}{\sqrt{n}} \sum_i w(S_i)\{g_i(l) - h\} + o_{\mathbb{P}}(1).$$

Because $\widehat{Q}$ is quadratic, its exact Taylor expansion yields

$$\widehat{h} = \widehat{Q}(\text{LATE}) + \widehat{Q}'(\text{LATE})(\widehat{\text{LATE}} - \text{LATE}) + \widehat{m}_2(\widehat{\text{LATE}} - \text{LATE})^2.$$

Here $\widehat{Q}'(\text{LATE}) = 2(\widehat{m}_2\text{LATE} - \widehat{r}) \xrightarrow{\mathbb{P}} 2(m_2\text{LATE} - r)$, and the last term is $O_{\mathbb{P}}(n^{-1})$. It follows that

$$\sqrt{n}(\widehat{h} - h) = \frac{1}{\sqrt{n}} \sum_i \psi_{h,i} + o_{\mathbb{P}}(1).$$

The same joint expansion and the product rule give

$$\sqrt{n}(\widehat{d} - d) = \frac{1}{\sqrt{n}} \sum_i \psi_{d,i} + o_{\mathbb{P}}(1), \quad \sqrt{n}(\widehat{\delta} - \delta) = \frac{1}{\sqrt{n}} \sum_i \psi_{\delta,i} + o_{\mathbb{P}}(1).$$

Applying the delta method to $(h, d, \delta, \text{FS}) \mapsto (h + d^2/\delta)/\text{FS}$, whose denominators are strictly positive, yields

$$\psi_{\underline{\sigma}_{\text{LATE},i}^2} = \frac{1}{\text{FS}} \left\{ \psi_{h,i} + \frac{2d}{\delta} \psi_{d,i} - \frac{d^2}{\delta^2} \psi_{\delta,i} - \underline{\sigma}_{\text{LATE}}^2 \psi_{\text{FS},i} \right\}.$$

This proves (55). The joint central limit theorem implied by (96) gives the stated normal limit.

C.10 Proof of Proposition 6

First note that from (19),

$$\mathbb{E}_w[\mu_{\text{FS}}(S)\{\mu_{\text{LATE}}(S) - \text{LATE}\}] = 0.$$

Therefore,

$$\sigma_{\text{LATE,FS}} = \mathbb{E}_w \left[\frac{\mu_{\text{FS}}(S)}{\text{FS}} \{ \mu_{\text{LATE}}(S) - \text{LATE} \} \{ \mu_{\text{FS}}(S) - \text{FS} \} \right].$$

Then, using the fact that $\mu_{\text{ITT}}(S) = \mu_{\text{FS}}(S)\mu_{\text{LATE}}(S)$,

$$\begin{aligned} \text{FS } \sigma_{\text{LATE,FS}} &= \mathbb{E}_w [\mu_{\text{FS}}(S) \{ \mu_{\text{LATE}}(S) - \text{LATE} \} \{ \mu_{\text{FS}}(S) - \text{FS} \}] \\ &= \text{Cov}_w(\mu_{\text{FS}}(S), \mu_{\text{ITT}}(S)) - \text{LATE } \text{Var}_w(\mu_{\text{FS}}(S)) \\ &= \text{Var}_w(\mu_{\text{FS}}(S)) (\beta_{\text{ITT,FS}} - \text{LATE}), \end{aligned}$$

which proves (29).

C.11 Proof of Proposition 7

For Point 1, for every $j, \ell \in \{1, \dots, J\}$,

$$\begin{aligned} &\mathbb{E}_w \left[\{ \mu_{\text{ITT}_{d,j}}(S) - \mu_{\text{FS}_d}(S) \text{LATE}_{d,j,\text{FS}_d^2} \} \{ \mu_{\text{ITT}_{d,\ell}}(S) - \mu_{\text{FS}_d}(S) \text{LATE}_{d,\ell,\text{FS}_d^2} \} \right] \\ &= \mathbb{E}_w [\mu_{\text{FS}_d}(S)^2] \text{Cov}_{\text{FS}_d^2}(\mu_{\text{LATE}_{d,j}}(S), \mu_{\text{LATE}_{d,\ell}}(S)). \end{aligned}$$

Stacking the identities for $j, \ell \in \{2, \dots, J\}$ shows that the second-moment matrix of the transformed regressors is

$$\mathbb{E}_w [\mu_{\text{FS}_d}(S)^2] \text{Var}_{\text{FS}_d^2}(\boldsymbol{\mu}_{\text{LATE},d,-1}(S)),$$

while their cross-moment vector with the transformed outcome is

$$\mathbb{E}_w [\mu_{\text{FS}_d}(S)^2] \text{Cov}_{\text{FS}_d^2}(\boldsymbol{\mu}_{\text{LATE},d,-1}(S), \mu_{\text{LATE}_{d,1}}(S)).$$

The coefficient from the no-intercept regression is the inverse of the first expression multiplied by the second. Therefore, it is equal to

$$\begin{aligned} &\left\{ \mathbb{E}_w [\mu_{\text{FS}_d}(S)^2] \text{Var}_{\text{FS}_d^2}(\boldsymbol{\mu}_{\text{LATE},d,-1}(S)) \right\}^{-1} \mathbb{E}_w [\mu_{\text{FS}_d}(S)^2] \text{Cov}_{\text{FS}_d^2}(\boldsymbol{\mu}_{\text{LATE},d,-1}(S), \mu_{\text{LATE}_{d,1}}(S)) \\ &= \text{Var}_{\text{FS}_d^2}(\boldsymbol{\mu}_{\text{LATE},d,-1}(S))^{-1} \text{Cov}_{\text{FS}_d^2}(\boldsymbol{\mu}_{\text{LATE},d,-1}(S), \mu_{\text{LATE}_{d,1}}(S)) = \beta^d. \end{aligned}$$

For Point 2, cancellation of the two first stages gives

$$\mathbb{E}_{\text{FS}_1 \text{FS}_2} [\mu_{\text{LATE}_{1,k}}(S) \mu_{\text{LATE}_{2,k}}(S)] = \frac{\mathbb{E}_w [\mu_{\text{ITT}_{1,k}}(S) \mu_{\text{ITT}_{2,k}}(S)]}{\mathbb{E}_w [\mu_{\text{FS}_1}(S) \mu_{\text{FS}_2}(S)]}.$$

Similarly,

$$\mathbb{E}_{\text{FS}_1 \text{FS}_2} [\mu_{\text{LATE}_{1,k}}(S)] = \frac{\mathbb{E}_w [\mu_{\text{FS}_2}(S) \mu_{\text{ITT}_{1,k}}(S)]}{\mathbb{E}_w [\mu_{\text{FS}_1}(S) \mu_{\text{FS}_2}(S)]}, \quad \mathbb{E}_{\text{FS}_1 \text{FS}_2} [\mu_{\text{LATE}_{2,k}}(S)] = \frac{\mathbb{E}_w [\mu_{\text{FS}_1}(S) \mu_{\text{ITT}_{2,k}}(S)]}{\mathbb{E}_w [\mu_{\text{FS}_1}(S) \mu_{\text{FS}_2}(S)]}.$$

Under squared-first-stage-1 weights,

$$\mathbb{E}_{\text{FS}_1^2} [\mu_{\text{LATE}_{1,k}}(S)^2] = \frac{\mathbb{E}_w [\mu_{\text{ITT}_{1,k}}(S)^2]}{\mathbb{E}_w [\mu_{\text{FS}_1}(S)^2]}, \quad \mathbb{E}_{\text{FS}_1^2} [\mu_{\text{LATE}_{1,k}}(S)] = \frac{\mathbb{E}_w [\mu_{\text{FS}_1}(S) \mu_{\text{ITT}_{1,k}}(S)]}{\mathbb{E}_w [\mu_{\text{FS}_1}(S)^2]}.$$

Subtracting the products of the corresponding means and taking the covariance–variance ratio yields (33), proving Point 2.

D Survey of Multi-Site RCTs

Table 8: Multi-site RCTs in AEJ: Applied Economics 2014-2016

Title	Units of Observation	Units of Randomization	Sites	Randomization Units per Site
Keeping It Simple: Financial Literacy and Rules of Thumb	Individual clients	1,193 individual clients	107 barrios	11.1
Improving Educational Quality through Enhancing Community Participation: Results from a Randomized Field Experiment in Indonesia	Students	520 schools	44 subdistricts	11.8
The Demand for Medical Male Circumcision	Individuals	1,634 individuals	29 enumeration areas	56.3
Should Aid Reward Performance? Evidence from a Field Experiment on Health and Education in Indonesia	Individuals	300 subdistricts (kecamatan)	20 districts (kabupaten)	15.0
Private and Public Provision of Counseling to Job Seekers: Evidence from a Large Controlled Experiment	Individuals	43,977 individuals	216 employment offices	203.6
Estimating the Impact of Microcredit on Those Who Take It Up: Evidence from a Randomized Experiment in Morocco	Households	162 villages (81 matched pairs)	47 branches	3.4
Microcredit Impacts: Evidence from a Randomized Microcredit Program Placement Experiment by Compartamos Banco	Households	250 geographic clusters	45 superclusters	5.6
The Impacts of Microcredit: Evidence from Bosnia and Herzegovina	Individuals	1,196 individuals	282 city/towns or 14 branches	85.4
Social Networks and the Decision to Insure	Households	5,335 households	185 villages	28.8
Inputs in the Production of Early Childhood Human Capital: Evidence from Head Start	Individuals	4,442 individuals	353 Head Start centers	12.6
The Returns to Microenterprise Support among the Ultrapoor: A Field Experiment in Postwar Uganda	Individuals	904 individuals	60 villages	15.1
The Impact of High School Financial Education: Evidence from a Large-Scale Evaluation in Brazil	Students	892 schools (matched pairs)	6 states (municipalities not reported)	148.7
Median			46	15.0

Notes: "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. In "The Impact of High School Financial Education", randomization was within matched pairs of schools stratified by municipality; the paper does not report the number of municipalities or pairs, so we count the six states in which the schools are located as sites.