Optional ERIC Coversheet — Only for Use with U.S. Department of Education Grantee Submissions

This coversheet should be completed by grantees and added to the PDF of your submission if the information required in this form **is not included on the PDF to be submitted**.

INSTRUCTIONS

- Before beginning submission process, download this PDF coversheet if you will need to provide information not on the PDF.
- Fill in all fields—information in this form **must match** the information on the submitted PDF and add missing information.
- Attach completed coversheet to the PDF you will upload to ERIC [use Adobe Acrobat or other program to combine PDF files]—do not upload the coversheet as a separate document.
- Begin completing submission form at https://eric.ed.gov/submit/ and upload the full-text PDF with attached coversheet when indicated. Your full-text PDF will display in ERIC after the 12-month embargo period.

GRANTEE SUBMISSION REQUIRED FIELDS

Title of article, paper, or other content

IDENTIFYING AND ESTIMATING PRINCIPAL CAUSAL EFFECTS IN MULTI-SITE TRIALS

All author name(s) and affiliations on PDF. If more than 6 names, ERIC will complete the list from the submitted PDF.

Last Name, First Name	Academic/Organizational Affiliation	ORCID ID
Yuan, Lo-Hua	Harvard University	
Feller, Avi	University of California, Berkeley	
Miratrix, Luke	Harvard University	

Publication/Completion Date—(if *In Press,* enter year accepted or completed) Sept. 2019

Check type of content being submitted and complete one of the following in the box below:

- ☑ If article: Name of journal, volume, and issue number if available
- ☐ If paper: Name of conference, date of conference, and place of conference
- ☐ If book chapter: Title of book, page range, publisher name and location
- ☐ If book: Publisher name and location
- ☐ If dissertation: Name of institution, type of degree, and department granting degree

Annals of Applied Statistics, 13(3)

DOI or URL to published work (if available) 10.1214/18-AOAS1235

Acknowledgement of Funding— Grantees should check with their grant officer for the preferred wording to acknowledge funding. If the grant officer does not have a preference, grantees can use this suggested wording (adjust wording if multiple grants are to be acknowledged). Fill in Department of Education funding office, grant number, and name of grant recipient institution or organization.

"This work was supported by U.S. Department of Education **[Office name]** ______ through **[Grant number]** ______ to **Institution]** ______. The opinions expressed are those of the authors and do not represent views of the **[Office name]** ______ or the U.S. Department of Education.

IDENTIFYING AND ESTIMATING PRINCIPAL CAUSAL EFFECTS IN MULTI-SITE TRIALS*

By Lo-Hua Yuan[†] Avi Feller[‡] and
Luke W. Miratrix[†]

Harvard University[†], *University of California, Berkeley*[‡]

Randomized trials are often conducted with separate randomizations across multiple sites such as schools, voting districts, or hospitals. These sites can differ in important ways, including the site's implementation, local conditions, and the composition of individuals. An important question in practice is whether—and under what assumptions—researchers can leverage this cross-site variation to learn more about the intervention. We address these questions in the principal stratification framework, which describes causal effects for subgroups defined by post-treatment quantities. We show that researchers can estimate certain principal causal effects via the multi-site design if they are willing to impose the strong assumption that the site-specific effects are uncorrelated with the site-specific distribution of stratum membership. We motivate this approach with a multi-site trial of the Early College High School Initiative, a unique secondary education program with the goal of increasing high school graduation rates and college enrollment. Our analyses corroborate previous studies suggesting that the initiative had positive effects for students who would have otherwise attended a low-quality high school, although power is limited.

1. Introduction Randomized trials are often conducted at multiple physical sites, with separate randomizations across, for example, schools, voting districts, or hospitals (Raudenbush and Bloom, 2015). These sites can differ in important ways, including the site's implementation quality, local conditions, and the composition of individuals. Intuitively, researchers should be able to leverage such differences across sites to learn more about the intervention. For instance, if impacts are systematically larger at sites with higher student attendance, what can we conclude about dosage effects?

*We gratefully acknowledge funding from the Spencer Foundation through a grant entitled "Using Emerging Methods with Existing Data from Multi-site Trials to Learn About and From Variation in Educational Program Effects" and from the Institute for Education Sciences, U.S. Department of Education, through Grant #R305D150040. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Keywords and phrases: Principal causal effects, principal stratification, covariate restrictions, multi-site randomized trials, noncompliance, Early College High School

More broadly, what questions can researchers answer using this approach and what assumptions are required?

This paper explores the use of cross-site variation to estimate causal effects defined by individual-level post-treatment behavior. Our motivating example is a randomized evaluation of an alternative high school program in North Carolina, known as Early College High Schools (ECHS; Edmunds et al., 2012). ECHS is an innovative approach that aims to increase college readiness and college completion rates among students typically underrepresented in post-secondary education. Edmunds et al. (2017) find meaningful, positive impacts on a range of key academic outcomes, including ninth-grade success, high school graduation, and college enrollment. These positive results raise additional questions about expanding the program. In particular, is it more effective for certain types of students or in certain settings?

Our analysis focuses on the quality of the school each student would attend in the absence of the program. In general, we expect to see larger impacts of ECHS for students who would otherwise attend low-quality public schools than for those who would otherwise attend high-quality public schools. The goal is to assess whether this indeed holds in practice, which would help guide the expansion of the program. We make this question precise via the *principal stratification* framework of Frangakis and Rubin (2002) and define subgroups, known as principal strata, determined by each student's school quality in both the observed treatment condition and the counterfactual condition. While membership in these endogenous subgroups is only partially observed, the corresponding causal effects are nonetheless well defined.

Although principal stratification is a powerful framework for defining causal effects of interest, estimating these impacts can be elusive (Page et al., 2015). In the context of multi-site trials, we show that estimation is possible via a *zero correlation assumption*: the site-specific distribution of principal strata (e.g., the proportion of Compliers) is uncorrelated with the site-specific impacts for these principal strata. This is a very strong assumption, roughly implying that the interaction between randomization and site indicator functions as a "second instrument" (the first being treatment randomization) that is predictive of principal stratum membership, but is uncorrelated with the treatment impact within any stratum. As we argue, multi-site trials differ from more general stratified randomized trials because we can appeal to a (super) population of sites. Thus, rather than assume that certain quantities are constant and equal to zero for all sites, we can instead assume that these quantities equal zero *on average* across sites (see

Kolesár et al., 2015). We describe this zero-correlation assumption in the context of principal stratification in the ECHS study. We also address estimation and discuss the weaker assumption that zero correlation only holds conditional on a set of auxiliary covariates.

To the best of our knowledge, this is the first paper that brings together the otherwise disparate literatures of multi-site trials and covariate restrictions for principal stratification. We mention several highly relevant papers, and explore the connections in more depth in Section 7. First, Reardon and Raudenbush (2013) outline nine assumptions required to estimate mediation (rather than principal stratification) effects via cross-site variation (see also Raudenbush and Bloom, 2015; Reardon et al., 2014). Second, Kolesár et al. (2015) explore related questions from an econometric perspective and consider estimation with "many invalid instruments." Both Reardon and Raudenbush (2013) and Kolesár et al. (2015) impose a zero correlation assumption very similar to the one we explore here, though our setup gives researchers greater flexibility by requiring fewer necessary conditions for identification and estimation. Third, Jiang, Ding and Geng (2016) discuss identifying principal causal effects by leveraging results from multiple studies. They impose the much stronger assumption that these effects are constant ("homogeneous") across studies (see also Kline and Walters, 2016, for additional discussion). Many other papers impose restrictions on covariates to identify principal causal effects, including Jo (2002), Peck (2003), Ding et al. (2011) and Mealli, Pacini and Stanghellini (2016). Finally, Miratrix et al. (2018) investigate the same substantive question that we explore here, but use covariates to sharpen bounds rather than to obtain point estimates.

The paper proceeds as follows. Section 2 describes the multi-site Early College High School study. Section 3 formulates the principal strata and associated estimands for ECHS. Section 4 gives the key methodological results, including identification and estimation. Section 5 extends these results to incorporate auxiliary covariates. Section 6 presents the results for the ECHS study. Sections 7 and 8 discuss connections to other methods and conclude. The supplementary materials contain implementation details, an extensive simulation study, and additional discussion of other methods, especially ASPES (Peck, 2003).

2. Early College High Schools The Early College High School (ECHS) Initiative was launched in 2002 with support from the Bill and Melinda Gates Foundation. The program partners small, autonomous public high schools with two- or four-year colleges to give students the opportunity to earn an associate's degree or up to two years of transferable college credit, as well as a high school diploma. Early Colleges are designed to increase college readiness and graduation rates by exposing high school students to college-style courses, building students' confidence in their ability to succeed in a college environment, and lessening the financial burden of college by giving students the option to earn college credits while still in high school. These programs are targeted at individuals generally under-represented in college, including low income, first generation, and minority students. Early College programs were oversubscribed at some sites, which then allocated slots to applicants randomly, creating a de-facto randomized trial.

We analyze data from the Evaluation of Early College High Schools in North Carolina (Edmunds et al., 2010). This study tracked a sample of 4,004 students who began ninth grade between 2005 and 2010 and who entered in one of 44 lotteries to gain entry into one of 19 different Early College programs. These ECHS programs are spread across the state, such that it was only feasible for a student to enter into a single lottery. Within each lottery, students were randomized either to receive or not receive an offer to attend an ECHS. Following Miratrix et al. (2018), we limit our analytic set to students who could be linked to the North Carolina Department of Instruction (NCDPI) databank, had school enrollment data in ninth grade, and had transcript data or End of Course exam data from NCDPI. We subset our sample to students whose ninth grade school was within 20 miles of their eighth grade school, under the assumption that a large distance between a student's middle and high schools indicates that the student moved between eighth and ninth grade, and was therefore effectively dropped from the trial. We also exclude students for whom we do not have complete information on race, gender, free or reduced-price lunch eligibility, first generation college student status, and eighth grade math and reading scores. Finally, to avoid unnecessary technical complications in the main text, we exclude the six lotteries that have no variability in our outcome measure of interest. We report the same analysis with all 44 lotteries in the supplement, which yields nearly identical conclusions.

Given these inclusion criteria, our final ECHS analysis sample consists of 3,477 students ($N_t = 2,021$, $N_c = 1,456$) across 38 lotteries in 18 ECHS schools, each with up to 6 cohorts. Throughout, we use the term 'site' to denote a specific lottery rather than a specific school. A key reason for this choice is that the proportion of principal strata can vary meaningfully within a school year to year, which complicates school-level analyses.

Outcomes. The North Carolina ECHS data set contains a battery of outcome measures. Our outcome of interest is a binary indicator of whether a student is "on track" to complete the Future-Ready Core Graduation Requirements set by the state of North Carolina at the end of ninth grade. This measure is based on compelling descriptive evidence that students who do well in ninth grade are more likely to excel in and graduate from high school (Allensworth, 2005).[1]

Covariates. Student baseline covariates include race, gender, free or reduced-price lunch eligibility, first generation college student status, and standardized eighth grade math and reading scores. Table 1 in the supplementary materials shows balance checks, stratified by lottery. Early College High Schools target students who would traditionally not enroll in college, and several schools in the study gave priority to groups underrepresented in higher education. As such, the ECHS sample is relatively disadvantaged, with around half of all students in the lottery eligible for free or reduced-price lunch. We also see slight imbalances in racial categories, with the treatment group comprised of more Black/African American students than the control group. We do not detect imbalance in any of the other baseline covariates.

Student sampling weights. In the ECHS study, students had unequal but known probabilities of winning a lottery. Some lotteries were more selective overall. Some lotteries gave certain students higher chances of a slot for equity reasons. All the calculations we perform on the ECHS data set use student-level sampling weights that reflect each student's probability of entering and winning a lottery based on demographics and other factors. In particular, we apply the same Hàjek estimator sample weighting approach discussed and used by Miratrix et al. (2018).

School quality. We label each school in the North Carolina Early College Study as one of three school types: high-quality public high school, low-quality public high school, or Early College High School. The high- and low-quality ratings are based on a composite of school-level measures, including achievement metrics, growth, and adequate yearly progress, as tracked by a centralized State of North Carolina school-report-card system. Schools classified by the state as "priority schools", "low performing schools", and "schools receiving no recognition" are categorized as low-quality schools. "Schools making high growth", "schools making expected growth", "honor schools of excellence", "schools of excellence", and "schools of progress" are classified as high-quality schools.[2] While the state also rates Early Colleges

[1]Details of the Future-Ready Core's requirements for math and English language reading and writing are at `http://www.dpi.state.nc.us/docs/gradrequirements/resources/gradchecklists.pdf`.

[2]See `http://www.ncpublicschools.org/docs/accountability/reporting/abc/2005-06/execsumm.html` for classification details.

TABLE (1)
Distribution of high school type by treatment status

School type	Treatment ($N_t = 2,021$)	Control ($N_c = 1,456$)
Early College HS (e)	85.4%	2.7%
High-Quality Public HS (hq)	2.4%	12.4%
Low-Quality Public HS (lq)	12.3%	85.0%

as either low- or high-quality, we treat ECHSs as their own quality category because an ECHS operates on principles that are distinct from a traditional public high school and provides students with a unique education environment that may not be captured by standard school rating measures.

Table 1 shows the distribution of ninth grade students in our data set across these three school types. In the treatment group, 85.4% of students attended an ECHS; 2.4% attended a high-quality school; 12.3% attended a low-quality school. In the control group, only 2.7% percent were able to cross over and register in an ECHS; 12.4% attended a high-quality school; 85% attended a low-quality school.

3. Setup and estimands We now describe the setup and estimands for the ECHS study using the principal stratification framework. Let Z_i be the treatment indicator for whether student i is randomly assigned to the active intervention, i.e., wins the lottery and is invited to enroll in an ECHS. Let Y_i^{obs} denote student i's observed outcome, i.e., the student's on-track status at the end of her ninth grade academic year. We assume randomization was valid within each lottery and that lotteries are independent. We also invoke SUTVA (Rubin, 1980), assuming that there is no interference between units and that there is one version of each treatment level; this precludes murky communication of whether someone wins the lottery and is invited to enroll in an ECHS. With these assumptions, we can then write down the potential outcomes for student i as $Y_i(1)$ and $Y_i(0)$, which are student i's on-track status depending on whether or not she receives an Early College enrollment offer. Her observed on-track status is $Y_i^{\text{obs}} = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$.

Given this setup, the overall Intent-to-Treat (ITT) effect is therefore

$$\text{Overall ITT } = \mathbb{E}[Y_i(1) - Y_i(0)],$$

the average impact of the ECHS enrollment offer on students' on-track status. For ease of exposition, we initially regard expectations and probabilities as being taken over a super-population of individuals, with individuals from a specific lottery as a random sample of this super-population. We discuss a corresponding super-population of sites in Section 4.

TABLE (2)
The nine possible principal strata in the ECHS study. We assume that strata (A) - (D) *do not exist, leaving five principal strata. The two highlighted cells indicate the strata of interest.*

		No ECHS offer ($Z_i = 0$)		
		$D_i(0) = e$	$D_i(0) = lq$	$D_i(0) = hq$
ECHS offer ($Z_i = 1$)	$D_i(1) = e$	**ECHS Always Taker**	**Low-Quality Complier**	**High-Quality Complier**
	$D_i(1) = lq$	*(A)*	**Low-Quality Always Taker**	*(C)*
	$D_i(1) = hq$	*(B)*	*(D)*	**High-Quality Always Taker**

We can now go beyond the overall impact of randomization using the principal stratification framework. Let $D_i(z) \in \{e,\ lq,\ hq\}$ denote the quality of school a student would attend if assigned to treatment level $Z_i = z$, where e, lq, and hq are abbreviations for ECHS, low-quality, and high-quality, respectively. We now define our principal strata S_i by the pair of school types a student would attend if assigned to treatment, $D_i(1)$, and if assigned to control, $D_i(0)$.

Table 2 shows the $3^2 = 9$ possible principal strata; rows indicate school type for students when assigned to treatment and columns indicate school type when assigned to control. The analysis becomes unwieldy without restrictions on the possible principal strata (see, e.g., Page et al., 2015). We therefore make structural assumptions that imply that strata *(A)* through *(D)* do not exist, which reduces the number of possible strata from nine to five. First, we assume that there are no Defiers (Angrist, Imbens and Rubin, 1996); that is, there are no individuals who only enroll in ECHS if denied the opportunity to do so.

ASSUMPTION 3.1 (No Defiers, or Monotonicity). *There are no individuals with* $\{D_i(1) = lq,\ D_i(0) = e\}$ *or* $\{D_i(1) = hq,\ D_i(0) = e\}$.

This eliminates strata *(A)* and *(B)*. To eliminate strata *(C)* and *(D)* we need an additional assumption:

ASSUMPTION 3.2 (No Flip-Floppers). *There are no individuals with* $\{D_i(1) = lq,\ D_i(0) = hq\}$ *or* $\{D_i(1) = hq,\ D_i(0) = lq\}$.

This assumption states that individuals do not switch the type of non-ECHS school as a result of the ECHS lottery. Kline and Walters (2016) refer to this as an independence of irrelevant alternatives assumption. Applying Assumptions 3.1 and 3.2 leaves five remaining strata: ECHS Always Takers (*eat*), Low-Quality Compliers (*lc*), High-Quality Compliers (*hc*), Low-Quality Always Takers (*lat*), and High-Quality Always Takers (*hat*), as shown in Table 2. As we show in the supplementary materials, we can use these assumptions to identify the distribution of principal strata, π_s.

Next, we extend the standard exclusion restrictions (e.g., Angrist, Imbens and Rubin, 1996) to the three "Always" strata in the more general setup:

ASSUMPTION 3.3 (Exclusion restrictions). *There is no impact of randomization for individuals in the Always ECHS, Always Low-Quality, or Always High-Quality strata. That is,*

$$ITT_{eat} = ITT_{lat} = ITT_{hat} = 0.$$

The logic here is identical to the simpler noncompliance setting. That is, since randomization has no impact on school quality for students in these groups, we assume that randomization also has no impact on their later outcomes. Finally, we can decompose the overall ITT effect into stratum-specific ITTs. Under Assumptions 3.1, 3.2, and 3.3:

$$\begin{aligned} \text{Overall ITT } &= \pi_{lc}\text{ITT}_{lc} + \pi_{hc}\text{ITT}_{hc} + \pi_{eat}\text{ITT}_{eat} + \pi_{lat}\text{ITT}_{lat} + \pi_{hat}\text{ITT}_{hat} \\ &= \pi_{lc}\text{ITT}_{lc} + \pi_{hc}\text{ITT}_{hc}. \end{aligned} \tag{3.1}$$

We can simplify this slightly by normalizing by the overall proportion of Compliers, $\pi_{lc} + \pi_{hc}$:

$$\begin{aligned} \text{Overall LATE} = \text{ITT}_c& \\ &= \frac{\pi_{lc}}{\pi_{lc}+\pi_{hc}}\text{ITT}_{lc} + \frac{\pi_{hc}}{\pi_{lc}+\pi_{hc}}\text{ITT}_{hc} \\ &= (1-\phi)\ \text{ITT}_{lc} + \phi\ \text{ITT}_{hc}, \end{aligned} \tag{3.2}$$

where $\phi = \frac{\pi_{hc}}{\pi_{lc}+\pi_{hc}}$ is the proportion of Compliers that have a High-Quality alternative.

We now have one equation and two unknowns. Without additional restrictions, we can only "set identify" the two impacts of interest, ITT_{lc} and ITT_{hc}, as in Miratrix et al. (2018). In the next section, we discuss the use of cross-site variation to achieve point identification. Other approaches are possible. First, Feller et al. (2016a) use a Bayesian model-based approach to estimate similar effects, though Feller et al. (2016b) suggest that such estimates might be unstable. Second, Mealli, Pacini and Stanghellini (2016) explore the use of multiple outcomes and other covariate restrictions. Finally, Kline and Walters (2016) identify these effects by imposing restrictions on the school type selection process.

4. Identification and estimation via zero site-level correlation We now turn to methods that exploit the multi-site experimental design to identify causal effects. We introduce the core identifying assumption and the super-population of sites, and briefly discuss estimation, deferring many details to the supplementary materials.

4.1. *Super-population of sites and the zero correlation assumption* We slightly extend our notation to emphasize the data's multi-site structure. Let $k = 1, 2, \ldots, K$ index the K sites of the experiment, where $X_i = k$ denotes that student i belongs to experimental site k. Let $\text{ITT}_{s|k} = \mathbb{E}\left[Y_i(1) - Y_i(0) | S_i = s, X_i = k\right]$ be the impact of randomization for principal stratum s in site k, with $\text{LATE}_k = \text{ITT}_{c|k}$; let $\pi_{s|k} = \mathbb{P}\{S_i = s | X_i = k\}$ be the proportion of individuals in principal stratum s in site k; and let $\phi_k = \pi_{hc|k}/(\pi_{lc|k}+\pi_{hc|k})$ denote the proportion of Compliers in site k who are of High-Quality type. Our parameters of interest are the population average treatment impacts for Low-Quality Compliers and High-Quality Compliers, ITT_{lc} and ITT_{hc}, for all students across all sites.

A key conceptual advance and statistical advantage of the multi-site setting, relative to a setting with a generic categorical covariate, is that we can envision a super-population of sites from which the K observed sites are drawn. This is sometimes referred to as a random effects formulation (see, for example, Kolesár et al., 2015), though we prefer to focus on the existence of a super-population. Specifically, we assume that we sample sites represented as triples of parameters $\left(\text{ITT}_{lc|k},\ \text{ITT}_{hc|k},\ \phi_k\right)$ from an infinite super-population of sites with mean vector $(\text{ITT}_{lc}, \text{ITT}_{hc}, \phi)$ and a 3×3 correlation matrix $\mathbf{\Sigma}$:

$$\begin{pmatrix} \text{ITT}_{lc|k} \\ \text{ITT}_{hc|k} \\ \phi_k \end{pmatrix} \overset{iid}{\sim} \left[\begin{pmatrix} \text{ITT}_{lc} \\ \text{ITT}_{hc} \\ \phi \end{pmatrix},\ \begin{pmatrix} \Sigma_{11} & & \\ \Sigma_{21} & \Sigma_{22} & \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{pmatrix} \right] \tag{4.1}$$

Under this interpretation, we extend the single super-population of individuals described in Section 3 to instead have two stages of sampling: first, we sample a site from an infinite super-population of sites; second, we sample an individual from the site-specific super-population.

Given this setup, it is natural to re-frame the main problem in terms of regression. First, re-write Equation (3.2) separately for each site, re-arrange terms, and add zero twice to obtain

$$\begin{aligned}
\text{LATE}_k &= (1-\phi_k)\ \text{ITT}_{lc|k} + \phi_k\ \text{ITT}_{hc|k} \\
&= (1-\phi_k)\ \text{ITT}_{lc} + \phi_k\ \text{ITT}_{hc}\ + \\
&\qquad (1-\phi_k)\ (\text{ITT}_{lc|k} - \text{ITT}_{lc}) + \phi_k\ (\text{ITT}_{hc|k} - \text{ITT}_{hc}) \\
&= (1-\phi_k)\ \text{ITT}_{lc} + \phi_k\ \text{ITT}_{hc} + (1-\phi_k)\ \epsilon_{lc|k} + \phi_k\ \epsilon_{hc|k}, \qquad (4.2)
\end{aligned}$$

where $\epsilon_{lc|k} = \text{ITT}_{lc|k} - \text{ITT}_{lc}$ and $\epsilon_{hc|k} = \text{ITT}_{hc|k} - \text{ITT}_{hc}$. Across all K sites, we therefore have a system of K linear equations:

$$\begin{aligned}
\text{LATE}_1 &= (1-\phi_1)\text{ITT}_{lc} + \phi_1\text{ITT}_{hc} + \eta_1 \\
\text{LATE}_2 &= (1-\phi_2)\text{ITT}_{lc} + \phi_2\text{ITT}_{hc} + \eta_2 \\
&\ \ \vdots \\
\text{LATE}_K &= (1-\phi_K)\text{ITT}_{lc} + \phi_K\text{ITT}_{hc} + \eta_K, \qquad (4.3)
\end{aligned}$$

where we condense the final terms: $\eta_k = (1-\phi_k)\ \epsilon_{lc|k} + \phi_k\ \epsilon_{hc|k}$.

This is a bivariate linear regression with no intercept, in which ITT_{lc} and ITT_{hc} are regression coefficients and η_k is the regression error term. Since we have a super-population of sites, we can identify the causal effects of interest under the classical assumption that the regression errors, η_k are uncorrelated with the regressors, ϕ_k and $1-\phi_k$, in the super-population. Specifically, we can identify the regression coefficients under the assumptions that $\text{Cov}(\epsilon_{lc|k}, \phi_k) = 0$ and $\text{Cov}(\epsilon_{hc|k}, \phi_k) = 0$, with the additional normalization that $\mathbb{E}\left[\epsilon_{lc|k}\right] = 0$ and $\mathbb{E}\left[\epsilon_{hc|k}\right] = 0$; or combining terms, $\text{Cov}(\eta_k, \phi_k) = 0$ and $\mathbb{E}\left[\eta_k\right] = 0$.

ASSUMPTION 4.1 (Zero site-level correlation between principal stratum distribution and principal causal effects). *The site-specific relative share of High-Quality Compliers is uncorrelated with the site-specific impacts for High-Quality Compliers and for Low-Quality Compliers.*

$$\text{Cov}(\epsilon_{lc|k}\ ,\ \phi_k) = 0 \quad \textit{and} \quad \text{Cov}(\epsilon_{hc|k}\ ,\ \phi_k) = 0. \qquad (4.4)$$

This is equivalent to assuming that $\Sigma_{31} = \Sigma_{32} = 0$ in Equation (4.1). In addition, we require that $\text{Var}(\phi_k) > 0$, that is $\Sigma_{33} > 0$, which is analogous to the relevancy assumption in standard instrumental variables. We combine all these assumptions into the following proposition.

Proposition 4.2 (Identification of principal causal effects via zero site-level correlation). *For a multi-site trial with $K \geq 2$ sites, under assumption 4.1, $\mathrm{Var}(\phi_k) > 0$, and the normalization that $\mathbb{E}\left[\epsilon_{lc|k}\right] = 0$ and $\mathbb{E}\left[\epsilon_{hc|k}\right] = 0$, the principal causal effects, ITT_{lc} and ITT_{hc}, are identified.*

The proof for Proposition 4.2 follows immediately from standard regression theory.[3] Importantly, while these results do not strictly require an underlying super-population of sites, it is difficult to imagine these conditions holding for a generic categorical covariate.

In the context of ECHS, the zero correlation assumption states that the impact of the program on High-Quality Compliers' ninth grade performance in a site does not systematically vary according to the relative proportion of High-Quality versus Low-Quality Compliers in a site; with the same assumption for Low-Quality Compliers. This strong assumption precludes factors that may differ across sites — such as the average academic preparedness of incoming ninth grade students — from influencing both the student compliance make-up of a site and the magnitude of impact ECHS has on students within the site. Intuitively, students who are more academically prepared might have more resources and support, such that they would attend a High-Quality public school if they did not attend an ECHS. In addition, students who enter ninth grade with a stronger academic background might experience ECHS differently from incoming students who have weaker academic foundations. To accommodate this kind of scenario, we discuss relaxing the zero-correlation assumption to hold conditional on covariates, such as prior academic preparedness, in Section 5.

Finally, it is useful to re-frame this setup in terms of the contrast $\mathrm{ITT}_{hc} - \mathrm{ITT}_{lc}$. We can re-write Equation (4.3) to highlight this directly:

$$\mathrm{LATE}_k = \mathrm{ITT}_{lc} + \phi_k(\mathrm{ITT}_{hc} - \mathrm{ITT}_{lc}) + \eta_k, \text{ for } k = 1, \ldots, K. \tag{4.5}$$

This yields a particularly simple form when there are only two sites, j and k:

$$\mathrm{ITT}_{hc} - \mathrm{ITT}_{lc} = \frac{\mathrm{LATE}_j - \mathrm{LATE}_k}{\phi_j - \phi_k}. \tag{4.6}$$

This is the slope of a line based on two points. It is also identical in form to the standard ratio estimator in instrumental variables, which underscores

[3]These zero correlation and marginal zero expectation conditions are precisely the moment conditions needed to identify the regression coefficients in a linear regression model. A stronger assumption often cited for regression is *strict exogeneity*, which states that the conditional mean of the error terms given the regressor equals zero, $\mathbb{E}[\epsilon_{s|k}|\phi_k] = 0$. This assumption implies the two moment conditions above, but the reverse is not true; see Reardon and Raudenbush (2013) for additional discussion in this context.

the connection to using the interaction of "site by randomization" as an additional instrument. See the supplementary materials for additional discussion of restrictions with a binary covariate, including a discussion of the ASPES approach of Peck (2003).

4.2. *Estimation* In order to estimate these effects, we begin with an overly simplistic approach that uses plug-in estimators for the site-specific moments, $\widehat{\mathrm{LATE}}_k$ and $\widehat{\phi}_k$. Let $\widehat{Y}_{zd} = \frac{1}{N_{zd}} \sum_{i \in \{Z_i = z,\, D_i^{obs} = d\}} Y_i^{obs}$ be the finite sample average observed outcome for students assigned to $Z_i = z$ with observed take up $D_i^{\mathrm{obs}} = d$, and let $\widehat{Y}_{zd|k}$ be the corresponding estimate for students in site k. $\widehat{Y}_{z\cdot|k}$ indicates a summation over d; that is, the average observed outcome for students at site k who were randomized to study arm z. Let $\widehat{\pi}_s$ denote the estimated proportion of individuals in principal stratum s, with $\widehat{\pi}_{s|k}$ the corresponding estimate for students in site k. (See the supplementary materials for details.) We then estimate the site-specific LATE as

$$\widehat{\mathrm{LATE}}_k = \frac{\widehat{Y}_{1\cdot|k} - \widehat{Y}_{0\cdot|k}}{\widehat{\pi}_{lc|k} + \widehat{\pi}_{hc|k}},$$

where $\widehat{\pi}_{lc|k} + \widehat{\pi}_{hc|k}$ is the estimated proportion of Compliers in site k. We can also estimate the relative proportion of High-Quality Compliers in site k:

$$\widehat{\phi}_k = \frac{\widehat{\pi}_{hc|k}}{\widehat{\pi}_{lc|k} + \widehat{\pi}_{hc|k}}.$$

With these site-aggregate statistics, we then estimate ITT_{lc} and ITT_{hc} via the regression coefficients from the site-level linear regression,

$$\widehat{\mathrm{LATE}}_k = \beta_{lc}\ (1 - \widehat{\phi}_k) + \beta_{hc}\ \widehat{\phi}_k + \eta_k\ , \tag{4.7}$$

where $\widehat{\beta}_{lc}$ and $\widehat{\beta}_{hc}$ are estimators for ITT_{lc} and ITT_{hc}, respectively. Taking the site-specific estimates, $\widehat{\mathrm{LATE}}_k$ and $\widehat{\phi}_k$, as fixed, we can account for uncertainty with the usual heteroskedastic-robust standard errors for linear regression (MacKinnon and White, 1985).

Measurement error. The plug-in approach ignores the fact that $\widehat{\mathrm{LATE}}_k$ and $\widehat{\phi}_k$ are estimated rather than known. This leads to two key complications. One complication is that conventional estimates of the standard error will under-estimate the true sampling variance. Also, the nominal point estimates could be biased; in particular, error in $\widehat{\phi}_k$ will attenuate the estimate of $\mathrm{ITT}_{hc} - \mathrm{ITT}_{lc}$. To account for the increased uncertainty due to measurement error, we therefore propose a straightforward case-resampling bootstrap approach that randomly samples students with replacement within each site.

For each bootstrap sample and independently for each site, we re-calculate $\widehat{\text{LATE}}_k^*$ and $\widehat{\phi}_k^*$ and then estimate ITT_{lc}^* and ITT_{hc}^* via the linear model 4.7. Finally, we apply standard multiple imputation combining rules (Rubin, 1987) to obtain a single point estimate and standard error for each principal causal effect.

Extensive simulation studies (see supplementary materials) show that this procedure has meaningfully smaller RMSE than the naive procedure, but that bias in the point estimate is still problematic. Many alternatives are possible, such as a parametric bootstrap, which repeatedly draws $\widehat{\text{LATE}}_k^*$ and $\widehat{\phi}_k^*$ via a multivariate Normal with means and covariances estimated from each site. See the discussion in Section 8.

Varying site size. Finally, site sizes typically vary in practice, which introduces additional complications. Specifically, the super-population means $(\text{ITT}_{lc}, \text{ITT}_{hc}, \phi)$ discussed in Section 4.1 correspond to site-level averages. If all sites have the same number of students, then the average over all sites equals the average over all students. If site sizes vary, however, we must choose whether to weight sites equally (site average) or weight individuals equally (population average). Following Raudenbush and Schwartz (2017), when sites have different numbers of Compliers, the unweighted linear model 4.7 estimates the average principal causal effects across sites, rather than across individuals. If, in addition to the conditions listed in Proposition 4.2, we also assert that $\text{ITT}_{lc|k}$ and $\text{ITT}_{hc|k}$ are independent of N_k, the number of Compliers in a site, then the population- and site-weighted estimates are equal. We return to this issue in the next section.

5. Conditional zero-correlation In practice, we often observe a rich set of individual- and site-level covariates. While potentially helpful for increasing efficiency, such covariates are particularly useful for relaxing the unconditional zero correlation of Assumption 4.1. Let $\mathbf{W}_k$ be a w-length vector of site-level covariates, which includes inherently site-level quantities, such as community type (urban, suburban, rural), as well as aggregate individual-level covariates, such as percent Free or Reduced-Price Lunch. We can then relax the zero correlation assumption such that it only holds *conditionally*:

$$\text{Cov}\left(\epsilon_{lc|k}\ ,\ \phi_k\ |\mathbf{W}_k\right) = 0 \quad \text{and} \quad \text{Cov}\left(\epsilon_{hc|k}\ ,\ \phi_k\ |\mathbf{W}_k\right) = 0\ , \tag{5.1}$$

with $\mathbb{E}\left[\epsilon_{s|k}|\mathbf{W}_k\right] = 0$, for $s \in \{lc, hc\}$. In the context of ECHS, this says, for example, that among sites of the same community type containing students of the same average level of academic preparedness, the impact of the ECHS program on different Complier types does not systematically vary according to the ratio of High- to Low-Quality Compliers in a site. In general, to obtain consistent estimates for the principal causal effects, we want to condition on confounding factors of compliance and treatment impacts; that is, baseline covariates that are predictive of the distribution of principal strata in a site, and, separately, are predictive of the site-specific principal causal effects.

There are several possible estimation procedures that incorporate auxiliary covariates under Assumption 5.1. The most straightforward, given our regression setup, is to include (grand-mean centered) site-aggregate values of confounders as additional regressors in the site-level linear regression. Specifically, instead of fitting model 4.7, we fit

$$\widehat{\text{LATE}}_k = \beta_{lc}^{adj}\ (1 - \widehat{\phi}_k) + \beta_{hc}^{adj}\ \widehat{\phi}_k + \boldsymbol{\gamma}\mathbf{W}_k + \eta_k^{adj}\ . \tag{5.2}$$

As above, $\mathbf{W}_k$ is a vector of site-aggregate covariate values, which could also include N_k, the total number of Compliers in site k.

The simple regression-adjusted model, however, restricts the possible treatment effect variation; see supplementary materials for additional discussion. For example, if we believe a baseline covariate $W_{1,k}$ influences the impact of ECHS on student on-track status differently for a predominately High-Quality Complier site compared to a site with mostly Low-Quality Compliers, then we may prefer the interaction adjusted model

$$\widehat{\text{LATE}}_k = \beta_{lc}^{int}\ (1-\widehat{\phi}_k) + \beta_{hc}^{int}\ \widehat{\phi}_k + \boldsymbol{\gamma}\ \mathbf{W}_{-1,k} + \delta_{lc}(1-\widehat{\phi}_k)W_{1,k} + \delta_{hc}\ \widehat{\phi}_k W_{1,k} + \eta_k^{int}\ , \tag{5.3}$$

where appropriate combinations of $\widehat{\beta}_s^{int}$ and $\widehat{\delta}_s$ yield estimates of the site-average impacts.

Finally, when site sizes vary, we can re-weight the regression coefficient estimates from Eqs. (5.2) or (5.3) to obtain population-average impacts under the assumption that $\text{ITT}_{lc|k}$ and $\text{ITT}_{hc|k}$ are *conditionally* independent of N_k, the number of Compliers in a site, given $\mathbf{W}$. For High-Quality Compliers, we have the following weighted average:

$$\widehat{\text{ITT}}_{hc}^{pop} = \sum_{k=1}^{K} \left(\widehat{\beta}_{hc}^{int} + \widehat{\boldsymbol{\gamma}}\ \mathbf{W}_{-1,k} + \widehat{\delta}_{hc}W_{1,k}\right) \frac{\hat{\phi}_k N_k}{\sum_{k=1}^{K} \hat{\phi}_k N_k}, \tag{5.4}$$

with an analogous estimate for Low-Quality Compliers.

6. Analysis of ECHS

6.1. *Main analysis* We investigate the impact of ECHS on the ninth grade on-track status of High-Quality Complier and Low-Quality Complier students. As we discuss in Section 4.1, we initially assume that the average impact of the Early College program on High-Quality Compliers' ninth grade performance is the same, in expectation, across all sites, and does not systematically vary according to the relative proportion of High-Quality versus Low-Quality Compliers in a site (with the the same for Low-Quality Compliers). We then relax this assumption by conditioning on standardized eighth grade reading score, which is predictive of both the relative proportion of High-Quality Compliers and of on-track percentages in sites.[4]

As described in the supplementary materials, we estimate impacts without covariate adjustment, with simple linear adjustment for site-average reading score, and with an interaction adjustment for site-average reading score. We account for different site sizes by taking weighted averages of predicted site-level impacts.

Figure 1 shows scatterplots of the estimated site-specific Complier impacts of ECHS on proportion on-track versus the estimated relative proportion of High-Quality Compliers in each site, before and after adjusting for site-average eighth grade reading score. As the left panel shows, 22 of the 38 sites have an estimated $\widehat{\phi}_k = 0$, meaning that we estimate that all of the Compliers at these sites are Low-Quality Compliers. Since the Low-Quality Compliers are also the much larger group, we therefore anticipate more precise estimates of ITT_{lc} than ITT_{hc}.

Figure 2 shows the corresponding point estimates and 95% confidence intervals for ITT_{lc} and ITT_{hc}. All the point estimates are positive, between 5.7 and 8.5 percentage points. There is no noticeable difference between the unadjusted versus simple adjusted or interaction adjusted point estimates for ITT_{lc}; nor is there a meaningful difference between the naive and bootstrap point estimates. Reading score adjustment has a more noticeable effect on point estimates for ITT_{hc}, with $\widehat{\mathrm{ITT}}_{hc}$ decreasing by about 1.3 percentage points under both simple linear adjustment and interaction adjustment.

The standard errors for both $\widehat{\mathrm{ITT}}_{lc}$ and $\widehat{\mathrm{ITT}}_{hc}$ increase slightly under interaction adjustment, compared to no adjustment or simple adjustment. For ITT_{lc} and ITT_{hc}, respectively, the bootstrap CI for each adjustment

[4] Eighth grade reading score is also highly correlated with many of the other available covariates (see also Miratrix et al., 2018). Adjusting for all six available baseline covariates—student race, gender, free or reduced-price lunch eligibility, first generation college student status, and standardized eighth grade reading and math scores—yields meaningfully noisier estimates. An additional complication is that many of these lotteries are for the same ECHS program over multiple years. In principle, we could restrict the sample to schools with multiple lotteries and condition our analysis on the specific ECHS or specify a hierarchical model. In practice, this is infeasible with our limited number of sites.

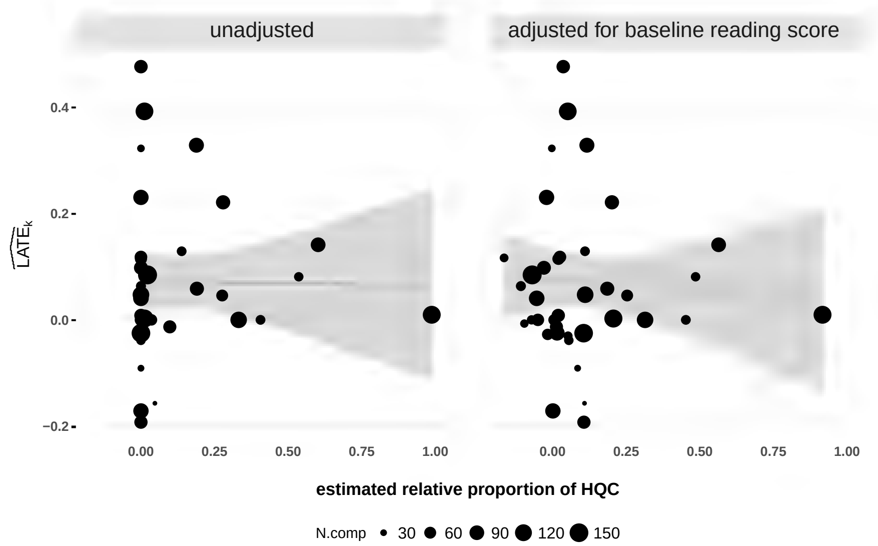

Fig (1) ECHS site-level data. Scatterplots of estimated site-specific Complier impacts (proportion on-track) versus (left panel) estimated relative proportion of High-Quality Compliers in a site, and (right panel) estimated *residual* relative proportion of High-Quality Compliers in a site, after regressing $\hat{\phi}_k$ on eighth grade reading score. The size of the points indicate the number of Compliers in a site. The lines fit to the points correspond to linear regressions with a free intercept; the y-intercept for each line is an estimate for ITT_{lc}, while the slope of each line is an estimate for the contrast $\text{ITT}_{hc} - \text{ITT}_{lc}$. The shaded grey regions are 95% confidence intervals for the conditional mean outcome.

method is roughly 23% and 40% wider than the CI of the corresponding naive estimate. This aligns with our simulation study finding that the bootstrap method produces overly conservative confidence intervals. Although we do not illustrate the results here, we note that adjusting for any single baseline covariate produces results that are substantively the same as those for reading score adjustment. Finally, we assess whether there are meaningful differences between ITT_{hc} and ITT_{lc} using the re-parameterization in Equation (4.5), which is illustrated by Figure 1, in which the y-intercept is an estimate for ITT_{lc} and the slope for $\widehat{\phi}_k$ is an estimate for the difference $\text{ITT}_{hc} - \text{ITT}_{lc}$. We do not find meaningful differences in stratum impacts for High- vs Low-Quality Compliers.

Overall, we find that the estimated impacts are quite similar for both Low- and High-Quality Compliers and that these estimates are stable across different models. Partly because the Low-Quality Complier group is larger,

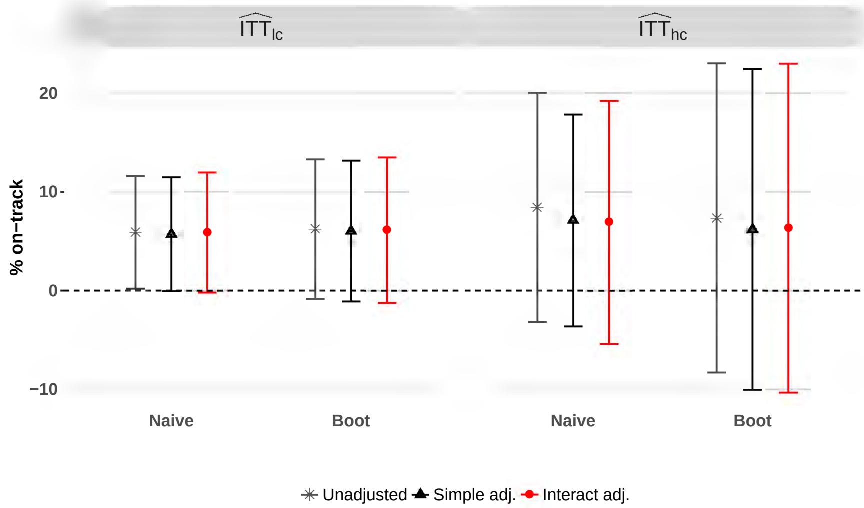

Fig (2) **Estimates of principal causal effects.** Point estimates and 95% confidence intervals for Low- and High-Quality Complier principal causal effects are plotted for each estimation method.

we are much more confident that the impact for this group is positive. By contrast, the estimated impact for High-Quality Compliers is much noisier. These results are consistent with the bounds in Miratrix et al. (2018).

6.2. *Model checking* An advantage of using a regression-based approach is that we can assess key identifying assumptions using standard regression diagnostics. In particular, the zero site-level correlation between principal stratum membership and stratum-specific impacts (Assumption 4.1) implies that $\mathbb{E}\left[\eta_k\right] = 0$ and $\mathrm{Cov}(\eta_k, \phi_k) = 0$. We can use the fitted residuals from the site-level regression to assess the evidence against these assumptions, though power might be limited. Importantly, the zero-correlation assumption is restricted to *mean* independence of the residual, rather than full stochastic independence. Thus, we would reject the identifying assumptions if there is a strong linear association, but would fail to reject even if there is, for example, meaningful evidence of heteroskedasticity. This approach is similar in spirit to tests for over-identifying restrictions in IV models (see, for example, Kolesár et al., 2015).

Figure 3 shows studentized residual plots corresponding to the unadjusted and simple adjusted linear models (Equations 4.7 and 5.2) fit to the site-aggregate ECHS data shown in Figure 1. As indicated by the blue best-fit line for each residual plot, there is no strong positive or negative linear

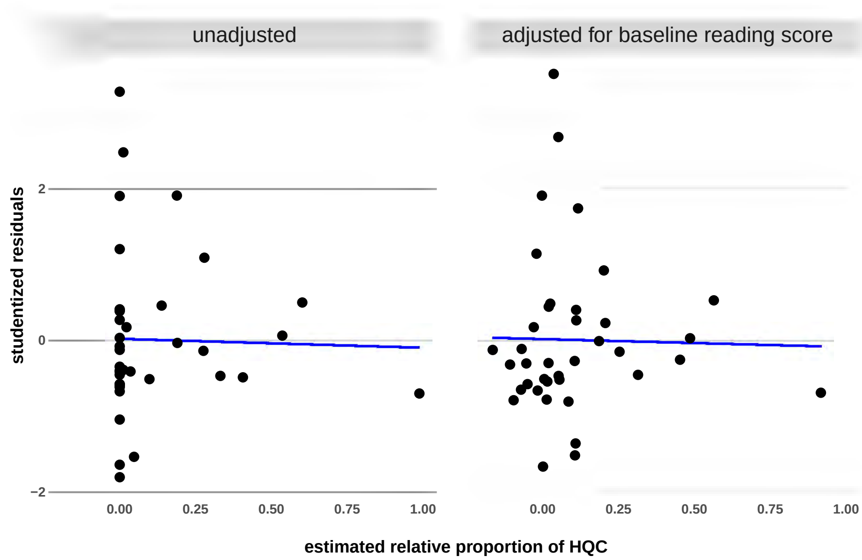

Fig (3) **Residual plots.** Studentized residuals versus estimated proportion of High-Quality Compliers for the Naive LATE model, where there is no baseline covariate adjustment (left panel) and where there is regression adjustment for eighth grade reading score (right panel). The blue lines are best-fit lines; one with a steep slope would indicate a violation of the (conditional) zero site-level correlation assumption needed to identify ITT_{lc} and ITT_{hc}.

pattern to the residuals, and the means of the residuals for each model are close to zero. Thus, there is no evidence against the identifying zero correlation assumptions, Assumptions 4.1 and 5.1. At the same time, the residual plots clearly invalidate a homogeneity assumption (Jiang, Ding and Geng, 2016) that the stratum-specific impacts are constant across sites, with large changes in the conditional variance of the residuals across $\widehat{\phi}_k$.

7. Connection to other methods Several approaches have the same setup as what we explore here, but rest on stronger assumptions. First, we can impose a stronger version of Assumption 4.1 by assuming that average impacts are *constant* across sites, rather than equal *in expectation* across sites. Specifically, instead of assuming $\mathbb{E}\left[\epsilon_{lc|k} \mid \phi_k\right] = 0$ for all k, we could instead require that $\epsilon_{lc|k} = 0$ for all k, or, equivalently, that $\text{ITT}_{lc|1} = \cdots = \text{ITT}_{lc|K}$. This clearly satisfies the requirements of Proposition 4.2, but is stronger than necessary for inference in our setting. Following the ecological inference literature, we refer to this as the *constancy* assumption; Gelman et al. (2001) provides a discussion of the *constancy* versus *zero correlation* assumptions in ecological regression. Jiang, Ding and Geng (2016) instead call this constancy assumption the homogeneity assumption; Wang, Zhou and Richardson (2017) relax this assumption by adjusting for baseline covariates; Kang et al. (2016) leverage this assumption to relax other requirements on possible effects.

One conceptual advantage of this constancy assumption is that we no longer need to posit the existence of a (hypothetical) super-population of sites. Instead, we can imagine sampling from an infinite super-population of individuals divided into K fixed sites. In fact, we no longer need multiple sites: the assumption of constant impacts could be applied to a single-site experiment where we imagine sampling from an infinite super-population of individuals divided into K fixed levels of any discrete covariate, such as grade level or racial group. In practice, the estimators for ITT_{lc} and ITT_{hc} would be the same as in Section 4.2, even though the underlying assumption is much stronger. See, for example, Hull (2018), who presents a similar setup as ours for a single site quasi-experiment with strata defined by a single (binary) covariate.[5]

The zero site-level correlation assumption we pose is also closely related to an important assumption in the multiple-site, multiple-mediator instrumental variables (MSMM-IV) literature. For a multi-site study in which a treatment may affect the outcome through multiple mediators, Reardon and Raudenbush (2013) delineate nine assumptions needed to identify the relevant causal effects using cross-site variation. Of the nine assumptions, the authors emphasize the critical assumption of *between-site compliance-effect independence*, in which the site-average compliance of each mediator is independent of the site Complier average effect of each mediator. This independence assumption is a closely related, but slightly stronger, version of the uncorrelatedness and marginal zero mean error conditions of Proposition 4.2.

Finally, we can re-frame much of the above discussion, such as Assumption 4.1, in terms of site-level *means* rather than site-level impacts. That is, we could assume that the site-specific mean outcome of Low-Quality Compliers assigned to treatment is uncorrelated with the site-specific relative share of Low-Quality Compliers. We view this as a slightly stronger assumption than what we propose. For example, it is conceivable that Low-Quality Complier students generally have less support and fewer resources that allow them to engage in academic activities, giving them a starting disadvantage compared to High-Quality Compliers. Thus, in schools with

[5]The core identifying assumption there is what Hull terms 'LATE homogeneity', which says stratum-specific LATEs are mean independent of the stratifying covariate.

a larger share of Low-Quality Compliers, students' academic performance under no intervention could be poorer, on average, than student academic performance at schools composed mostly of High-Quality Compliers. This scenario would violate a zero correlation in site-level means assumption. Assumption 4.1, on the other hand, permits control mean outcomes to co-vary with the relative proportion of Low-Quality Compliers in a site.

8. Conclusion The principal stratification literature largely focuses on randomized studies where there is only one experimental site. We extend this framework to the multi-site setting in the context of an evaluation of Early College High Schools and show how to identify and estimate key principal causal effects under a strong zero correlation assumption. We relax this assumption by incorporating auxiliary covariates and explore several issues that arise in estimation.

There are several directions for future work. The most important is to explore estimators that appropriately account for measurement error. First, we could adapt methods from the literature on multi-site, multi-mediator IV; specifically, Reardon et al. (2014) offer two bias-corrected instrumental variables estimators that could be extended to principal stratification. Second, we could further explore standard measurement error models or fully Bayesian hierarchical models as a way to simultaneously address both bias and sampling variance; Bloom et al. (2017) discuss relevant strategies in the multi-site setting, including under noncompliance.

Finally, it is useful to assess how to incorporate the zero correlation assumption into a broader principal stratification analysis, such as a bounds approach (Miratrix et al., 2018). Understanding the many possible identification and estimation approaches is increasingly important as more and more researchers use the principal stratification framework.

References

Allensworth, E. (2005). Graduation and Dropout Trends in Chicago: A Look at Cohorts of Students from 1991 through 2004. Report Highlights. *Consortium on Chicago School Research*.

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

Bloom, H. S., Raudenbush, S. W., Weiss, M. J. and Porter, K. (2017). Using multisite experiments to study cross-site variation in treatment effects: A hybrid approach with fixed intercepts and a random treatment coefficient. *Journal of Research on Educational Effectiveness* **10** 817–842.

Ding, P., Geng, Z., Yan, W. and Zhou, X.-H. (2011). Identifiability and estimation of causal effects by principal stratification with outcomes truncated by death. *Journal of the American Statistical Association* **106** 1578–1591.

Edmunds, J. A., Bernstein, L., Glennie, E., Willse, J., Arshavsky, N., Unlu, F., Bartz, D., Silberman, T., Scales, W. D. and Dallas, A. (2010). Preparing Students for College: The Implementation and Impact of the Early College High School Model. *Peabody Journal of Education* **85** 348-364.

Edmunds, J. A., Bernstein, L., Unlu, F., Glennie, E., Willse, J., Smith, A. and Arshavsky, N. (2012). Expanding the Start of the College Pipeline: Ninth-Grade Findings From an Experimental Study of the Impact of the Early College High School Model. *Journal of Research on Educational Effectiveness* **5** 136-159.

Edmunds, J. A., Unlu, F., Glennie, E., Bernstein, L., Fesler, L., Furey, J. and Arshavsky, N. (2017). Smoothing the Transition to Postsecondary Education: The Impact of the Early College Model. *Journal of Research on Educational Effectiveness* **10** 297-325.

Feller, A., Grindal, T., Miratrix, L., Page, L. C. et al. (2016a). Compared to what? Variation in the impacts of early childhood education by alternative care type. *The Annals of Applied Statistics* **10** 1245–1285.

Feller, A., Greif, E., Miratrix, L. and Pillai, N. (2016b). Principal stratification in the Twilight Zone: Weakly separated components in finite mixture models. Arxiv 1602.06595.

Frangakis, C. E. and Rubin, D. B. (2002). Principal Stratification in Causal Inference. *Biometrics,* **58** 21-29.

Gelman, A., Park, D. K., Ansolabehere, S., Price, P. N. and Minnite, L. C. (2001). Models, assumptions and model checking in ecological regressions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **164** 101–118.

Hull, P. (2018). IsoLATEing: Identifying Counterfactual-Specific Treatment Effects with Cross-Stratum Comparisons. Working paper.

Jiang, Z., Ding, P. and Geng, Z. (2016). Principal causal effect identification and surrogate end point evaluation by multiple trials. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **78** 829–848.

Jo, B. (2002). Estimation of Intervention Effects with Noncompliance: Alternative Model Specifications. *Journal of Educational and Behavioral Statistics* **27** 385–409.

Kang, H., Zhang, A., Cai, T. T. and Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American Statistical Association* **111** 132–144.

Kline, P. and Walters, C. R. (2016). Evaluating public programs with close substitutes: The case of Head Start. *The Quarterly Journal of Economics* **131** 1795–1848.

Kolesár, M., Chetty, R., Friedman, J., Glaeser, E. and Imbens, G. W. (2015). Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics* **33** 474–484.

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics* **29** 305–325.

Mealli, F., Pacini, B. and Stanghellini, E. (2016). Identification of principal causal effects using additional outcomes in concentration graphs. *Journal of Educational and Behavioral Statistics* **41** 463-480.

Miratrix, L., Furey, J., Feller, A., Grindal, T. and Page, L. C. (2018). Bounding, An Accessible Method for Estimating Principal Causal Effects, Examined and Explained. *Journal of Research on Educational Effectiveness* **11** 133-162.

Page, L. C., Feller, A., Grindal, T., Miratrix, L. and Somers, M.-A. (2015). Principal stratification: A tool for understanding variation in program effects across endogenous subgroups. *American Journal of Evaluation* **36** 514–531.

Peck, L. R. (2003). Subgroup Analysis in Social Experiments: Measuring Program Impacts Based on Post-Treatment Choice. *American Journal of Evaluation* **24** 157–187.

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

Raudenbush, S. and Schwartz, D. (2017). Identification and Estimation in Multisite Randomized Trials with Heterogeneous Treatment Effects. *submitted*.

Reardon, S. F. and Raudenbush, S. W. (2013). Under what assumptions do site-by-treatment instruments identify average causal effects? *Sociological Methods & Research* **42** 143–163.

Reardon, S. F., Unlu, F., Zhu, P. and Bloom, H. S. (2014). Bias and bias correction in multisite instrumental variables analysis of heterogeneous mediator effects. *Journal of Educational and Behavioral Statistics* **39** 53–86.

Rubin, D. B. (1980). Comment on "Randomization Analysis of Experimental Data: The Fisher Randomization Test". *Journal of the American Statistical Association* **75** 591–593.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Wang, L., Zhou, X.-H. and Richardson, T. S. (2017). Identification and estimation of causal effects with outcomes truncated by death. *Biometrika* **104** 597–612.