The Annals of Applied Statistics
2019, Vol. 13, No. 3, 1348–1369
https://doi.org/10.1214/18-AOAS1235

IDENTIFYING AND ESTIMATING PRINCIPAL CAUSAL EFFECTS IN A MULTI-SITE TRIAL OF EARLY COLLEGE HIGH SCHOOLS[1]

BY LO-HUA YUAN, AVI FELLER, AND LUKE W. MIRATRIX

Airbnb, Inc., University of California, Berkeley and Harvard University

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 quality, 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 independent of 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? 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

Received March 2018; revised December 2018.

[1]Supported by 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.

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

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 under-represented 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 the program 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 *between-site distribution-effect independence assumption*: the site-specific distribution of principal strata, for example, the proportion of Compliers, is (mean) independent of the site-specific impacts for these principal strata (Reardon and Raudenbush (2013)). This is a very strong assumption, roughly implying that the interaction between randomization and site indicator functions as a "second instrument" (the first being the treatment randomization itself) that is predictive of principal stratum membership, but is independent of the treatment impact within any stratum. We describe this independence assumption and the corresponding estimation in the context of principal stratification in the ECHS study. We also weaken this assumption such that independence only needs to hold conditional on a set of auxiliary covariates. Finally, we discuss how this assumption arises naturally in multi-site trials as compared to more general stratified randomized trials. Specifically, we appeal to the idea of sampling sites from some underlying population, in this case the population of high schools in North Carolina.

To the best of our knowledge, this is the first paper that brings together the otherwise disparate literatures of multi-site trials and principal stratification. We mention several relevant papers, and explore the connections in more depth in Section 7. First, the approach we outline here has the same form as the Multi-Site, Multi-Mediator Instrumental Variable (MSMM-IV) framework of Reardon and Raudenbush (2013), though some underlying assumptions and concepts differ.

As we highlight in the Supplementary Material, our key independence assumption maps directly to the same critical assumption in the MSMM-IV framework (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." 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. Many other papers impose similar restrictions on covariates to identify principal causal effects, including Jo (2002), Peck (2003), Ding et al. (2011), and Mealli, Pacini and Stanghellini (2016). Fourth, Bowden, Davey Smith and Burgess (2015) apply so-called *Egger regression* (Egger et al. (1997)) to meta-analyses of Mendelian randomization, which has a similar form to the setting we consider. Fifth, 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. Finally, this approach also has deep links to ecological regression (Gelman et al. (2001)) and ASPES (Peck (2003)). We believe we are the first to connect MSMM-IV and related methods to these other areas.

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 Material (Yuan, Feller and Miratrix (2019)) contain implementation details, an extensive simulation study, and additional discussion of other methods.

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.

We analyze data from the Evaluation of Early College High Schools in North Carolina (Edmunds et al. (2010)). The Early College programs in the study were over-subscribed and allocated slots to applicants via lottery, creating de-facto randomized trials. Overall, the study tracked a sample of 4004 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. Lastly, 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 Supplementary Material, which yields nearly identical conclusions.

Given these inclusion criteria, our final ECHS analysis sample consists of 3477 students ($N_t = 2021$ won an ECHS voucher, $N_c = 1456$ did not win an ECHS voucher) across 38 lotteries in 18 ECHS schools, each school with up to six 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 relative availability of high versus low quality alternative high schools varied from year to year, which complicates analyses that pool across lotteries.

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)).[2]

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 Material 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

[2]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.

TABLE 1
Distribution of high school type by treatment status

School type	Treatment ($N_t = 2021$)	Control ($N_c = 1456$)
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%

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.[3] While the state also rates Early Colleges as either low- or high-quality, we categorize ECHSs separately since they are the school type of interest.

Table 1 shows the distribution of ninth grade students in our data set across these three school types. In the treatment group, 85% of students attended an ECHS; 2% attended a high-quality school; 12% attended a low-quality school. In the control group, only 3% percent were able to cross over and register in an ECHS; 12% attended a high-quality school; 85% attended a low-quality school.

As we discuss next, the goal of our analysis is to explore treatment effect variation based on the quality level of the high school a student would attend if she does *not* enroll in an ECHS. In many settings, a student's traditional high school—and

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

thus the quality level of that high school—is fixed prior to randomization. For instance, if there is only a single alternative high school in each region, there would be no uncertainty in counterfactual school quality and standard subgroup analysis would be sufficient for assessing treatment effect variation. This is insufficient in our setting, however. In the sample we consider, students often applied to attend an ECHS program for which they were not zoned, and, more generally, have some flexibility in choosing high schools (Edmunds et al. (2012)). Thus, simply using the default fails to capture important heterogeneity. Even in cases where student school choice is fixed in practice, we do not necessarily have access to this information, which again makes standard subgroup estimation impossible. The principal stratification framework is well suited to this formulation: as we discuss next, we regard principal stratum membership as a quantity that is fixed but unknown at baseline, mirroring more traditional analyses of treatment effect variation. Thus, principal stratification is a moderation analysis on partially observed (or latent) subgroups.

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, that is, wins the lottery and is invited to enroll in an ECHS. Let Y_i^{obs} denote student i's observed outcome, that is, 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. 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

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

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.

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

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 ($Z_i = 0$)		
ECHS offer ($Z_i = 1$)	$D_i(0) = e$	$D_i(0) = lq$	$D_i(0) = hq$
$D_i(1) = e$	**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**

Table 2 shows the $3^2 = 9$ possible principal strata; rows indicate school type attended 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)*. As with other lottery studies, this assumption seems reasonable in the context of ECHS, where students have little incentive to directly counteract the randomization. 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, also known as an independence of irrelevant alternatives assumption. This would be violated if, for example, a family loses the ECHS lottery and then updates their preferences about the relative strengths of traditional high schools, such as prioritizing instructional quality over logistical considerations. While possible, this is unlikely in practice. Thus, excluding these categories is a sensible simplifying 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 Material, we can use these assumptions to identify the distribution of principal strata, π_s, for $s \in \{\text{eat}, \text{lc}, \text{hc}, \text{lat}, \text{hat}\}$.

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,

$$\text{ITT}_{\text{eat}} = \text{ITT}_{\text{lat}} = \text{ITT}_{\text{hat}} = 0.$$

The logic here is identical to the simpler noncompliance setting: since randomization has no impact on school quality for students in these groups—likely because students are not induced to change schools—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:

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

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

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

where $\phi = \frac{\pi_{\text{hc}}}{\pi_{\text{lc}}+\pi_{\text{hc}}}$ 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. 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. Mealli, Pacini and Stanghellini (2016) explore the use of multiple outcomes and other covariate restrictions. Kline and Walters (2016) identify these effects by imposing restrictions on the selection process.

4. Identification and estimation via between-site independence. 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 Material.

4.1. *Super-population of sites and the between-site independence 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 $\text{ITT}_{s|k} = \mathbb{E}[Y_i(1) - Y_i(0) | S_i = s, X_i = k]$ 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 $\phi_k = \pi_{\text{hc}|k}/(\pi_{\text{lc}|k} + \pi_{\text{hc}|k})$ denote the proportion of Compliers in site k who are of the 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, e.g., Kolesár et al. (2015)), though we prefer to focus explicitly on a super-population. Specifically, we assume that we sample sites represented as triples of parameters $(\text{ITT}_{\text{lc}|k}, \text{ITT}_{\text{hc}|k}, \phi_k)$ from an infinite super-population of sites with mean vector $(\text{ITT}_{\text{lc}}, \text{ITT}_{\text{hc}}, \phi)$ and a 3×3 covariance matrix $\mathbf{\Sigma}$:

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

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 reframe the main problem in terms of regression. First, rewrite equation (3.2) separately for each site, rearrange terms, and add zero twice to obtain

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

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

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

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

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 (mean) independent of the regressors, ϕ_k and $1-\phi_k$, in the super-population. Specifically, we can identify the regression coefficients under the assumptions that $\mathbb{E}[\epsilon_{\text{lc}|k} \mid \phi_k] = 0$ and $\mathbb{E}[\epsilon_{\text{hc}|k} \mid \phi_k] = 0$.

ASSUMPTION 4.1 (Independence between principal stratum distributions and principal causal effects). The site-specific relative share of High-Quality Compliers is (mean) independent of the site-specific impacts for High-Quality Compliers and for Low-Quality Compliers:

$$\mathbb{E}[\epsilon_{\text{lc}|k} \mid \phi_k] = 0 \quad \text{and} \quad \mathbb{E}[\epsilon_{\text{hc}|k} \mid \phi_k] = 0. \tag{4.4}$$

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. Assumption 4.1 is also equivalent to assuming that $\Sigma_{31} = \Sigma_{32} = 0$ in expression (4.1) together with a Normality assumption.

We combine all these assumptions into the following proposition.

PROPOSITION 4.2 (Identification of principal causal effects). *For a multi-site trial with $K \geq 2$ sites, under Assumption 4.1 and $\text{Var}(\phi_k) > 0$, the principal causal effects, ITT_{lc} and ITT_{hc}, are identified.*

The proof for Proposition 4.2 follows immediately from standard regression theory. See Reardon and Raudenbush (2013) for additional discussion. 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 between-site independence 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, and that the analogous assumption holds 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 independence assumption to hold conditional on covariates, such as prior academic preparedness, in Section 5. In the Supplementary Material, we show that site-specific stratum membership weakly varies with measures of middle school academic preparation, suggesting that the unconditional independence assumption may be violated in practice.

Finally, it is useful to reframe this setup in terms of the contrast $\text{ITT}_{\text{hc}} - \text{ITT}_{\text{lc}}$. We can rewrite equation (4.3) to highlight this directly:

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

This yields a particularly simple form when there are only two sites, 1 and 2:

$$\text{ITT}_{\text{hc}} - \text{ITT}_{\text{lc}} = \frac{\text{LATE}_1 - \text{LATE}_2}{\phi_1 - \phi_2}. \tag{4.6}$$

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 the connection to using the interaction of "site by randomization" as an additional instrument. This simple form reveals a danger of this identification approach: if the sites do not differ substantially in the proportion of High-Quality Compliers, then estimates for this treatment impact contrast will be unstable, akin to the weak instrument problem (Kolesár et al. (2015), Raudenbush, Reardon and Nomi (2012)). See the Supplementary Material 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, $\widehat{\text{LATE}}_k$ and $\widehat{\phi}_k$. Let $\widehat{Y}_{\text{zd}} = \frac{1}{N_{\text{zd}}} \sum_{i \in \{Z_i = z, D_i^{\text{obs}} = d\}} Y_i^{\text{obs}}$ be the finite sample average observed outcome for students assigned to $Z_i = z$ with observed take up $D_i^{\text{obs}} = d$, and let $\widehat{Y}_{\text{zd}|k}$ 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 Material for details.) We then estimate the site-specific LATE as

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

where $\widehat{\pi}_{\text{lc}|k} + \widehat{\pi}_{\text{hc}|k}$ is the estimated proportion of Compliers in site k.[4] We can also estimate the relative proportion of High-Quality Compliers in site k:

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

[4]In applications where the site-specific proportion of Compliers is small, researchers might instead use the ITT parameterization in Equation (3.1). See the Supplementary Materials.

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

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

where $\widehat{\beta}_{\text{lc}}$ and $\widehat{\beta}_{\text{hc}}$ are estimators for ITT_{lc} and ITT_{hc}, respectively. Taking the site-specific estimates, $\widehat{\text{LATE}}_k$ 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 $\widehat{\text{LATE}}_k$ and $\widehat{\phi}_k$ are estimated rather than known. This leads to two key complications. First, conventional estimates of the standard error will under-estimate the true sampling variance. Second, the nominal point estimates could be biased; in particular, error in $\widehat{\phi}_k$ will attenuate the estimate of $\text{ITT}_{\text{hc}} - \text{ITT}_{\text{lc}}$. To account for the increased uncertainty due to measurement error, we 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 recalculate $\widehat{\text{LATE}}_k^*$ 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 Material) show that this procedure has meaningfully smaller RMSE than the naïve 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. Site sizes typically vary in practice, which introduces additional complications. Specifically, the super-population means $(\text{ITT}_{\text{lc}}, \text{ITT}_{\text{hc}}, \phi)$ 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. And if site impact varies with precision, then the estimate can again be biased. If, in addition to the conditions listed in Proposition 4.2, we also assert that $\text{ITT}_{\text{lc}|k}$ and $\text{ITT}_{\text{hc}|k}$ are independent of N_k, the (estimated) 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 independence. 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 independence 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, or the total number of Compliers in a site. We can then relax the independence assumption such that it only holds *conditionally*:

$$\mathbb{E}[\epsilon_{\mathrm{lc}|k} \mid \phi_k, \mathbf{W}_k] = 0 \quad \text{and} \quad \mathbb{E}[\epsilon_{\mathrm{hc}|k} \mid \phi_k, \mathbf{W}_k] = 0. \tag{5.1}$$

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

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

with $\mathbf{W}_k$ as above and with $\boldsymbol{\zeta}$ as the corresponding vector of regression coefficients for $\mathbf{W}_k$.

The simple regression-adjusted model, however, restricts the possible treatment effect variation; see Supplementary Material 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

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

where appropriate combinations of $\widehat{\beta}_s^{\mathrm{int}}$ and $\widehat{\delta}_s$ yield estimates of the site-average impacts. Here $\mathbf{W}_{-1,k}$ is $\mathbf{W}_k$ with the $W_{1,k}$ covariate removed; $\boldsymbol{\gamma}$ is the corresponding vector of regression coefficients for $\mathbf{W}_{-1,k}$.

When site sizes vary, we can reweight the regression coefficient estimates from equations (5.2) or (5.3) to obtain population-average impacts under the assumption that $\mathrm{ITT}_{\mathrm{lc}|k}$ and $\mathrm{ITT}_{\mathrm{hc}|k}$ are *conditionally* independent of N_k, given $\mathbf{W}_k$. For High-Quality Compliers, we have the following weighted average:

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

with an analogous estimate for Low-Quality Compliers.

When the covariate space is high-dimensional and conditioning on the full vector of **W** is infeasible, one may instead condition on a balancing score, namely, the expected principal score across sites, as a way to satisfy the conditional independence assumption (Yuan (2018)).

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 same assumption 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.[5]

As described in the Supplementary Material, we estimate impacts three ways: without covariate adjustment, with a 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 $\widehat{\phi}_k = 0$, meaning 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 anticipate more precise estimates of $\mathrm{ITT}_{\mathrm{lc}}$ than $\mathrm{ITT}_{\mathrm{hc}}$.

Figure 2 shows the corresponding point estimates and 95% confidence intervals for $\mathrm{ITT}_{\mathrm{lc}}$ and $\mathrm{ITT}_{\mathrm{hc}}$. 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 $\mathrm{ITT}_{\mathrm{lc}}$; nor is there a meaningful difference between the naïve and bootstrap point estimates. Reading

[5]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.

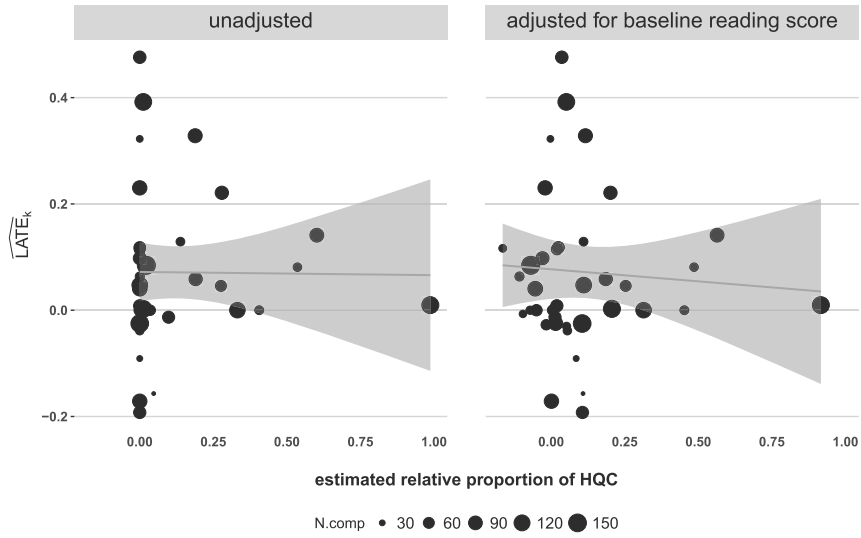

FIG. 1. *ECHS site-level data. Scatterplots of estimated site-specific Complier impacts* (*proportion on-track*) *versus* (*left panel*) *estimated relative proportion of High-Quality Compliers in a site, and* (*right panel*) *estimated* residual *relative proportion of High-Quality Compliers in a site,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

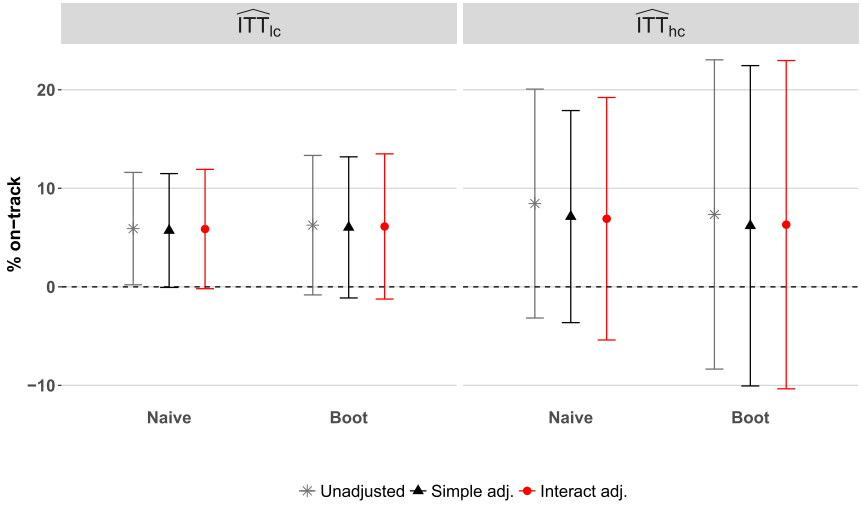

FIG. 2. *Estimates of principal causal effects. Point estimates and* 95% *confidence intervals for Low- and High-Quality Complier principal causal effects are plotted for each estimation method.*

score adjustment has a more noticeable effect on point estimates for ITT_{hc}, with $\widehat{\text{ITT}}_{\text{hc}}$ decreasing by about 1.3 percentage points under both simple linear adjustment and interaction adjustment.

The standard errors for $\widehat{\text{ITT}}_{\text{hc}}$ are considerably larger than those for $\widehat{\text{ITT}}_{\text{lc}}$, which reflects the fact that the estimated number of Low-Quality Compliers is roughly seven times larger than the estimated number of High-Quality Compliers. The standard errors for both $\widehat{\text{ITT}}_{\text{lc}}$ and $\widehat{\text{ITT}}_{\text{hc}}$ 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 method is roughly 23% and 40% wider than the CI of the corresponding naïve estimate. This aligns with our simulation study finding that the bootstrap method produces overly conservative confidence intervals. Although we do not include 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 is a meaningful difference between ITT_{hc} and ITT_{lc} using the reparameterization in equation (4.5), also shown graphically in Figure 1. While we do not find a significant difference in stratum impacts for High- vs Low-Quality Compliers, this is under a scenario of limited power due to the relatively small share of High-Quality Compliers.

Overall, we find that the estimated impacts are quite similar for Low- and High-Quality Compliers and that these estimates are stable across different models. We interpret these estimates with caution, however. As we discuss above, measurement error can meaningfully attenuate differences in the estimated impacts between the two groups. Also, there is considerable uncertainty around the impact for High-Quality Compliers. In the Supplementary Material, we conduct a more detailed power analysis, finding that there is low power for estimating impacts for this comparatively small group. Thus, in this context, we can only draw limited conclusions for the ECHS evaluation.

Our findings differ slightly from the treatment effect bounds obtained by Miratrix et al. (2018), who ignore the multi-site study design. Like us, Miratrix et al. (2018) find evidence of a positive effect of ECHS on the on-track status of Low-Quality Complier students. However, whereas Miratrix et al. (2018) also estimate very wide bounds for the impact of ECHS on High-Quality Compliers, our analysis is more suggestive of a positive impact than theirs. See the Supplementary Material for additional comparisons.

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 between-site independence Assumption 4.1 implies that the fitted residuals should be mean independent of the site-specific proportion of High-Quality Compliers. While power might be limited, we can use the fitted residuals from the site-level regression to directly assess the evidence against these assumptions. Importantly, the relevant assumption is restricted to *mean* independence of

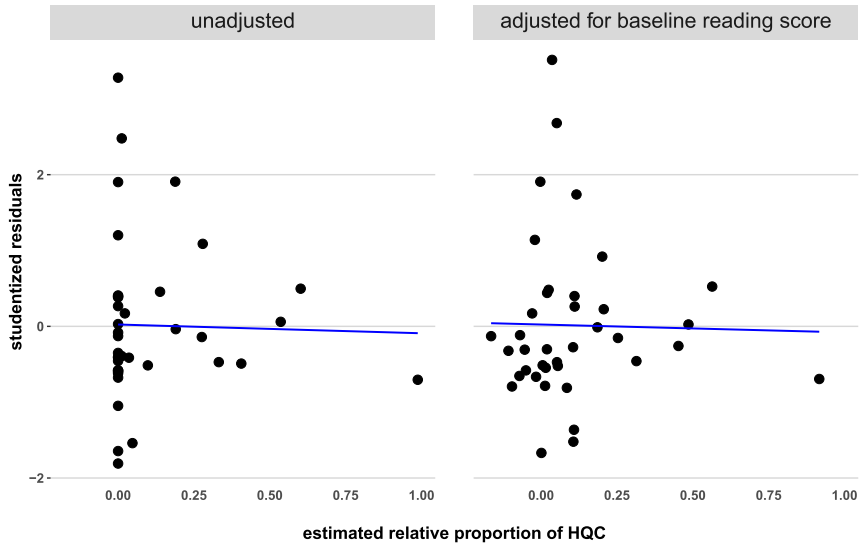

FIG. 3. *Residual plots. Studentized residuals versus estimated proportion of High-Quality Compliers for the Naive LATE model, where there is no baseline covariate adjustment* (*left panel*) *and where there is regression adjustment for eighth grade reading score* (*right panel*). *The darker lines are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

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, e.g., 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 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 independence assumptions, Assumption 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 $\widehat{\phi}_k$.

While straightforward, this approach to model checking has some limitations. Specifically, measurement error at the site level could increase the variability in the residuals, thereby decreasing power to detect violations of the independence assumption. One alternative is the so-called tomography plot, which is common in ecological inference (Gelman et al. (2001)). More broadly, Carnegie, Harada and

Hill (2016) propose a model-based approach to assessing sensitivity to unmeasured confounding that could be adapted to this setting.

7. Connection to other methods. The approach we explore here has connections to a broad range of other methods. First, the between-site independence Assumption 4.1 maps directly to the *between-site compliance-effect independence assumption* in the multiple-site, multiple-mediator instrumental variables (MSMM-IV) literature (Reardon and Raudenbush (2013)). The Supplementary Material include a detailed comparison of the assumptions necessary for principal stratification versus mediation in this setting. While the underlying quantities of interest differ, both our approach and MSMM-IV crucially rely on between-site variation for identification and estimation.

Second, 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 $\mathbb{E}[\epsilon_{s|k} \mid \phi_k] = 0$ for all k, we could instead require that $\epsilon_{s|k} = 0$ for all k, or, equivalently, that $\text{ITT}_{s|1} = \cdots = \text{ITT}_{s|K}$ for $s = \text{lc}$ and $s = \text{hc}$. This clearly satisfies the requirements of Proposition 4.2, but is stronger than necessary for identification in our setting. Following the ecological inference literature, we refer to this as the *constancy* assumption; see Gelman et al. (2001) for additional discussion. 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 could 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.[6]

Lastly, we can reframe 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 independent of 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 possible that, in schools with a larger share of Low-Quality Compliers,

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

the academic performance of Low-Quality Compliers under no intervention could be lower, on average, than the academic performance of Low-Quality Compliers at schools with a small share of this group. This scenario would violate an assumption that site-level means are independent of site-level distributions. Assumption 4.1, on the other hand, permits control mean outcomes to depend on the relative proportion of Low-Quality Compliers in a site.

8. Discussion. 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 between-site independence assumption. We relax this assumption by incorporating auxiliary covariates and explore several issues that arise in estimation, including quantifying uncertainty and model checking.

Consistent with the original experimental analysis, we find that enrolling in ECHS has a positive impact on students' ninth-grade on-track status, a proxy for high school graduation. We fail to find differential impacts based on whether students would otherwise attend a low-quality or high-quality traditional high school. Statistical power, however, is limited overall. Specifically, there are relatively few students who would otherwise attend high-quality high schools, which complicates the analysis. Thus, the approach of exploiting between-site variation to estimate principal causal effects is of general interest but yields limited conclusions in the ECHS evaluation.

More broadly, the method we explore here has some drawbacks. First, measurement error is a primary concern, attenuating the effect estimates and reducing power, especially with many small sites. While we propose a bootstrap approach for incorporating uncertainty, addressing measurement error is an important direction for future work. In particular, in the context of multi-site, multi-mediator IV, Reardon et al. (2014) propose two bias-corrected instrumental variables estimators that could be extended to principal stratification. We could also 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. These same issues can also arise with large sites if the principal strata of interest are rare, such that there is a "weak" instrument within each site. Raudenbush, Reardon and Nomi (2012) and Kolesár et al. (2015) suggest possible paths forward in this "many weak instruments" setting. Rather than give general advice, we recommend that researchers conduct simulation studies calibrated to their specific settings, analogous to the simulations we perform in the Supplementary Material.

Second, while a strength of the principal stratification framework is that it imposes relatively few assumptions, especially compared to mediation (Page et al. (2015)), the corresponding conclusions are typically weaker than those from mediation. For instance, the Low- and High-Quality Compliers in our application differ

across both observable and unobservable characteristics. Thus, if we had observed differences in impacts between these groups, we could not solely attribute those differences to counterfactual school quality. Researchers interested in exploring these mechanisms should instead use a mediation framework as in Reardon and Raudenbush (2013). Further comparing these approaches in a multi-site setting is a promising research direction.

Finally, it is useful to assess how to incorporate the between-site independence 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.

SUPPLEMENTARY MATERIAL

Supplement to: "Identifying and estimating principal causal effects in a multi-site trial of Early College High Schools" (DOI: 10.1214/18-AOAS1235SUPP; .pdf). The Supplementary Material includes additional analyses, proofs and other technical materials.

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. Available at https://files.eric.ed.gov/fulltext/ED486035.pdf.

ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **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.

BOWDEN, J., DAVEY SMITH, G. and BURGESS, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44** 512–525.

CARNEGIE, N. B., HARADA, M. and HILL, J. L. (2016). Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness* **9** 395–420.

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. *J. Amer. Statist. Assoc.* **106** 1578–1591. MR2896858

EDMUNDS, J. A., BERNSTEIN, L., GLENNIE, E., WILLSE, J., ARSHAVSKY, N., UNLU, F., BARTZ, D., SILBERMAN, T., SCALES, W. D. et al. (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.

EGGER, M., SMITH, G. D., SCHNEIDER, M. and MINDER, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ* **315** 629–634.

FELLER, A., GRINDAL, T., MIRATRIX, L. and PAGE, L. C. (2016a). Compared to what? Variation in the impacts of early childhood education by alternative care type. *Ann. Appl. Stat.* **10** 1245–1285. MR3553224

FELLER, A., GREIF, E., MIRATRIX, L. and PILLAI, N. (2016b). Principal stratification in the Twilight Zone: Weakly separated components in finite mixture models. Available at arXiv:1602.06595.

FRANGAKIS, C. E. and RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58** 21–29. MR1891039

GELMAN, A., PARK, D. K., ANSOLABEHERE, S., PRICE, P. N. and MINNITE, L. C. (2001). Models, assumptions and model checking in ecological regressions. *J. Roy. Statist. Soc. Ser. A* **164** 101–118. MR1819025

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. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 829–848. MR3534352

JO, B. (2002). Estimation of intervention effects with noncompliance: Alternative model specifications. *J. Educ. Behav. Stat.* **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. *J. Amer. Statist. Assoc.* **111** 132–144. MR3494648

KLINE, P. and WALTERS, C. R. (2016). Evaluating public programs with close substitutes: The case of head start. *Q. J. Econ.* **131** 1795–1848.

KOLESÁR, M., CHETTY, R., FRIEDMAN, J., GLAESER, E. and IMBENS, G. W. (2015). Identification and inference with many invalid instruments. *J. Bus. Econom. Statist.* **33** 474–484. MR3416595

MACKINNON, J. G. and WHITE, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *J. Econometrics* **29** 305–325.

MEALLI, F., PACINI, B. and STANGHELLINI, E. (2016). Identification of principal causal effects using additional outcomes in concentration graphs. *J. Educ. Behav. Stat.* **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. W., REARDON, S. F. and NOMI, T. (2012). Statistical analysis for multisite trials using instrumental variables with random coefficients. *Journal of Research on Educational Effectiveness* **5** 303–332.

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? *Sociol. Methods Res.* **42** 143–163. MR3190727

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. *J. Educ. Behav. Stat.* **39** 53–86.

RUBIN, D. B. (1980). Comment on "Randomization analysis of experimental data: The Fisher randomization test". *J. Amer. Statist. Assoc.* **75** 591–593.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. MR0899519

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. MR3694585

YUAN, L.-H. (2018). Regressions for estimating main and principal causal effects in multi-site randomized trials and small sample designs. Doctoral dissertation. Available at https://dash.harvard.edu/handle/1/40050155.

YUAN, L.-H., FELLER, A. and MIRATRIX, L. (2019). Supplement to "Identifying and Estimating Principal Causal Effects in a Multi-site Trial of Early College High Schools." DOI:10.1214/18-AOAS1235SUPP.

LO-HUA YUAN
AIRBNB, INC.
888 BRANNAN ST.
SAN FRANCISCO, CALIFORNIA 94103
E-MAIL: lohuayuan@alumni.harvard.edu

A. FELLER
GOLDMAN SCHOOL OF PUBLIC POLICY
UNIVERSITY OF CALIFORNIA, BERKELEY
2607 HEARST AVE.
BERKELEY, CALIFORNIA 94720
E-MAIL: afeller@berkeley.edu

L. W. MIRATRIX
GRADUATE SCHOOL OF EDUCATION
HARVARD UNIVERSITY
14 APPIAN WAY
CAMBRIDGE, MASSACHUSETTS 02138
E-MAIL: lmiratrix@g.harvard.edu