Identification and Counterfactuals for Program Evaluation of Career and Technical Education

SEPTEMBER 2020

Stephen L. Ross, University of Connecticut | Eric Brunner, University of Connecticut | Rachel Rosen, MDRC

September 2020

The work of the CTE Research Network Lead is supported by the Institute of Education Sciences (IES) at the U.S. Department of Education with funds provided under the Carl D. Perkins Career and Technical Education Act through Grant R305N180005 to the American Institutes for Research (AIR). The content of this publication and the opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Ross, S. L., Brunner, E., & Rosen, R. (2020). *Identification and counterfactuals for program evaluation of career and technical education.* Arlington, VA: American Institutes for Research, Career and Technical Education Research Network. Retrieved from https://cteresearchnetwork.org/resources/cte-evaluation-counterfactuals

This report is available on the CTE Research Network website at https://cteresearchnetwork.org/resources/cte-evaluation-counterfactuals.

Abstract

This paper considers recent efforts to conduct experimental and quasi-experimental evaluations of career and technical education programs. It focuses on understanding the counterfactual, or control population, for these program evaluations, discussing how the educational experiences of the control population might vary from those of the treated population and the ways in which the treatment and control populations used for evaluation may differ from each other. The paper begins by discussing the key identification strategies and the associated assumptions used to identify program effects, including regression and propensity score matching, instrumental variables, regression discontinuity designs, randomized controlled trials, and lottery-based admissions. It then presents a series of case studies evaluating the counterfactual for specific studies that used each identification strategy.

Contents

Page

Section I. Introduction

Career and technical education (CTE) encompasses a broad array of educational programs that provide students with specialized skills in trades, applied sciences, and modern technologies, directly preparing these students for specific careers. Formerly referred to as vocational education, CTE has seen a resurgence of interest in recent years as an alternative pathway for K–12 students. In 2015, 11 million students enrolled in secondary or postsecondary CTE programming (Passarella, 2018). In 2018, the Carl D. Perkins Career and Technical Education Act was reauthorized, providing $1.2 billion in funding for CTE programs and job training for students. Even within secondary education, these programs vary considerably, from stand-alone CTE-focused high schools, to programs within traditional high schools, to programs where high school students receive training outside the regular high school facility, often in cooperation with local community colleges. Furthermore, in addition to a sequence of CTE courses, CTE often is a multifaceted experience for students, including elements such as work-based learning, associated academic coursework, career exploration, and preparation for industry certification and licensure examinations, to name a few.

Interpreting the results of program evaluations can be quite challenging in the context of CTE because of the following: (a) the broad opportunities available for selection into and out of largely voluntary programs, (b) the variety of contexts in which CTE is delivered, and (c) the multiple avenues often available to students for obtaining CTE experiences. First (and foremost), rigorous program evaluations require a comparison of students in a CTE program to control group students who did not participate in that program. Carefully identifying the specific set of students that operates as a control group—and understanding the ways and reasons why these control group students might differ systematically from the students in the CTE program—is critical to evaluating and interpreting estimated treatment effects (i.e., assessing internal validity). With most CTE delivered at the level of individual high schools, schools typically adjust capacity according to the historic and current demand for electives, with access to limited electives managed in a decentralized process. Students who select CTE options from their high school offerings may be different from students who did not make these selections along both observable (e.g., interest in CTE) and unobservable (e.g., grit or determination) dimensions. Furthermore, when access to electives is rationed, treatment students may be selected based on the private knowledge of teachers or administrators about a student's aptitude and motivation. Finally, the heterogeneity of these programs can make it difficult to identify the places where the program can have an impact. Some programs focus on career readiness and may reduce college attendance as students transition to employment, whereas other programs that focus on science, technology, engineering, or mathematics often improve students' preparation for higher education.

Second, the heterogeneous contexts within which CTE is offered raise significant concerns about the generalizability or external validity of any study results. As noted previously, most students who participate in CTE classes do so by taking electives within their home high school. On the other hand, most opportunities for identifying the causal effect of CTE on student outcomes arise in the context of career academies or stand-alone career high schools that often face demand that exceeds the supply of seats in the program or school. Therefore, research has provided extremely limited evidence on the effectiveness of CTE when offered as electives within a traditional high school curriculum. Further, the selection of students into a program might influence the external validity of any findings if the treatment effects are heterogeneous within the population and the analysis design systematically selects a sample of specific types of students to receive treatment. For example, if CTE attracts a population of students that is especially motivated to pursue CTE, then expanding CTE options might attract a broader population of students. In that case, the effects of expanding the existing program might be substantially smaller than evaluation estimates based on the current population of highly motivated students.

Third, researchers need to understand the experiences of the control group relative to students in the CTE program being evaluated. To what extent do control group students have the opportunity to take advantage of other CTE experiences outside the program being evaluated? Many traditional high schools offer some type of CTE programming, and most students take at least one CTE course. Do the opportunities and environment of control group students reflect the environment that program participants would have experienced if they had not been in the program (i.e., the counterfactual)? For example, CTE program participants might be more likely to seek out other CTE opportunities. If the program is outside a traditional high school, then program participants might be more likely to participate in other school choice opportunities. To take an extreme example, consider two separate, well-identified CTE program evaluations that both have strong internal validity. In the first program, if students were not selected to participate in the CTE program, they returned to the traditional high school that offered few (if any) CTE experiences. In the second program, if students were not selected to participate in the CTE program, they had the opportunity through school choice or a related policy to attend a very similar school that also offered extensive CTE experiences. Under this scenario, even if both CTE programs were successful at increasing the educational and labor market outcomes of participants, we might observe positive treatment effects for only the first program. This outcome occurs even though the evaluations of both programs provide unbiased estimates of treatment effects. In this case, what really matters is what defines treatment relative to the counterfactual.

Further, the experiences of the control group might not reflect the experiences that the treatment group would have received if there had been no program. In general, winning access to a selective program can have positive effects on students, whereas losing access may negatively affect attitude and persistence. Also, opportunities to experimentally evaluate a program often arise when new programs are created or services are expanded. In the case of CTE, students may not enroll in the limited number of CTE electives when attending a traditional high school, but the creation of a new CTE program at a high school may raise awareness about CTE and increase enrollment, leading the control group to change its behavior as well.

Fourth, CTE programs often involve a vector of treatments, such as greater exposure to CTE coursework, career-themed pathways, work-based learning and other work-related learning activities, exposure to different peers, and greater integration of core courses such as mathematics and English courses with CTE material. In most cases, researchers will be able to identify only the effect of the combined vector of treatments. Consequently, it often is difficult to identify exactly which elements of a CTE program impact outcomes such as high school graduation, college enrollment, and adult earnings. Further, researchers often cannot identify which members of the control group receive different components of the overall treatment outside the program, again making it difficult to interpret the results of a CTE program evaluation. In addition, CTE programs may have different impacts that depend on the career cluster or program of study that students choose. Obviously, understanding how outcomes such as adult earnings vary with a student's career cluster choice is extremely important from a policy perspective. However, providing convincing causal evidence on how the choice of a career cluster affects outcomes is extremely challenging because students are almost never randomly assigned to clusters. The fact that students typically choose the career cluster they wish to pursue raises obvious concerns about nonrandom sample selection into particular career clusters.

Although numerous studies have attempted to evaluate the impact of CTE programs (see Rosen, Visher, & Beal, 2018, for a review), few studies provide convincing quasi-experimental or experimental evidence on the causal effect of such programs on student outcomes, and no study has yet estimated the causal impact of the separate elements that may or may not be part of a CTE experience. Using lottery-based admissions, Kemple (2008) and Page (2012) examined labor market outcomes of students at nine different career academies, and Hemelt, Lenard, and Paeplow (2019) examined the effect of attending a single career academy in North Carolina. Dougherty (2018) studied three regional vocational and technical high schools in Massachusetts using a regression discontinuity (RD) strategy that compared similar students just above and just below admission cutoffs, and Brunner, Dougherty,

and Ross (2019) used the same approach to examine 16 stand-alone technical high schools in Connecticut. Finally, Cullen, Jacob, and Levitt (2005) estimated the effects of 10 large career academies in Chicago Public Schools (CPS) using an instrumental variables (IV) identification strategy. They used student residential proximity to career academies as an instrument for school choice (i.e., students who live closer to one of the academies end up attending those high schools at higher rates). Witzen (2019) used a similar IV approach based on driving time to a CTE center to estimate the effects of completing a CTE program on college attendance and earnings at 6 years after high school graduation and at the time of first employment.[1]

This paper examines the problem of understanding the counterfactual population that forms the basis of CTE program evaluations by taking a case study approach through the lens of typical methods used to study CTE program effects. Section II briefly discusses the basic assumptions required to identify the causal effect of a program for all four methods considered. In Section III, we select a paper that uses one (or more) of the four methods and discuss the relevant issues for evaluating each paper's findings. We begin our case studies with the most common and simple approach of controlling for observables, namely, regression and matching methods. There we focus on the well-known evaluation by SRI International (Warner et al., 2016) of the California Linked Learning District Initiative (LLDI). We then turn to IV identification strategies, with a particular focus on work by Cullen et al. (2005). For RD, we evaluate and discuss a recent paper by Brunner et al. (2019) that examined admission to stand-alone technical high schools in Connecticut. Finally, for randomized controlled trials (RCTs) and lottery-based admissions, we focus on MDRC's evaluation of career academies housed within nine individual high schools located across the United States (Kemple, 1997, 2004, 2008). We conclude with a summary and synthesis of the papers and methods we reviewed and a discussion of their implications for policy, practice, and future research.

Section II. Methods

This section describes the various research methods used to identify the causal effects of CTE on educational and labor market outcomes and how those methods influence the counterfactual condition. Specifically, we focus on four empirical research methods: (a) conditioning on key observables (e.g., regression or propensity score matching), (b) IV approaches, (c) RD designs, and (d) lottery-based admissions. Another common method, called difference-in-differences, examines the effects of creating a program by comparing changes across time between places implementing a program and similar places that do not implement a specific program. To our knowledge, however, difference-in-differences methods have not been applied within the CTE context.

We motivate these methods and their implications for describing the counterfactual condition using a potential outcome framework. Specifically, let Y_{0i} denote individual *i*'s potential outcome without treatment and Y_{1i} denote individual *i*'s potential outcome with treatment T_i. The average causal effect of treatment (i.e., participation in CTE) would then be the expected value or average difference between an individual's outcome with and without treatment:

$$E[Y_{1i} - Y_{0i}] = E[Y_{1i}] - E[Y_{0i}] \quad (1)$$

[1] In the European context, Silliman and Virtanen (2019) examined the effect of admission to the secondary school vocational track in Finland on adult earnings. Using an RD design that exploited the centralized admission process used for secondary education, they found that students admitted to the vocational track have annual earnings that are approximately 7% higher at age 31. Bertrand, Mogstad, and Mountjoy (2019) examined the implementation of a vocational program in Norway that increased the pathways to college and expanded the opportunities for apprenticeship. Using an RD design that exploited eligibility based on birth date, they found that access to the program increased social mobility, especially for men.

In practice, we can observe only one of these two outcomes. We use the potential outcome framework to motivate and describe how different empirical research methods identify a control group for assessing the effects of treatment on the treated population to help inform our discussion of the counterfactual experiences of this control group.

To simplify the exposition, we use a single dichotomous indicator to denote treatment. Additional indicators would be needed to account for the fact that CTE and non-CTE students may or may not experience, for example, career-themed pathways, work-based learning, associated academic coursework, or career exploration and planning. However, in almost all cases, the source of identifying information will usually isolate only one or a limited number of treatment domains, such as admission to an overall program based on meeting a threshold score or one or two alternative experimental treatments. Even in the case of propensity score matching, the matching on observables takes place for a single treatment, such as participation in a CTE program. As a result, we focus on the case where participation in a CTE program can be modeled as one overall encompassing treatment.

Propensity score matching methods are a common approach for using observational data to provide an estimate of

$$E\left[Y_{1i} - Y_{0j} | p(x_i) \approx p(x_j)\right] \quad (2)$$

where $p(x_i)$ is a function that predicts the likelihood of treatment based on observable attributes of the treatment and potential control groups. The control group for a specific treated individual *i* is other untreated individuals *j* who had the same (or close) predicted likelihood of participating in the treatment based on observables. This approach addresses bias that might arise because selection into the treatment based on observables might follow a very complex, nonlinear process. However, propensity score approaches do nothing to address the concern that unobservables correlated with outcomes of interest also might affect selection into treatment and thus bias the estimated treatment effects. In other words, under the propensity score matching approach, the counterfactual consists of students who are similar to treated students along observable dimensions but may look very different from treated students along unobservable dimensions such as interest in participating in CTE programs, motivation, or aptitude. Further, it is worth noting that if a linear model of x_i does a good job of predicting treatment, then the results of a propensity score analysis are likely to be very similar to ordinary least squares estimates that condition on the same controls.

The traditional IV approach for addressing selection into treatment based on unobservables is to identify a variable (instrument) that is highly correlated with treatment status but is otherwise as good as randomly assigned and only affects outcomes of interest through its effect on treatment status (commonly referred to as the exclusion restriction). For example, suppose that a policy lowered the cost of participating in a CTE program for some individuals, not others, but had no other impact on the individuals. If eligibility for the program or exposure to the instrument (z_i) appeared random or was based entirely on variables observed by the researcher, then one might use this instrument to identify the effect of treatment by applying a standard IV Wald estimator, as follows:

$$\frac{(E[y_i|z_i = 1] - E[y_i|z_i = 0])}{(E[T_i|z_i = 1] - E[T_i|z_i = 0])} \quad (3)$$

The numerator is simply the difference between the average outcomes of those exposed to the instrument and those who are not, which captures whether the program has an effect because exposure to treatment differs between the two groups. The denominator measures the effect of exposure to the instrument on actual treatment, scaling the effect of differential treatment incidence up to the expected effect arising from receiving treatment compared with not receiving treatment. The standard exclusion restriction assumption, that the instrument effects the outcome of interest only through its effect on treatment status, has a natural application to the consideration of the counterfactual. Perhaps, the experience that treated individuals would have had if they had not been treated

differs between those exposed to the instrument or program and those who are not. This situation would imply that counterfactual experiences for the two groups ($z_i = 0$ and $z_i = 1$) are different, creating differences between the outcomes of the two groups that are unrelated to treatment. As a result, this situation would violate the exclusion restriction assumption.

RD designs are a relatively recent approach for identifying causal treatment effects, where treatment is either determined or at least influenced heavily by crossing some type of threshold c. In this case, the individuals just above the threshold are nearly identical to the individuals just below the threshold, but those above receive treatment and those below do not; alternatively, at a minimum, those above the threshold have a substantially higher incidence of treatment. If the threshold or cutoff is truly arbitrary and cannot be manipulated by the participants or program administrators, then the RD approach overcomes the threat of selection bias because all individuals near the threshold are expected to be similar along both observable and unobservable dimensions, implying that treatment is as good as randomly assigned, conditional on being close to the threshold. However, although RD designs have strong internal validity, estimated treatment effects are measured explicitly only for those at the threshold:

$$E[Y_{1i} - Y_{0i} | x_i = c] \quad (4)$$

where *x* is the running or forcing variable that determines whether the individual is above or below the cutoff *c*. In practice, there is usually insufficient information to estimate treatment effects exactly at the threshold; hence, observations that fall within a bandwidth b of the threshold will be used. The differences between the individuals above and below the threshold are captured by conditioning on the actual value of the running variable under the assumption that the unobservables vary continuously, rather than jumping discontinuously at the threshold like treatment does. The estimated treatment effects can then be expressed as follows:

$$E[Y_{1i} - Y_{0i} | x_i, c + b \geq x_i \geq c - b] \quad (5)$$

In this case, the counterfactual question reduces to comparing the experiences of those above the threshold who are more likely to receive the treatment to the experiences of those below the threshold. If the experiences of nontreated individuals vary continuously across the threshold, then these differences will be captured by the running variable, but this would need to be empirically verified. A second concern is that RD often is estimated with a small, relevant sample because internal validity usually requires a narrow bandwidth. In this case, the sample within the bandwidth might provide only very noisy information on the counterfactual experiences of those who were not treated, and one might be forced to rely on the counterfactual experiences of a larger sample of individuals, which may imply bias in the documented counterfactual.

Finally, we turn to situations where treatment is effectively random, perhaps because of application to an oversubscribed program, where admission is determined via lottery, or as part of an RCT, where a sample is randomly drawn from a population of interest and then randomly allocated between the treatment and control groups. In either case, the control sample is drawn from the same population as the treatment sample, so the outcomes of the control sample will provide an unbiased estimate of the outcomes that the treatment sample would have experienced if they had been selected for the control group instead because both samples are drawn from the same population. Therefore, the estimated treatment effect is simply

$$E[Y_{1i} - Y_{0i}] = E[Y_{1i}] - E[Y_{0i}] \quad (6)$$

This situation also is ideal from the perspective of understanding the counterfactual. Again, because both samples are drawn from the same population, the experiences of the control group provide an unbiased estimate of the experiences that the treatment group would have had if they had not been treated. The larger issue in the context of a randomized study is the availability of information on the experiences of the control group. However, in the

context of an RCT study, this data problem is very tractable because both the control and treatment group populations are determined prior to the administration of any treatment, so following both groups across time should typically be feasible.

Section III. Case Studies

IIIa. Regression and Matching Methods

For our observational study, we discuss an evaluation by SRI International of the California LLDI (Warner et al., 2016). In 2009, The James Irvine Foundation launched the initiative to demonstrate this approach in nine districts. The nine districts participating in the initiative varied in size from slightly more than 5,000 high school students to more than 185,000 high school students and represented a variety of geographic regions across California. The districts had a high proportion of disadvantaged and non-White students, and in all nine districts, student achievement was below average.

The initiative was guided by a principle of rejecting traditional separation of students into vocational and academic tracks. Therefore, in addition to standard approaches, such as CTE courses in sequences emphasizing real-world applications and work-based learning, LLDI also emphasized rigorous academics to prepare students to succeed in college and provided comprehensive support services to help students who were traditionally marginalized succeed in their academic coursework. A key aspect of the program, however, was the creation of certified linked learning pathways that had successfully undergone an external review process of pathway quality managed by ConnectEd: The California Center for College and Careers or by NAF (previously the National Academy Foundation), a national network of college and career academies. As of July 2016, 46 pathways were certified in the nine districts.

Focusing on the evaluation analysis of administrative data, SRI examined dropout and high school graduation rates, completion of college preparatory credits, and college attendance and persistence. The primary treatment group was any student enrolled in a certified pathway. The decision to enroll in a certified pathway was entirely up to the individual student but could be influenced by school personnel, including faculty involved in developing or teaching within the pathway. Critically, students in participating districts had to have access to all pathways. In most cases, this access was provided by allowing open choice across high schools in the participating districts.

Given this district-level model for the provision of pathways, the control group, or the counterfactual comparison, was initially based on a sample of all students in the district who did not participate in certified or noncertified CTE pathways. However, this assumed that students who participated in certified pathways would be evenly distributed across other alternatives if the pathway did not exist. In many of these districts, students had several alternatives, including alternative and continuation high schools where average levels of student achievement were lower. An alternative assumption might be that pathway students would have attended a traditional high school if they had not been part of the path. In fact, many of these students completed pathways within their traditional high school. This counterfactual difference matters for the estimated effects of belonging to a certified pathway. Warner et al. (2016) found that pathway participants had a 13 percentage point advantage in terms of high school graduation compared with all other students in the district, but that advantage was reduced to 11 percentage points when including only those students who attended traditional high schools in the control group. Further, many students attend their assigned school by default, and if they participate in a certified pathway, that participation is likely at their assigned school. Given the large variation across individual schools within midsized and large districts, perhaps students in the same assigned school would have made a better control group for analyzing the effects of the program.

Also, the selection of schools into the initiative may have substantial implications for the counterfactual experiences of students in the linked learning certified pathways. A key criterion for school selection was "a demonstrated track record [of] implementing career pathways and evidence of existing capacity on which to develop a larger system of multiple pathways" (Warner et al. 2016, p. 2). Most pathways included in the study were, in fact, preexisting career academies that chose to become part of the initiative. If students systematically selected into CTE programs regardless of the initiative's activities, many students in the certified linked learning pathways would have been pursuing a very similar CTE focus area—even without the initiative—because the prior established pathways were the pathways that would likely have been certified the earliest. If so, it is impossible to know whether the linked learning process (i.e., the initiative) leads to more productive CTE education, although these analyses might be informative on the effect of exposure to a well-developed, possibly preexisting CTE curriculum. Although the study did find smaller positive effects on high school graduation for students in noncertified pathways, this result could arise simply because the certified CTE pathways were the most effective pathways at the school.

In this context, it would have been useful for the report to examine whether the effects of certified pathways depended on the order and timing of certification. The pathways that were certified later were likely newer or less-developed pathways at the beginning of the initiative. If the initiative had never occurred, those pathways might never have been developed. If the newest paths have similar effects as earlier pathways developed at the same school, such evidence would be supportive of a broader impact of the initiative and its associated certification process.

The possible sorting of students into CTE pathways also creates a potential threat to identifying the causal effect of the program, regardless of how the counterfactual treatment is defined. In observational studies similar to this analysis, the researcher can only condition on the observable student attributes, and if students select into the program based on unobservable attributes, then treatment effect estimates may contain the influence of those unobservable attributes on outcomes. As shown by Warner et al. (2016), the pathway student sample appeared to be more female, more Latino, less Asian, higher achieving, and less likely to be receiving special education services than the sample of district students attending traditional high schools. To address this sorting, the report authors conducted a regression analysis using administrative data, where outcomes were related to both participation in a certified pathway or treatment indicator (T_i) and the predetermined variables ($\boldsymbol{X_i}$) that the administrative data contained on the individual students. The following equation illustrates their model:

$$y_{idc} = \boldsymbol{\alpha}^T \boldsymbol{X_i} + \beta T_i + \boldsymbol{\tau_d} + \boldsymbol{\gamma_c} + \varepsilon_{idc} \quad (7)$$

where y_{idc} denotes an outcome of interest for student *i* in district *d* and cohort *c*, $\boldsymbol{\tau_d}$ is a vector of district fixed effects, $\boldsymbol{\gamma_c}$ is a vector of cohort fixed effects, and ε_{idc} is a stochastic error term representing factors unobserved in the data. The district fixed effects ($\boldsymbol{\tau_d}$) assured that the control group students for pathway participants were students from the same district. The cohort fixed effects ($\boldsymbol{\gamma_c}$) allowed outcomes to vary across cohorts of incoming ninth graders, breaking any correlation between trends in student outcomes and the growing availability of certified pathways. The list of control variables included gender, whether a student participated in the free or reduced-price lunch program, race and ethnicity, high or low achievement on the English Language Arts California Standards Test, whether the student was classified as gifted or as special needs, and a series of controls on the student's English proficiency. If one wished to use students in the same school as the counterfactual, then Equation 7 could be easily modified by replacing district fixed effects with school fixed effects.[2]

[2] The authors allowed ε_{idc} to be correlated between students in the same pathways; however, allowing for such correlation has no systematic influence on the estimated effect of participating in a certified pathway. Specifically, the authors estimated a hierarchical linear model with random effects by pathway and student. The estimates arising from ordinary least squares and hierarchical linear models theoretically converge to the same value.

When evaluators assessed the effects of pathway participation conditional on these control variables, the estimated effect of the program fell by more than half from 11% to 5% on high school graduation. Students appeared to sort into the certified pathways based on observable attributes that correlated with high school graduation, so there is every reason to believe that students also may sort based on important unobservables such as determination, initiative, and family support. Therefore, substantial erosion in the treatment effect by adding student controls suggested that adding additional, currently unobserved information on students might erode most of the estimated program effects. Unfortunately, the research report did not provide these types of comparisons for any outcomes other than high school graduation, so one cannot tell how sensitive the effects on college attendance and perceptions of college readiness were to controls for student observables.

Further, the researchers did not provide information on how well the control variables explained outcomes. This information is important to understand. If the observables explained much of the variation in student outcomes while leaving most of the outcome differences in place, then the risk of bias from unobservables was likely limited. Specifically, because the relationship between observed student attributes and outcomes related to program participation was weak, it is reasonable to assume that the relationship for unobservables was likely weak as well. This idea typically relied on the assumption that observable controls are at least as important in controlling for bias as the unobserved factors that influence both participation and outcomes. This assumption seems reasonable given that typically the information collected was likely collected because it was considered to be important (see Oster, 2019, and Altonji, Elder, & Tabor, 2005).

The administrative data was analyzed using linear controls for the observables, whereas the survey data on college readiness was examined using a propensity score approach. For the analysis of survey responses, the authors included additional student controls for actual standardized English language arts and mathematics test scores and the educational attainment of parents. However, the use of a propensity score approach did not change the key concerns about bias from unobserved student attributes. To illustrate this point, consider how a traditional propensity score analysis might be conducted by first modelling treatment and then measuring outcomes by comparing students who had a similar likelihood of being treated. The following equation represents the discrete outcome of whether a student receives treatment:

$$T_i = \begin{cases} 1 & \text{if } \boldsymbol{\rho}^T \boldsymbol{X_i} + \mu_i \geq 0 \\ 0 & \text{if } \boldsymbol{\rho}^T \boldsymbol{X_i} + \mu_i < 0 \end{cases} \quad (8)$$

Using Equation 8, we can obtain an estimate of the probability of treatment, $Prob[y_i|\widehat{\boldsymbol{\rho}}^T \boldsymbol{X_i}]$ which is a function of observables and the estimated parameters from the equation. Then, if a propensity score matching approach is used, treatment and control group students can be divided into *B* bins based on having similar likelihoods of participating in a certified pathway. For example, the effect of treatment in a given school and cohort can be estimated by comparing only those students in the same bin:

$$\hat{\beta} = \sum_{b=1}^{B} (Mean[y_i|T_i = 1, b] - Mean[y_i|T_i = 0, b]) * \theta_b \quad (9)$$

where θ_b is the fraction of students in a bin.[3] However, a very similar estimate can be obtained by simply including bin fixed effects ($\boldsymbol{\pi_b}$) into a simplified version of Equation 7.

$$y_{ib} = \beta T_i + \boldsymbol{\pi_b} + \varepsilon_{ib} \quad (10)$$

Because bin membership is simply a nonlinear function of $\boldsymbol{X_i}$, propensity score analysis simply allows for nonlinearities in selection. Thus, the value of propensity score analysis is that it delivers robust estimates when

[3] We could make similar estimation and calculations by bin, school, and cohort.

student selection based on observable attributes is very complex and highly nonlinear, but it does nothing to address bias from unobserved student attributes.[4] As with the administrative data, the authors did not show how inclusion of the administrative controls or the additional survey controls affected the magnitude of the treatment estimates, but given the sensitivity of the high school graduation results, one should be careful interpreting their findings.

In observational studies, researchers can condition effect estimates on only the observed information. A simple linear regression approach may not completely control for selection into the program based on the observables used in regression. However, in practice, many studies find that treatment effect estimates are very similar when using either a linear regression or a propensity score approach, suggesting that selection into many programs is well approximated by linear models in student attributes (see Zanutto, 2006, for example). Therefore, the primary concern about systematic selection into programs typically arises from unobserved participant attributes. Consequently, regardless of the statistical technique used, program effect estimates should be assessed in part on how sensitive the treatment effect is in controlling for student attributes compared with how well those attributes explain student outcomes.

In summary, our analysis of the LLDI study identified several issues that should be considered when interpreting the findings of any study that uses a standard regression or propensity score approach to estimate treatment effects. First, because students have access to all pathways within a district, the LLDI study used the sample of all other district students as a control group. However, if pathway participants would typically attend their assigned high school and would likely have attended that high school without the program, then students in the same school might form a better counterfactual, or control group. In fact, Warner et al. (2016) found that pathway participants had a 13 percentage point higher high school graduation rate, but the effect was reduced to 11 percentage points using only students attending traditional high schools. Second, a key criterion for school selection was “a demonstrated track record [of] implementing career pathways and evidence of existing capacity on which to develop a larger system of multiple pathways” (Warner et al. 2016, p. 2). Many of the well-developed and well-run CTE focus areas in selected high schools prior to the initiative likely became the pathways that were eventually certified, and these programs would have been available to students in a counterfactual world with no LLDI. Therefore, any evaluation of the LLDI would have struggled to separate out the effects of the initiative from the effect of CTE more generally. Finally, observational studies can condition only on the observables available, whether analysts use regression or propensity score approaches. The effects of participation on high school graduation conditional on these controls was less than half of the unconditional estimates (5% relative to 11%), suggesting that controls for unobservables might explain the entire remaining effect.

IIIb. Instrumental Variables

As noted previously, Cullen et al. (2005) used an IV identification strategy to examine the effect of attending a career academy in CPS on high school graduation rates. In 1980, CPS created a school choice system as part of a consent decree requiring the district to address racial segregation. In this system, students were guaranteed a spot in an assigned neighborhood school, but they could apply to any school within the district and could submit an unlimited number of applications. By the mid-1990s, approximately half of all high school students opted out of their assigned high school to attend a different school within the district.

Cullen et al. (2005) examined a sample of more than 60,000 students who began ninth grade within the district in 1993, 1994, and 1995. After conditioning on student observables, they found that high school students who opted out of their assigned neighborhood school were 7.6 percentage points more likely to graduate from high school

[4] See Morgan and Winship (2015) for a comparison of linear regression and matching methods to identify causal effects.

than their peers who did not opt out of the same school. However, they also showed that students who attended their assigned area school had substantially lower eighth-grade test scores than students who attended choice schools. This strong sorting on ability raises concerns that students also might select out of their assigned schools based on unobserved attributes, such as motivation or family support, which might lead to higher graduation rates even if the choice program did not improve student performance. As discussed earlier, selection into treatment on unobservables represents the most significant concern associated with program evaluations based on observational data.

To address this concern, Cullen et al. (2005) exploited the distance between where a student lived and alternative, nonassigned schools, given the strong role physical access plays when students and their parents choose schools. Specifically, they measured the distance in miles from a student's home to the closest nonassigned school for three categories: career academy, high-achieving high school, and any regular high school. These distances can be used as instruments to predict whether students opt out of their assigned school overall or to a specific type of school, and this prediction can be used to estimate the effect of opting out on student outcomes, such as graduating from high school. Under an exclusion restriction assumption that these distances (instruments) do not have any direct influence on or correlation with the relevant outcomes beyond the effect of distance on whether the student opts out (treatment), the correlation between the instrument and outcomes can uncover the effect of treatment on outcomes. Cullen et al. found that being closer to a career academy both increased the likelihood of attending the career academy and raised high school graduation rates. For example, their estimates implied that opting out of an assigned neighborhood school and into a career academy raised graduation rates between 4 and 5 percentage points for students in the middle of the ability distribution. On the other hand, being closer to traditional high schools, even high-performing schools, had a much weaker and statistically insignificant relationship with high school graduation rates.

In a multivariate context, we can express this identification strategy for the effect of treatment (T_i) on outcomes (y_i) using the following equations:

$$y_i = \boldsymbol{\alpha}^T \boldsymbol{X_i} + \beta T_i + \varepsilon_i \tag{11}$$

$$T_i = \begin{cases} 1 & \text{if } \boldsymbol{\rho}^T \boldsymbol{X_i} + \delta Z_i + \mu_i \geq 0 \\ 0 & \text{if } \boldsymbol{\rho}^T \boldsymbol{X_i} + \delta Z_i + \mu_i < 0 \end{cases} \tag{12}$$

where Z_i is the instrument and μ_i is distributed based on the underlying selection into the treatment model (e.g., logit, probit, or linear probability model). This two-stage approach involved estimating Equation 12 first and using the estimates to create a predicted value or probability of treatment; then the second stage, Equation 11, could be estimated via ordinary least squares, replacing T_i with its predicted probability. Similar to Warner et al. (2016), Equation 11 can be extended to include assigned school fixed effects so that students are compared with not only other students who were assigned to the same school but also those who reside at different distances from their closest alternative schools.

The primary difference between Equations 11 and 12 and the propensity score approach discussed in the previous section is that Equation 12 contains an exclusion restriction in that Z_i influences the predicted probability of T_i and does not have a direct influence on the outcome y_i. Essentially, students exposed to a low value of the instrument serve as control or counterfactual students for treatment group students exposed to a high value of the instrument. As the standard IV Wald estimator shows, differences in the instrument imply a change in treatment exposure across different values of the instrument based on δ, and the effect of that change on outcomes is scaled by the implied change in treatment exposure (δ):

$$\hat{\beta} = \frac{(E[y_i|z_i = 1] - E[y_i|z_i = 0])}{(E[T_i|z_i = 1] - E[T_i|z_i = 0])} \tag{13}$$

To illustrate this for a multivariate context, Equation 14 presents the reduced form estimate of the relationship between the instrument and outcomes:

$$y_i = \tilde{\boldsymbol{\alpha}}^T \boldsymbol{X_i} + \theta Z_i + \varepsilon_i \quad (14)$$

If treatment or Equation 12 follows a simple linear probability model

$$Prob[T_i] = \boldsymbol{\rho}^T \boldsymbol{X_i} + \delta Z_i + \mu_i \quad (15)$$

then the causal estimate from Equation 11 is

$$\hat{\beta} = \frac{\hat{\theta}}{\hat{\delta}} \quad (16)$$

A key question for interpreting any IV analysis is which individual in the sample identifies θ. IV approaches capture what is known as the local average treatment effect (LATE). Although "average" is self-explanatory, "local" has a very specific meaning: The estimate captures the effects of treatment on marginal individuals who are near the threshold of taking-up treatment (i.e., individuals who were pushed into treatment by exposure to the instrument). For example, imagine that Z_i is a binary (0/1) variable, where 1 indicates a higher incidence of treatment. In the Cullen et al. (2005) context, students are either close to 1 or far from 0 in choosing a career academy. Then, based on Equation 12, exposure to $Z_i = 1$ would have the following effects:

$\boldsymbol{\beta}^T \boldsymbol{X_i} - \delta + \mu_i < 0$	No effect, remain untreated (never taker)	(17)
$0 \leq \boldsymbol{\beta}^T \boldsymbol{X_i} - \delta + \mu_i < \delta$	Change from untreated to treated (complier)	(18)
$\boldsymbol{\beta}^T \boldsymbol{X_i} + \mu_i \geq 0$	No effect, remain treated (always taker)	(19)

Exposure to $Z_i = 1$ increased treatment for only the sample of compliers, individuals who fell into the narrow range between the threshold 0 and the magnitude of the effect of exposure. Never takers have too low a propensity to take up the treatment, and exposure to the instrument cannot reverse that decision. In contrast, always takers have such a strong propensity that they participate whether exposed to the instrument or not. Finally, the IV estimate is based on the magnitude of the effect on compliers and then scaled up to a population effect by dividing by the share of individuals moved from being untreated to treated.

In the case of distance to a career academy as an instrument, many students did not attend the academy even when they were right next door with an effective distance of zero. Cullen et al. (2005) almost certainly observed distances where the likelihood of attending a career academy was quite small. Under those circumstances, the LATE captured the treatment effects for individuals who had relatively strong preferences for attending a career academy. Specifically, individuals with a strong preference for CTE were those for which modest changes in distance could push them into attending such an academy (compliers), whereas individuals with weaker preferences would likely never attend a career academy even if they lived next door (never takers). If the treatment effects were weaker for individuals with weaker preferences, then the LATE may substantially overstate the treatment effect within the entire population. Therefore, it is important (if feasible) to test whether the magnitude of treatment effects varies with the value of the instrument. If treatment effects are heterogeneous across the instrument, researchers should document the distribution of treatment take-up over the domain of the instrument to assist in the interpretation of their estimates.

Next, even for the compliers, interpretation of the treatment effect depends on the counterfactual experiences of the marginal individuals who did not opt out of their assigned school for a career academy. Therefore, researchers need to document the many factors that may change as students select into career academies and away from their assigned school. For example, selecting into a career academy may change other factors, such as the quality of peers or the school disciplinary system. If so, it is hard to disentangle the effects of those

changes from the effects of the career-specific training provided at the career academy. Cullen et al. (2005) addressed these concerns in part by examining the impact of predicted peer quality based on the closest choice schools. They found that predicted peer quality did not affect high school graduation, and their estimates of the effect of attending a career academy on high school graduation were robust to the inclusion of this control.

A final concern in terms of the counterfactual experience of students in a career academy is the extent to which those students can take similar career training electives even if they do not attend a career academy. The focus of Cullen et al. (2005) was on school choice, as opposed to career academies, so they did not provide any discussion or comparison between CTE options in the career academies and the options available in other schools, including the student's home school. However, in other studies of stand-alone high schools with a CTE focus, the availability of CTE options was much higher, and stand-alone CTE schools had a much broader integration of CTE into the educational environment and experience than traditional high schools (see, for example, Brunner et al., 2019). This aspect of stand-alone academies suggests that they might provide quite different experiences from the residentially assigned school.

As in observational studies, IV studies must be evaluated in terms of how good a counterfactual is provided by students who have low values on the instrument (e.g., Are students who are close to career academies different on unobservables than students who are far from career academies?). Such omitted variables violate the exclusion restriction because the instrument (distance) is then predictive of outcomes (high school graduation) for reasons other than the instrument's effect on treatment (attending a career academy). For example, distance from a career academy may be correlated with the quality of the assigned school if career academies tend to be located closer to schools with higher test scores or more advantaged populations.

This specific problem might be solved by including assigned school fixed effects so that distance comparisons are made only for students assigned to the same schools. However, even in this case, the career academy may be located in (i.e., closer to) higher socioeconomic status neighborhoods. Thus, being closer to a career academy also exposes a student to those neighborhoods and affects their outcomes through this additional mechanism as well as through attending a career academy. This concern about omitted variables often leads researchers to examine instruments for balance over observables under the argument that if the instrument is uncorrelated with observed neighborhood or student attributes, then it is likely also uncorrelated or at worst weakly correlated with the unobserved attributes. For example, Cullen et al. (2005) showed that their results were robust to the inclusion of assigned school fixed effects, with no correlation between distance to schools and the eighth-grade mathematics and reading scores of individual students.[5] These tests of balance were closely related to examinations of the how treatment effect estimates change when observables are included in the model, as discussed earlier for the LLDI, because the inclusion of observables will affect treatment effect estimates only if those observables correlate with treatment.

Finally, for this specific instrument (distance), the exclusion restriction might also fail because the attribute matters directly, which is a problem that cannot be addressed by balancing tests. For example, Cullen et al. (2005) found that proximity to one's assigned or home school, not just career academies, influenced high school graduation. Perhaps proximity to the school attended always raises graduation rates. If that were the case, the effect of attending the career academy might be driven by the fact that students who attend the academy reside near the academy, not the effect of attending the career academy itself. Another potential explanation may be that a higher likelihood of students attending the career academy (further from the home school and closer to the academy) also implies a lower likelihood of attending the home school, which leads to changes at the home school. For

[5] However, for career academies, the balancing test relating test scores and distance did fail at the 10% level of confidence, which is somewhat concerning.

example, if the home school has a nearby career academy, the home school may lose a lot of students to the career academy. This may disrupt negative social networks that might have interfered with completing coursework and graduating for students in both the home school and the career academy.[6] As a result, all these students do better, and based on location, most of these students reside near a career academy.

In summary, the IV approach of Cullen et al. (2005) identified the effect of CTE for students whose attendance at a career academy was influenced by physical proximity. Many students did not attend the academy even when they were right next door with an effective distance of zero. Under those circumstances, IV estimation captured the treatment effects for individuals who had relatively strong preferences for attending a career academy. Further, selecting into a career academy may change other factors, such as a given student's peer group or the school disciplinary system. If so, it is hard to disentangle the effects of those changes from the effects of the career-specific training provided at the career academy. Further, IV studies must be evaluated in terms of how good a counterfactual is provided by students who have low values on the instrument (e.g., Are students who are close to career academies different on unobservables than students who are far from career academies?). Such omitted variables can violate the exclusion restriction because the instrument (distance) is then predictive of outcomes (high school graduation). Finally, for this specific instrument (distance), the exclusion restriction might also fail if proximity to a school directly influences student success.

IIIc. Regression Discontinuity Designs

Brunner et al. (2019) used an RD design to examine the effect of attending one of 16 stand-alone technical high schools within the Connecticut Technical High School System (CTHSS) on student educational and labor market outcomes. CTHSS is a quasi-independent district of choice, where all students who attend one of the 16 schools in the system participate in some form of CTE. Schools in CTHSS are spread across the state and serve approximately 11,000 students, or 7% of all high school students in Connecticut, each year. Although schools in CTHSS are located throughout the state, approximately 30% of all CTHSS students come from one of Connecticut's six poorest central cities, and, on average, students who apply to CTHSS tend to be more disadvantaged and exhibit lower academic performance than the overall population of Connecticut high school students.

Students apply for admission to up to three CTHSS schools of their choice in rank order during the second half of their eighth-grade year. Students are assigned an application score that is based on their eighth-grade mathematics and reading test scores plus grade point average and attendance in middle school.[7] Schools then admit students in descending order of application scores until all seats in a school are full. Given that all schools in CTHSS are oversubscribed each year, the admission process leads to a threshold score for which students with scores above the threshold receive an offer of admission and students with scores just below the threshold do not receive an offer of admission. Students are admitted to any school where their admission score exceeds the threshold for that school. Given the intent to treat (ITT) nature of the estimate, the sample includes one observation for each student application, and for that observation, the student is treated if admitted to that school.[8]

---

[6] See Billings, Deming, and Ross (2019) for evidence of the effect of such disruptions.
[7] In more recent years, CTHSS has added points for extracurricular activities and a written statement of interest to the application score.
[8] There is some deviation between a student's application score and whether the student receives an offer of admission. Specifically, some students with scores lower than the threshold score may be admitted to promote a diverse student body. Other students with scores above the threshold may not be admitted based on information in their disciplinary file or because they applied late.

Given both the complex process of student admission and the potential systematic differences between students who accept or turn down admission offers, a simple comparison of applicants who attend schools in CTHSS to applicants who do not would be insufficient to provide causal evidence on the effects of CTHSS on student outcomes. Therefore, Brunner et al. (2019) employed an RD identification strategy to estimate these effects. RD designs identify the causal effect of treatment on outcomes by exploiting the fact that individuals who are close to the admission threshold—but on opposite sides of the threshold—are likely similar based on both observable and unobservable factors. In fact, given that the researcher typically observes all the factors used for admission, individual student unobservables should be completely uncorrelated with whether a student's position is above or below the threshold. As a result, by comparing individuals who are close to but on opposite sides of the admission threshold, RD designs avoid the potential bias associated with nonrandom selection into treatment by comparing students who exhibit small differences on observable factors and, on average, should exhibit no differences on unobservables.

The RD design employed by Brunner et al. (2019) was a fuzzy RD design, where the fuzziness arose because (a) not all students with scores (S_i) above the threshold chose to attend a school in the CTHSS (even though they received an offer of admission), and (b) some students with a score below the admission threshold were admitted to promote a diverse student body. The fuzzy RD design can be expressed essentially as an IV analysis for outcome y_i, as discussed in the previous subsection, where a dummy variable for whether the score is above the threshold (D_i) serves as an instrument for treatment (T_i). Specifically, the fuzzy RD design can be expressed as follows:

$$y_i = \alpha_0 + \beta T_i + \omega_1 S_i + \omega_2 (D_i \cdot S_i) + \varepsilon_i \quad (20)$$

$$T_i = \rho_0 + \delta D_i + \upsilon_1 S_i + \upsilon_2 (D_i \cdot S_i) + \mu_i \quad (21)$$

where treatment (T_i) is modelled as a linear probability model. The key differences between the fuzzy RD design given by Equations 20 and 21 and a standard IV design given by Equations 18 and 19 are as follows: (a) The fuzzy RD design typically does not control for a standard vector of observables, which are replaced by simple intercepts, and (b) the design adds additional controls to directly account for the minor observed differences in scores between the samples above and below D_i. Specifically, because the regression considers only those observations near the threshold, the potentially nonlinear relationship between score S_i and outcome y_i can be approximated by a linear or higher order polynomial relationship. Further, this relationship is allowed to vary on either side of the admission cutoff by interacting the score and the cutoff, possibly because admitted student unobservables affect take-up or compliance in ways that vary based on the score.[9]

Equations 20 and 21 are estimated using two-stage least squares, whereby Equation 21, the first stage, is initially estimated to predict treatment (T_i) based on whether an individual's application score is above the threshold or cutoff score. The predicted value is then used in replacement of T_i in Equation 20 in the second stage to obtain treatment on the treated (TOT) estimates, which are given by the estimated coefficient $\hat{\beta}$ from Equation 20. Alternatively, one can obtain ITT estimates by estimating the reduced form equation:

$$y_i = \tilde{\alpha}_0 + \rho D_i + \tilde{\omega}_1 S_i + \tilde{\omega}_2 (D_i \cdot S_i) + \varepsilon_i \quad (22)$$

where ρ provides an estimate of the effect of being offered the treatment (admission to a school in the CTHSS) on the outcome of interest.

[9] The RD specifications used by Brunner et al. (2019) are a bit more extensive in the sense that they pool many admissions cohorts and students applying to many schools in CTHSS. As a result, the score variable is centered by subtracting the threshold for each cohort and school. In addition, their specification included CTHSS school-by-cohort fixed effects to capture differences between schools and town of residence fixed effects.

As noted previously, under fairly weak assumptions, RD models provide an estimate of the treatment effect that can be expressed as

$$E[Y_{1i} - Y_{0i}|S_i = c] \quad (23)$$

where Y_{1i} and Y_{0i} are potential outcomes for individual *i* with and without treatment, respectively, and c is the cutoff or threshold score. Because one can never observe both Y_{1i} and Y_{0i} simultaneously, experimental and quasi-experimental designs exploit a control or counterfactual group to obtain Y_{0i}. Therefore, in practice, the theoretical RD estimator in Equation 23 must be modified as follows:

$$E[Y_{1i}|S_i, c + b \geq S_i \geq c] - E[Y_{0i}|S_i, c \geq S_i \geq c - b] \quad (24)$$

where b is the bandwidth or window around the cutoff score. Brunner et al. (2019) imposed a bandwidth of plus or minus 10 points on the admission score. Thus, their control or counterfactual group was given by individuals with application scores within the 10-point bandwidth and below the threshold or cutoff score.[10]

One obvious concern with causal estimates obtained from RD designs relates to their external validity or generalizability because the LATE is estimated only for those local students who are within the 10-point bandwidth. It is therefore important to carefully characterize the marginal individual for whom the treatment effects are being identified. This will typically involve comparing the observable characteristics of individuals who are close to the cutoff score to the broader population. Brunner et al. (2019) addressed this issue by comparing students within 10 points on either side of the admission cutoff to the broader population of students who applied for admission to a school in the CTHSS and the population of eighth-grade students in Connecticut as a whole. Further, in cases like Brunner et al., where the sample pools many subsamples (in this case, schools and application cohorts), where each subsample faced a different threshold, it is valuable to test whether the effect of treatment or admission varies across the admission threshold. If not, this is at least suggestive that the estimated effects could generalize to a broader population of applicants.

Brunner et al. (2019) showed that relative to the overall population of eighth-grade students in the state, students who applied for admission to a school in the CTHSS were more likely to be male (58% versus 51%) and almost twice as likely to be African American, Hispanic, or eligible for free or reduced-price lunch. CTHSS applicants also tended to have standardized seventh-grade mathematics and reading scores that were approximately two thirds of a standard deviation below the state average. On the other hand, students within 10 points of the admission threshold tended to be quite similar in terms of gender, race and ethnicity, and free or reduced-price lunch status relative to all CTHSS applicants, although they tended to have standardized mathematics and reading scores that were somewhat lower than the broader population of applicants. Thus, the LATE identified by Brunner et al. applied to a population of students that was lower achieving, of lower income, and more likely a student of color than the population of students in Connecticut overall. However, Brunner et al. also found no evidence of heterogeneity in the effect of attending a school in the CTHSS over the admission thresholds used at different schools and for different cohorts. Further, they found no heterogeneity in effects based on the demographic attributes of the students or the sending schools. As noted earlier, these heterogeneity analyses suggest that the effects of Brunner et al. are very broad based and could generalize to a significantly larger population.

[10] In the reduced form model given by Equation 22, Y_{1i} in Equation 24 represents the outcome for individuals within the specified bandwidth and above the admission threshold, regardless of whether they actually attend a school in the CTHSS. Similarly, Y_{0i} represents the outcome for individuals within the specified bandwidth and below the admission threshold, even if they do attend a school in the CTHSS. In the IV model given by Equations 20 and 21, Y_{1i} represents the outcome for individuals who were induced to attend a school in the CTHSS because they were above the admission threshold and within the specified bandwidth, and Y_{0i} represents the outcome for individuals within the specified bandwidth who did not attend a school in the CTHSS because they were below the admission threshold.

To interpret causal estimates obtained from RD designs, it is critical to understand and make clear the treatment control contrast. In the case of Brunner et al. (2019), this involved clearly describing the experiences of students who attended a school in the CTHSS (those with outcome Y_{1i}) and the counterfactual experiences of control group students who did not get into a school in the CTHSS (those with outcome Y_{0i}). In terms of the treatment experiences, rather than taking a set of electives, as is common in traditional high school settings, students attending a school in the CTHSS choose from a set of CTE courses that are typically grouped into one of 10–17 programs of study, such as heating, ventilation, and air conditioning; information systems technology; and health technology. In the fall of ninth grade, students at schools in the CTHSS explore different programs of study and are then assigned to a particular program based on their preferences and program availability. CTHSS students then spend the next three years with the same cohort of peers and instructors within their program of study and take a minimum of three (often more) aligned courses within their program of study. In ninth grade, all CTHSS students also must take a course that focuses on professional and noncognitive skills. In that course, students must develop a team social contract that serves as a contract between students and teachers regarding classroom behavior and professional etiquette. The class also includes team-building exercises designed to foster a sense of community among students from different towns. Instructors of CTE courses work closely with the instructors of core academic subjects to ensure overlap of content. All schools in CTHSS also offer students work-based-learning opportunities through partnerships with local companies, as well as career awareness activities, such as job shadowing and company visits.

The vast majority of students who fail to gain admission to a school in the CTHSS attend a traditional high school in their town of residence.[11] Although most of these traditional high schools offer some CTE courses, they tend to offer significantly fewer CTE courses and programs of study than do schools in the CTHSS. Specifically, traditional high schools offer at most two to four CTE programs, with fewer multicourse sequences, and students often take only one or two CTE courses, which may or may not be in the same program area. At schools in the CTHSS, nearly 90% of all electives are CTE courses, and approximately 57% of all elective courses are trade courses in the areas of architecture, transportation, and manufacturing. In contrast, at traditional high schools, which represents the counterfactual experience of most students who do not gain admission, only 45% of all electives include some form of CTE, and only 5% of all electives are trade courses. Thus, on average, students who attend a school in the CTHSS are exposed to significantly more CTE courses and programs of study and have more opportunities to participate in work-based-learning and other career awareness activities relative to the control group.

Given that male and female students tend to pursue very different programs of study in schools in CTHSS and tend to have different patterns of labor force participation, Brunner et al. (2019) conducted separate analyses for male and female students. They found that male students who attended a school in the CTHSS were approximately 10 percentage points more likely to graduate from high school. They also found that although male students were initially about 8 percentage points less likely to attend college, by age 23, there was no difference in college attendance rates or semesters of college attended between CTHSS students and their control group counterparts. Finally, Brunner et al. found that at age 23, male students who attended a CTHSS school had quarterly earnings that were approximately 30% higher than the control group.[12] These results appear to be broad based with little evidence of heterogeneity across characteristics such as free or reduced-price lunch status, race and ethnicity, or

[11] A small percentage of students attend other types of choice schools, such as magnet and charter schools, that may or may not offer CTE experiences, and others may choose to attend a private school.
[12] The initial negative effects on college attendance in spite of increases in high school graduation raises the possibility that the outcome of interest may vary depending on the specific program. For example, the LLDI discussed previously focused in part on college preparation. On the other hand, CTHSS focuses heavily on career readiness and work-based learning, where successful preparation may reduce the need for immediate investments in higher education.

residence in a poor central city. In contrast, Brunner et al. found relatively precisely estimated null effects of attending a school in the CTHSS across all outcomes for female students.

Brunner et al. (2019) then asked whether the gap in the number of CTE courses and programs offered in schools in CTHSS compared with the number of courses and programs in counterfactual high schools could explain their results. They found that reducing the gap in CTE offerings by 10% could explain approximately 2% to 4% of their estimated treatment effects. Thus, although differences in the number of CTE course offerings could explain some of the positive treatment effects, other features of the CTHSS experience were clearly very important. What are these other experiences? Unfortunately, Brunner et al. were unable to provide a definitive answer to that question. As noted previously, some of those experiences include greater access to work-based learning, closer alignment between CTE experiences, coursework, and core academic subjects; other career awareness activities; and the effect of progressing through a 3-year CTE program with a common group of faculty and students. Importantly, because Brunner et al.'s evaluation identified only the effect of attending a school in the CTHSS, with all the experiences and components that entails, the authors could not determine how much of each experience or component of attending a school in the CTHSS contributed to the overall treatment effects. Furthermore, other factors, such as the peers whom students are exposed to in schools in CTHSS relative to counterfactual high schools, differences in student self-esteem, or reductions in the stigma effects associated with CTE that might be avoided by attending a school in the CTHSS, also may explain some of Brunner et al.'s treatment effects. Although some of these experiences are potentially measurable, such as peer effects, others are quite difficult to measure and thus remain unobserved. Consequently, it is difficult to fully attribute treatment effects to CTE exposure.[13]

In situations where there is more variation in the elements of the CTE program within the treatment sample, studies with strong identification strategies might be followed up by more descriptive analyses. For example, regression analysis might examine the correlation between treatment effects and various elements of the CTE program. Further, qualitative analysis might examine the implementation of the CTE program to uncover unique aspects of especially effective programs.

As with observational and IV studies, RD identification strategies also require assumptions. In this case, the RD strategy exploits a form of local randomization, where the randomization is conditional on the running variable S_i falling within a narrow bandwidth around the cutoff or threshold score c_i and on the precise value of the running variable S_i. A natural way to test for local randomization is to examine whether the observables are assigned randomly on either side of the threshold. In the case of Brunner et al. (2019), the observable predetermined variables include race and ethnicity, free or reduced-price lunch status, and prior achievement measures such as standardized mathematics and reading test scores. To test for balance, Brunner et al. and others replaced the outcome y_i in Equation 22 with the elements x_i^k of the vector of predetermined variables ($\boldsymbol{X_i}$) and tested the hypothesis that the estimated coefficient on D_i equaled zero.[14] Specifically, they estimated specifications of the following form:

[13] In the case of peer effects, Brunner et al. (2019) presented evidence that worked strongly against a peer effect story. The estimated positive effects of attending a school in the CTHSS were relatively constant when comparing students from very low-performing central city districts to students from suburban districts. Further, as noted earlier, the estimated effects also are uncorrelated with the admission threshold. Students who come from low-performing central city schools or who attend technical high schools with the highest admission thresholds would be expected to have the greatest improvement in peer quality, yet those factors appear unrelated to the benefits of CTHSS.

[14] One concern with this approach is if the vector of predetermined variables is large, it is quite possible that one or more of the estimated coefficients on D_i will be statistically significant simply because of chance or type I error. As a result, an alternative approach to testing for balance involves making D_i the dependent variable and regressing it on the vector of predetermined

$$x_i^k = \alpha_0^k + \rho^k D_i + \omega_1^k S_i + \omega_2^k (D_i \cdot S_i) + \varepsilon_i^k \quad (25)$$

Finally, a critical decision in conducting RD models is the bandwidth. On one hand, if the bandwidth is too narrow, the analysis may not have enough power to detect treatment effects. On the other hand, if the bandwidth is too wide, then the IV estimate of β in Equation 20 or the reduced form estimate of ρ in Equation 22 is likely to be biased because the linear approximation of the relationship between the outcome and the running variable will begin to break down as the distance between the threshold and the observations increases. As a result, it is important to demonstrate that RD results are robust in magnitude to reducing the size of the bandwidth considered, even if these reductions lead to estimates that are somewhat noisy.

In summary, RD designs identify the causal effect of treatment by exploiting the fact that individuals who are close to the admission threshold—but on opposite sides—are likely similar. One obvious concern with RD estimates is external validity because the effect is estimated only for those local students who are within the chosen bandwidth. However, Brunner et al. (2019) found that the effects are similar across schools and years, even though the thresholds used in each school and year vary considerably, suggesting broad-based effects. CTHSS represents a substantially different counterfactual experience because selected students leave their assigned high school and enter a separately run statewide district, where students explore a wide variety of CTE programs and then study the selected program in substantial depth, including opportunities for work-based learning. Most students who do not gain admission attend a traditional town high school. Although most town high schools offer some CTE, they tend to offer significantly fewer CTE courses and programs of study than do schools in CTHSS, and few offer programs of study in traditional building trades or manufacturing. However, Brunner et al. found that the gap in the number of CTE courses and programs offered in schools in CTHSS can explain only about 30% of the treatment effects. The rest of the effects were likely the result of the unique nature of stand-alone technical high schools and the learning experiences offered.

IIId. Randomized Controlled Trials and Lottery-Based Admissions

One of the best-known RCTs in CTE is the evaluation of career academies conducted by MDRC (Kemple, 1997, 2004, 2008). Career academies are a CTE model that began in the 1970s and saw rapid expansion in the early 1990s largely because of the inclusion of the approach under the School-to-Work Opportunities Act of 1994. Career academies are a widely used high school reform initiative that aim to keep students engaged in school and prepare them for successful transitions to postsecondary education and employment. Typically serving between 150 and 200 students from Grades 9 or 10 through Grade 12, career academies are small learning communities within a traditional high school where students take several classes per year together, with the same group of teachers. The academies also combine academic and technical curricula related to a career theme and establish partnerships with local employers to provide work-based learning opportunities.

Beginning in 1993 and including cohorts of students who enrolled in career academies between 1993 and 1995, MDRC conducted a rigorous 10-year evaluation of the model using a random assignment research design in a diverse group of nine high schools across the United States. Located in medium- and large-sized districts, the schools all served low-income populations in urban settings.[15] The participating career academies were able to implement and sustain the core features of the approach, and they served a cross-section of the student populations in their host schools. Of the nine participating academies, approximately 80% of the students were enrolled in 3-year academies (six semesters) that began in 10th grade, and the remaining students were enrolled

variables and a linear control for the running variable. In which case, testing for balance now involves simply conducting a joint hypothesis test (*F*-test) of the null hypothesis that the vector of estimated coefficients on $\boldsymbol{X_i}$ are jointly equal to zero.

[15] The study began with 10 academies, but one site disbanded after the 1995–96 school year and was unable to meet the data needs of the evaluation. This site was ultimately dropped from the study.

in 4-year (eight semester) academies that began in ninth grade. The participating academies offered a range of occupational themes, including business and finance, technology, health, public service, and travel and tourism.

MDRC conducted random assignment in the eighth- or ninth-grade year prior to academy enrollment. The treatment group members were those who won the opportunity to enroll in an academy, and those who lost the same opportunity were the control, or counterfactual, group. Across the three cohort years of the study, 1,764 students were in the analytic sample of students.

Most quasi-experimental design studies, such as those discussed earlier in this paper, seek to mimic the properties of a random assignment design study by plausibly eliminating mechanisms of selection bias in which students with characteristics related to the outcome of interest are systematically sorted or selected into different program conditions. By contrast, with random assignment, the treatment and control groups are comparable, on average, by construction because the distribution of both observable and unobservable characteristics across groups is a function of the randomization, not student selection. For example, in many studies, student motivation to obtain and participate in a given treatment can be a concern because it cannot usually be observed by the researcher and is likely related to successful outcomes of the treatment program—students who want to engage in a program are more likely to excel in it.

In the career academies study, students expressed interest in the program by filling out an application or intake form to participate in the study and agreed to be entered into the randomization process in which all students had equal probabilities of selection into either of the assignment conditions. This process allowed students to reveal their level of interest in participating before receiving a placement in a random assignment group, so the treatment and control groups were drawn from the same population of applicants.[16] This process allowed researchers to assume that levels of interest and motivation were evenly distributed between the treatment and control groups. Further, because these samples were drawn from the same population, it is reasonable that without the program, these students would have made the same decisions, on average. For this reason, the random assignment study design assumed that the control groups' experience represented what would have happened to the treatment group students, in the absence of the treatment experience. In this way, random assignment created a convincing counterfactual group for the treatment group, and the difference in outcomes represented a credible estimate of the effect of the treatment. One possible flaw in this logic would be if the solicitation and application process made the career academy a more salient option to the applicants, and, as a result, the control group made different choices than they would have without the presence of the program.

Although pure random assignment studies can use a fairly simple estimation model to measure the effect of treatment, the model used in this study needed to account for a more complicated random assignment process. Specifically, because there were nine academies, and random assignment was conducted separately for each cohort of students within each academy, there were 20 separate site-by-year cohorts that participated in random assignment. These separate rounds of random assignment are known as "random assignment blocks," and the following model accounts for this study design feature (Kemple, 2004, 2008):

$$Y_i = \sum_n \gamma_{0n} S_{ni} + \boldsymbol{\gamma_1}^T \boldsymbol{X_i} + \beta_0 T_i + \varepsilon_i \qquad (26)$$

[16] In random assignment studies involving minors, parents rather than students must provide active consent to participate. However, students also usually provide their own assent. It is no longer clear what the active consent and assent processes were in this study, although they were approved by an institutional review board. Also, although the random assignment ratio in this study was 55% treatment and 45% control, each individual student had the same probability of selection into treatment as every other student.

In this model, Y_i is the outcome of interest for student *i*, S_{ni} is a random assignment block indicator variable for student *i* in random assignment block *n* (*n* = 1, . . ., 20), and *T* is the indicator variable for treatment group assignment. This model also included baseline covariates of student characteristics ($\boldsymbol{X_i}$) that served to improve the precision of the estimates.[17]

MDRC's study was a long-term follow-up study, and survey data were collected once during high school, as well as at 48 and 96 months after expected high school graduation. Multiple outcomes were collected at both post high school time points, including self-reported data on high school graduation, postsecondary enrollment, degree attainment, and earnings, as well as other outcomes such as marriage and being custodial parents (Kemple, 2004, 2008). Across time, the findings from this study indicated that the career academy treatment did not have impacts on educational outcomes such as high school graduation or postsecondary enrollment. However, this was in part because more than 90% of both treatment and control group students graduated high school or obtained a General Educational Development certificate, which was higher than the national averages for those outcomes during that time period. However, the study did find that at both 48 and 96 months, students assigned to the treatment group had substantially higher average earnings than those assigned to the control group, and the impacts were primarily driven by young men. Specifically, the study found that participation in the career academy program helped young men obtain higher paying jobs that allowed them to work more hours (i.e., they were more likely to be full time rather than part time), such that by the end of 96 months, students from the treatment group earned an average of 11% more per year (or $2,088) than students from the control group, for a total of an additional $16,704 in the 8-year follow-up period (in 2006 dollars). Additional findings indicated that career academies also led to a higher proportion of treatment group students living independently with children and a spouse or partner. Young men in the treatment group also were more likely to be married and serve as custodial parents (Kemple, 2008).

As noted earlier, a key advantage of RCT studies for internal validity is that treatment is randomly assigned after individuals have volunteered to participate in the program, so the treatment and control groups should be the same (on average) on observable and unobserved attributes. However, this raises an additional question concerning external validity. To the extent that volunteers systematically selected into the program, the estimated effects of treatment may not reflect the effects in the broader population. If the individuals who need the treatment most are the least likely to put in the effort or have the organizational skills to apply, then the estimates from the RCT will likely understate the effects in the general population. This might occur when at-risk populations are likely to benefit the most from a program. On the other hand, if individuals who will benefit most volunteer for the program at higher rates, then the trial will overstate the effects in the population. This failure to generalize might be especially important in an initial trial, where individuals volunteering to be part of an experimental program could be quite different from the population that applies for a well-known and established program.

One way to expand external validity is to reach out and recruit subjects who otherwise might not have been represented within the sample of volunteers. In the case of the career academies study, in an effort to ensure that the evaluation would produce results that were relevant for students from underserved communities, academies in low-income minority communities were selected for participation in the study. Within those sites, deliberate efforts were made to include students perceived to be at risk for dropping out (Kemple, 1997, 2008). Impacts were estimated for three risk subgroups—low, medium, and high—and the characteristics used to define these groups were collected prior to random assignment. Although there were statistically significant impacts on earnings for

[17] Because the observables should be uncorrelated with treatment resulting from random assignment, their inclusion in the regression equation will have no influence on the treatment effect estimates but may reduce standard errors if they substantially improve the model fit.

students from all three risk groups at various points in the follow-up period, positive impacts were most consistent across time for students from the high-risk group (Kemple, 2008).

Importantly, all the study impact estimates were ITT analyses, meaning that the outcomes were measured for those who were offered a seat in the academies, rather than just those who actually enrolled in them. Because random assignment determines the offer of a place in treatment (but cannot control who takes up the offer), ITT estimates do not suffer from selection bias issues, as discussed earlier. However, it is important to understand that the ITT estimate does not reflect the effect of receiving the treatment. In fact, not all students who were offered a seat in a career academy chose to enroll in the program. As noted earlier, these students are known as "never takers." In addition, some students not offered a seat may have found ways to receive the treatment anyway, perhaps by enrolling in another school offering an academy that was not in the study sample of schools. As discussed earlier, these students are known as "always takers."

In the MDRC study, the researchers reported that 8% of the control group were always takers, and 13% of the treatment group were never takers. Therefore, when thinking about the counterfactual, these ITT estimates compared a treatment population of which 87% received career academy training to a control population of which 8% received career academy training. As with IV estimates, these ITT estimates can be inflated by dividing the impact of assignment into treatment on the likelihood of treatment, in this case 81 percentage points to obtain a TOT estimate. However, if the effect of career academies on outcomes was heterogeneous, the sorting into compliance with random assignment within treatment and control groups can influence treatment effect estimates. For example, if individuals who will benefit the most from the treatment are most likely to take up an offer, then the effects on treated individuals (TOT) will overstate the effects for the population of applicants—much like selection into applying might lead to effect estimates that overstate effects for the eligible population, as discussed earlier. On the other hand, if always takers in the control group are likely to benefit more, the TOT estimates will understate the effects for applicants. In this study, the populations of never and always takers was a relatively small share of the total study population, so selection into compliance within the treatment and control groups was probably minimal.

In addition to measuring outcomes, the teacher and student surveys collected during high school were intended to measure service contrast, which is the difference in experiences of those in the treatment and control groups (Kemple, 1997). Students in both the treatment and control groups were surveyed, and both academy and nonacademy teachers in the schools also were surveyed. Often times, in random assignment studies, the counterfactual to treatment is described as the business as usual (BAU) condition, but this may not be the case in educational settings in which reforms and new interventions are frequently introduced simultaneously and in the same overlapping settings. To isolate the effects of a given treatment, it is important to know not only what the implementation of a treatment model looks like but also what control group students are experiencing. That is, service contrast helps explain exactly the "business" of the BAU condition. Measuring service contrast allows researchers to understand whether and how the treatment and control conditions differ during the study period, which can later be used to contextualize the impact findings. For example, if a random assignment study has null findings, perhaps the program was ineffective as designed. On the other hand, null findings may arise because the experience of those in the treatment group did not actually differ greatly from the experience of those in the control group because either the implementation of the treatment program was poor or the control group students received similar services and opportunities.

To measure service contrast, surveys were collected from both teachers and students in the academy and nonacademy settings. There were statistically significant and positive differences in how treatment group students rated their experiences of teacher and peer support, motivation for school, and perceived relevance of school work compared with control group students. This provided evidence that the treatment group experience did differ

from the control group experience in important ways. In addition, there were large and positive differences in how academy teachers rated their professional experiences and levels of job satisfaction than nonacademy teachers, which also provided evidence that the learning environment for students differed across groups (Kemple, 1997).

Although all the academies were found to have fairly high levels of service contrast between the treatment and control groups in terms of integrated curriculum and exposure to careers and work-based learning, variation occurred across academies in the level of interpersonal support contrast. To better understand the relationship between interpersonal support and student outcomes, MDRC assessed variation in outcomes across sites that had high and low levels of contrast on this measure. In particular, they examined the relationship between this measure and the three student risk categories described previously. Academies with a high level of contrast on the measure of support were associated with lower levels of dropout and higher levels of curriculum completion for students in the medium-risk group. Impacts were similar across high- and low-contrast academies for students in the high- and low-risk groups, with the exception that students in the high-risk group but in the low-contrast academies were somewhat less likely to drop out and more likely to engage in career-related course taking (Kemple & Snipes, 2000).

However, as mentioned earlier, one must be aware of the possibility that the existence of the lottery influenced the counterfactual experiences on control group students, so those experiences may not represent what treatment group students would have experienced without the program. For example, in documenting service contrast, the career academy study asked many questions that captured sentiment. If students who lost the lottery were less happy with the traditional school because of that loss, they might report lower levels of satisfaction with their experiences, or if they were simply more aware of what CTE offered, they may have evaluated the traditional opportunities more critically. In the case of interpersonal support, documentation of the resources provided for support, versus self-reports related to student impressions of support, may provide more reliable results that are less likely to be influenced by the experimental context. Even when researchers ask very detailed, specific questions about counterfactual experiences, the program's existence might influence those experiences. For example, the greater salience of CTE related to the creation of the career academy may cause lottery losers to be more active in seeking CTE type experiences.

In many RCT studies, including the career academies study, program implementation data also are collected, often using qualitative methods such as site visits, interviews, and focus groups. The implementation study piece usually provides information on how well a program was implemented in the treatment condition. The collection of data to understand implementation and service contrast differs, with the former used to understand whether a program or model is implemented with fidelity to its initial design or conceptualization. By definition, implementation data are collected about the treatment condition only, whereas service contrast data are collected about both the treatment and control conditions to measure differences between them. An early implementation study found that all participating sites were able to implement the main elements of the model, including the school-within-a-school approach, a curriculum that blended academic and career-themed coursework and partnerships with local employers (Kemple & Rock, 1996).

MDRC also used the data from the implementation study to examine the relationship between how well elements of the model hypothesized to produce impacts on student outcomes were implemented, plus student experiences across both academies and students. The three components examined included interpersonal support for students, combined academic and CTE curricula, and exposure to career awareness and work-based learning. The study found that regardless of assignment condition, students who experienced high levels of interpersonal support in Grades 9 and 10 were less likely to drop out of high school and more likely to complete the core curriculum. They also found that students who engaged in career awareness and work-based learning were more

engaged in school, prepared to graduate, and more likely to go to college. A weaker but still positive association was reported between students' engagement in school and an integrated curriculum (Kemple & Snipes, 2000).

Although MDRC's effort shed light on relationships between components of the career academy model and student outcomes, it could not make a causal connection between specific elements of the model and student outcomes because students were randomly assigned to an entire package of activities. Thus, the impact estimates represented only the effect of participating in the total career academy experience. Although the MDRC study advanced the field of knowledge about one CTE model, it would be useful to also know which elements of the program in particular might be related to specific outcomes. A weakness of the correlational analysis identified by MDRC is that it cannot provide causal evidence on which program elements might be the most useful if one wishes to expand or scale up a CTE program. For example, as noted earlier, the MDRC study found substantially larger treatment effects in career academies with better implementation of work-based learning. However, career academies that did a better job of implementing work-based learning may differ in unobserved ways that confound such descriptive analyses. As a result, it is unclear whether work-based learning is driving the positive outcomes or unobservables associated with schools that successfully implemented work-based learning. If incorporated into the design, RCTs can provide causal evidence related to the specific elements of treatment by (a) incorporating multiple treatment arms that include different aspects or treatment or (b) dividing treatment into many elements and randomizing the inclusion of specific elements, such as in factorial designs.

The final results of the study also generated additional interest in understanding the dosage level of the program that students would need to experience to receive a benefit from participating. Given that the academy experience is fairly intense, requiring participation for 3–4 years, it may be policy relevant to understand whether the same or similar impacts can be achieved with lower levels of commitment by students and investments by schools. This is a particularly important issue because large and sustained impacts on labor market outcomes were observed, despite the fact that only 53% of the treatment group participants reported actually completing the entire academy experience (i.e., still enrolled in the academy at the end of scheduled 12th-grade year; Kemple, 2004). To that end, Lindsey Page conducted a further analysis of MDRC's study data to try to isolate elements of the program that led to student success. Using a principal stratification analysis in which she measured outcomes for students with different levels of participation, Page found that the highest levels of monthly earnings were accrued by those students who remained enrolled in the program for the longest amount of time (Page, 2012).

It is important to note, however, that the counterfactual condition (i.e., the experience of the control group in this case) may be particular to this study and may not generalize to other random assignment studies of CTE programs or even replications of this or similar random assignment studies. For example, another kind of random assignment study with a very different comparison condition is one that relies on naturally occurring lotteries, rather than those designed and managed by the research team. An example of this kind of study is an evaluation of the P-TECH 9-14 model of schools in New York City, currently being conducted by MDRC (Rosen et al., 2020). P-TECH 9-14 schools are 6-year high schools in which students earn both a high school diploma and an associate's degree from a partner postsecondary institution. These schools are organized around a career theme, and each one has an industry partner that provides professional learning experiences, such as mentoring, job shadowing opportunities, and internships. Rather than researcher-managed random assignment, the P-TECH 9-14 study identified randomly assigned treatment and control groups by leveraging lotteries created by the high school placement algorithm used by New York City's centralized high school admissions process. In this case, treatment group students randomly won an opportunity to attend one of seven P-TECH 9-14 schools, and the control group students randomly lost the opportunity to attend these same schools.

Many of the research properties in the P-TECH 9-14 study are the same as those described earlier (e.g., those associated with the interval validity of random assignment), but the counterfactual condition to which control group

students were assigned was very different. The key difference is that in the career academy study, control group students were assigned a seat in the same high school as the treatment group students but received a different experience. By contrast, in the lottery-based study, treatment group students received an offer to enroll in a self-contained school, whereas control group students received offers of admission to 399 different high schools across New York City, which is almost all other high schools in the district. Further, about one fourth of these schools offered a range of CTE opportunities, but many others did not provide any CTE. Therefore, the experience that constituted BAU in the lottery-based random assignment design study was much vaster and more difficult to define, particularly in terms of understanding the service contrast and what control group students experienced compared with the career academy study.

Although random assignment design studies are generally the gold standard for research, one potential threat to internal validity is sample attrition. A large amount of sample attrition across time, or particularly uneven attrition from the treatment and control groups, can potentially bias the study findings. In this case, bias could happen if substantial numbers of students assigned to the control group left their schools all together because they did not receive a spot in the academy. This kind of large exodus would be problematic if it then inhibited the collection of outcome data from those students because that would mean impact estimates would become a comparison between those students assigned to treatment and those assigned to control and did not have the motivation to seek alternative schooling. In the career academy study, even though many treatment group students reported not graduating from the academy at the end of 12th grade, the fact that the researchers were still able to collect outcome data from 82% of the treatment group students and 80% of the control group students meant that attrition from the sample was neither uneven nor unreasonably high across the assignment groups.[18]

Finally, it also is worth noting that MDRC is currently undertaking a replication study of the original career academy study.[19] Broadly, the replication seeks to repeat an existing study with a different sample and setting in an effort to obtain the same or similar results. In this replication study of the original study, students were randomized and entered the academies between 2017 and 2020, which is 25 years after the sample of students in the original study experienced high school. Given that much has changed in terms of both education policy and high school experiences in the last several decades, one open question in the replication study is whether and how the counterfactual condition has changed across time. In the mid-1990s, No Child Left Behind and other federal accountability legislation for education had yet to be enacted, and the academic standards for the average high school were arguably less rigorous in many states than they are today. So even if today's career academies are operated in similar ways to how they were implemented in the past, if the changes wrought by the evolving federal and state educational policy landscape have changed the educational experience of students in the BAU condition or control group, this may impact the ability of the current study to replicate findings. For this reason, the "comparison to what" question is always critical for understanding and interpreting results from any random assignment design study.

In summary, with random assignment, the treatment and control groups are comparable on average by construction. However, this design raises concerns about external validity. For example, if the individuals who need the treatment most are the least likely to put in the effort or have the organizational skills to apply, the estimates from an RCT will likely understate the effects in the general population. A second issue that can arise with random assignment is compliance. Although selection was modest in the Career Academies study, the ITT will understate the effects of receiving treatment with certainty. Further, if the effect of Career Academies on outcomes is heterogeneous, sorting into compliance with assignment can influence treatment effect estimates

[18] For further discussion of acceptable levels of attrition bias, see the What Works Clearinghouse *Standards Handbook,* Version 4.0.
[19] See https://www.mdrc.org/project/next-generation-california-partnership-academies#overview for more information.

because people who actually receive the treatment are different from treated individuals who failed to comply. Finally, the MDRC study explicitly measured service contrast by capturing differences between treatment and control group students in their educational experiences. They found that treatment group students provided higher ratings on teacher and peer support, motivation for school, and perceived relevance of schoolwork. However, the experiences of these control group students may not be representative of the counterfactual for treatment group students if the creation of the program influenced either the decisions or responses of the control group students. Finally, in other environments with more school choice opportunities (especially with multiple CTE-oriented choices), the counterfactual experiences of control group students may be much more similar to those of treatment group students, possibly eroding the treatment effect of the specific program, even if CTE experiences improve student outcomes.

Section IV. Summary and Discussion

In this paper, we reviewed four studies that used different strategies in different contexts to uncover the effects of CTE on student outcomes. The first study examined the LLDI in California, where districts were systematically selected for the development of certified CTE pathways and students were free to select into or out of those programs. As a result, this study could control only for those student attributes captured in traditional administrative data. The second study examined stand-alone career academies in CPS, identifying the effect of attending an academy by instrumenting for attendance using physical proximity between each student's residence and the closest career academies. Next, CTHSS was evaluated using an RD approach, exploiting the fact that students are admitted in order based on a well-documented application score. Finally, we discussed MDRC's study of within school career academies, where MDRC randomized the admission of student applicants into the programs.

Several common themes arise from these case studies. First (and foremost) is the question of internal validity: Have the studies identified the causal effect of the program on participants? The LLDI is the weakest on internal validity because the evaluation did not contain any strategy for isolating program effect estimates from the unobservables that influenced selection into the program. Indeed, the study found that the inclusion of basic student controls from the administrative data eroded estimated treatment effects by roughly 50%. If estimating causal treatment effects is the goal, then future CTE studies that employ regression or matching methods with observational data should examine how sensitive treatment effect estimates are to the inclusion of control variables. If the inclusion of control variables significantly attenuates treatment estimates, this would suggest that the inclusion of additional (possibly unobserved) controls might completely eliminate any treatment effects.

The CPS study tried to avoid selection bias by using distance as an instrument for participation. However, residential location might correlate with many of the same unobservables that caused bias in a simple observational study like the LLDI, violating the assumptions in the IV model. Thus, researchers should examine instruments for balance over observables under the argument that if the instrument is uncorrelated with observed neighborhood or student attributes, then it is likely also uncorrelated or at worst weakly correlated with the unobserved attributes. If possible, researchers should examine whether the magnitude of treatment effects varies over the value of the instrument. If treatment effects are heterogeneous across the instrument, researchers should document the distribution of treatment take-up over the domain of the instrument to assist in the interpretation of their estimates.

Both the CTHSS and the MDRC career academy studies had much higher internal validity than the LLDI or CPS studies. The CTHSS study addressed the threat of selection bias by exploiting that fact that all 16 schools in CTHSS used a rule-based admission system and have been oversubscribed since 2006. This leads to an RD identification strategy that compares nearly identical students on either side of an admission threshold. The

MDRC career academy study addressed the threat of selection bias by randomly assigning students to the treatment and control conditions. As noted previously, the RD strategy used in the CTHSS study exploited a form of local randomization, where the randomization was conditional on the running variable falling within a narrow bandwidth around the cutoff or threshold. Researchers using RD methods to identify causal treatment effects should always test for balance by examining whether the observables are assigned randomly on either side of the threshold within the specified bandwidth. Researchers also should examine whether estimated treatment effects are relatively stable when estimated using the optimally chosen bandwidth and then several narrower bandwidths that are less prone to selection bias. Finally, although RD studies have strong internal validity, their external validity or generalizability may be a larger concern because the effect is estimated for only those local students who are within a given bandwidth. Thus, researchers should carefully characterize the marginal individual for whom the RD treatment effects are being identified. This will typically involve comparing the observable characteristics of individuals who are close to the cutoff score to the broader population of interest.

Both the CTHSS and the MDRC studies followed an ITT strategy, where identification arises from comparing admitted to nonadmitted students, even though students may not perfectly comply with these rule-based admission decisions. Take-up in the MDRC study was very high, so the subsample of compliers is quite close to the sample of treated students. However, in the CTHSS study, being above the threshold raised attendance by only 58 percentage points, so the treated students or compliers may differ in substantial ways from the never takers who were admitted and did not attend. Finally, all four studies may suffer from attrition bias, especially as researchers attempted to follow people into the labor market, but the use of high-quality administrative data as in the CTHSS study can mitigate those concerns by tracking individuals using permanent identifiers (e.g., Social Security numbers).

As noted previously, CTE programs often involve a vector of treatments, such as greater exposure to CTE coursework, career-themed pathways, work-based learning, and greater integration of core courses such as mathematics and English courses with CTE material. To the extent possible, any evaluation of CTE should keep track of which items on the CTE menu are, in fact, experienced by each CTE student and their counterfactual counterpart (Maxwell et al., 2017). Unfortunately, however, many strategies for uncovering the causal effects of programs are limited in their ability to address which aspects of the program contribute to success. In some cases, descriptive analyses can be used to make some progress in terms of identifying effective elements of treatment. For example, the CTHSS study examined the CTE opportunities available to students who were not admitted to a school in the CTHSS based on CTE offerings at their residential high school, concluding that a substantial share of the benefits of attending a school in the CTHSS arose from the stand-alone nature of the system. Similarly, the MDRC career academy study conducted an implementation analysis that captured variation in interpersonal support, work-based learning, and integration of CTE and academic curricula. Notably, they found substantially larger treatment effects in career academies with better implementation of work-based learning. Of course, academies that did a better job of implementing work-based learning may differ in unobserved ways that confound such descriptive analyses. Going forward, RCTs should, if possible, incorporate variation in the experimental treatment across students, sites, or programs to identify the most effective aspects of CTE programs, by either including multiple treatment arms or randomizing the elements included within treatment, such as in a factorial design.

The second important issue to consider is external validity (i.e., the generalizability of the study results). In many cases, studies with strong internal validity provide evidence that can be harder to generalize than studies that document outcomes from a broad population participating in a large, long-standing program. For example, the LLDI was a new program created to provide additional investments and raise the quality of existing CTE programs. As a result, the LLDI certified pathways almost certainly were drawn from well-established programs, and it is very difficult to know whether any documented effects would or would not have been found in those

preexisting programs if the LLDI had not existed. MDRC's investigation of career academies avoided the problems arising with new programs or an incremental modification to existing programs by conducting a study of the effects of participation in an existing within-school career academy. This strategy, however, carried its own limitation because the MDRC study effectively was an evaluation of a small number of volunteer schools. Whether similar effects might be found in a broader population of schools is unknown. The schools in the CTHSS study provided CTE education at scale, educating 7%–8% of all high school students in Connecticut. In this case, the limitation arose from the identification strategy itself because the authors could estimate the effect on only those students who were very near to the admission threshold. Luckily, the admission system exhibited considerable variation in the threshold across time and across schools, and the positive effects on students appeared relatively stable across those thresholds, thus supporting the generalizability of those estimates.

Finally, the third key issue that arises is the counterfactual experiences of the control group students, or what MDRC described as the service contrast, and whether those experiences are reflective of what the treatment group would have experienced without the program. The counterfactual experience was clearly a concern in the LLDI study, where the magnitude of estimates was sensitive to which students were included in the control group and the fact that treatment group students might have been much more likely to have participated in any preexisting noncertified CTE programs if their district had not been selected for LLDI. In CPS and CTHSS, the counterfactual was more clear because the studies involved well-established, stand-alone career or technical schools, and a substantial fraction of students attended their traditional, residentially assigned school. In several ways, lottery-based admission provides an excellent opportunity for documenting the counterfactual experiences of the treated students because the treatment and control groups are both randomly drawn from the same population. In principle, RD approaches have a similar advantage except that the focus on a small bandwidth often implies that the data are quite thin when trying to document control group experiences. In both cases, however, the introduction of a new program may create differences between control group experiences and the counterfactual experiences that would have occurred without the program. Regardless, careful documentation of differences in resources, peers, CTE classes, and other CTE-focused activities (e.g., work-based learning) is important for understanding the impacts of CTE. The need for such documentation becomes even more critical in districts with a wealth of choice options. As discussed earlier, the evaluation of P-TECH 9-14 schools in New York City may be difficult given the wide array of alternative options available, including options that provide CTE.

References

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy, 113*(1), 151–184.

Bertrand, M., Mogstad, M., & Mountjoy, J. (2019). *Improving educational pathways to social mobility: Evidence from Norway's "Reform 94"* (NBER Working Paper No. 25679). Cambridge, MA: National Bureau of Economic Research. Retrieved from https://www.nber.org/papers/w25679?sy=679

Billings, S., Deming, D., & Ross, S. L. (2019). Partners in crime. *American Economic Journal: Applied Economics, 11*(1), 126–150.

Brunner, E., Dougherty, S., & Ross, S. (2019). *The effects of career and technical education: Evidence from the Connecticut Technical High School System.* Providence, RI: Brown University, Annenberg Institute. Retrieved from https://www.edworkingpapers.com/sites/default/files/ai19-112.pdf

Cullen, J. B., Jacob, B. A., & Levitt, S. D. (2005). The impact of school choice on student outcomes: An analysis of the Chicago Public Schools. *Journal of Public Economics, 89*(5–6), 729–760.

Dougherty, S. M. (2018). The effect of career and technical education on human capital accumulation: Causal evidence from Massachusetts. *Education Finance and Policy, 13*(2), 119–148.

Hemelt, S. W., Lenard, M. A., & Paeplow, C. G. (2019). Building bridges to life after high school: Contemporary career academies and student outcomes. *Economics of Education Review, 68,* 161–178.

Kemple, J. J. (1997). *Career academies: Communities of support for students and teachers: Emerging findings from a 10-site evaluation.* New York, NY: MDRC. Retrieved from https://files.eric.ed.gov/fulltext/ED415403.pdf

Kemple, J. J. (with Scott-Clayton, J.). (2004). *Career academies: Impacts on labor market outcomes and educational attainment.* New York, NY: MDRC. Retrieved from https://www.mdrc.org/sites/default/files/full_49.pdf

Kemple, J. J. (with Willner, C. J.). (2008). *Career academies: Long-term impacts on labor market outcomes, educational attainment, and transitions to adulthood.* New York, NY: MDRC. Retrieved from https://www.mdrc.org/publication/career-academies-long-term-impacts-work-education-and-transitions-adulthood

Kemple, J. J., & Rock, J. L. (1996). *Career academies: Early implementation lessons from a 10-site evaluation.* New York, NY: MDRC. Retrieved from https://files.eric.ed.gov/fulltext/ED398401.pdf

Kemple, J. J., & Snipes, J. (2000). *Career academies: Impacts on student engagement and performance in high school.* New York, NY: MDRC. Retrieved from https://www.mdrc.org/sites/default/files/Career_Academies_Impacts_on_Students.pdf

Maxwell, N. L., Whitesell, E., Bellotti, J., Leshnick, S., Henderson-Frakes, J., & Berman, D. (2017). *Youth CareerConnect: Early implementation findings.* Princeton, NJ: Mathematica Policy Research.

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). New York, NY: Cambridge University Press.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business and Economic Statistics,* 37(2), 187–204.

Page, L. (2012). Understanding the impact of career academy attendance: An application of the principal stratification framework for causal effects accounting for partial compliance. *Evaluation Review, 36*(2), 99–132.

Passarella, A. (2018). *The necessary components of an effective career and technical (CTE) program* (Policy Brief). Baltimore, MD: Johns Hopkins School of Education. Retrieved from https://edpolicy.education.jhu.edu/wp-content/uploads/2019/01/CTE-Published-Final-CFC-FINAL.pdf

Rosen, R., Byndloss, D. C., Parise, L., Alterman, E., Dixon, M., & Medina, F. (2020). *Bridging the school to work divide: Interim implementation and impact findings from New York City's P-TECH 9-14 schools.* New York, NY: MDRC. Retrieved from https://www.mdrc.org/sites/default/files/P-TECH_Report_2020.pdf

Rosen, R., Visher, M. & Beal, K. (2018). *Career and technical education: Current policy, prominent programs, and evidence.* New York, NY: MDRC. Retrieved from https://www.mdrc.org/publication/career-and-technical-education

Silliman, M., & Virtanen, H. (2019). *Labor market returns to vocational secondary education* (ETLA Working Papers 65). Helsinki, Finland: ELTA Economic Research.

Warner, M., Caspary, K., Arshan, N., Stites, R., Padilla, C., Patel, D., . . . Adelman, N. (2016). *Taking stock of the California Linked Learning District Initiative: Seventh-year evaluation report.* Menlo Park, CA: SRI International. Retrieved from https://www.hsredesign.org/wp-content/uploads/2018/07/sri_year_7_linked_learning_evaluation_report_0.pdf

Witzen, B. H. (2019). *The effect of high school career and technology education on postsecondary enrollment and early career wages.* Baltimore, MD: Maryland Longitudinal Data System Center. Retrieved from https://mldscenter.maryland.gov/egov/Publications/ResearchReports/FinalCTEReportOctober2019.pdf

Zanutto, E. L. (2006). A comparison of propensity score and linear regression analysis of complex survey data. *Journal of Data Science, 4*(1), 67–91.

1400 Crystal Drive, 10th Floor
Arlington, VA 22202-3239 | www.air.org

cteresearchnetwork.org

The American Institutes for Research (AIR) and its partners—the Association for Career and Technical Education (ACTE), JFF, and Vanderbilt University—serve as the CTE Research Network Lead. The Network Lead provides network administration and coordination as well as research, training, and dissemination to increase the number and quality of CTE impact evaluations and strengthen the field's research capacity.

www.air.org

www.acteonline.org

www.jff.org

www.vanderbilt.edu

12670_09/20