Abstract Title Page

Title:
Methodological Complications of Matching Designs under Real World Constraints: Lessons from a Study of Deeper Learning

Authors and Affiliations:
Kristina Zeiser, American Institutes for Research
Jordan Rickles, American Institutes for Research
Michael S. Garet, American Institutes for Research

Abstract Body

Problem / Background / Context:
Educational leaders and practitioners often desire research evidence on multifaceted school and instructional reform strategies currently being implemented in heterogeneous settings. There is particular interest in high school initiatives targeting college and career readiness. But providing timely information on the impact of such initiatives can be challenging, especially since the effects may unfold over multiple years of high school and subsequent postsecondary years.

While in principle it would be possible to conduct a prospective random assignment study, in which students are randomly assigned to different high schools at entry to 9th grade (perhaps through lottery), such a prospective design would take many years to produce data on key outcomes. Thus, there is an interest in quasi-experimental matching designs, relying on retrospective data to match students at entry to high school. The research methodology literature provides direction on how to best design matching studies that facilitate causal inferences, but there is little discussion about the many methodological issues and trade-offs researchers face when designing and implementing such studies in complex field settings, especially when focusing on initiatives that unfold over time with multiple, diverse outcomes. Such research projects are further complicated by the fact that the schools in which we conduct studies are facing their own budgetary and time constraints, with burdens placed on them from district requirements and other external factors. Limited systematic dissemination of these complexities, solutions, and implications can handicap future research endeavors and limit research usefulness for practitioners and decision-makers.

Purpose / Objective / Research Question / Focus of Research:
To help researchers understand potential issues one can encounter when conducting propensity matching studies in complex settings, this paper describes methodological complications we faced when studying schools using deeper learning practices (described below) to improve college and career readiness. In particular, the paper focuses on five questions:

1. What complications in the sample selection process can arise when treatment conditions are not well defined?
2. What complications in the data collection process can arise when studying schools across multiple school districts (and states)?
3. What complications arise in a propensity-matching context when the study requires administering study measures in addition to relying on administrative record data?
4. What complications arise from missing data due to attrition, non-consent, and non-response, when students are followed longitudinally?
5. What implications do the above complications have for interpreting and communicating findings to decision-makers and practitioners?

In addressing these questions, we use our experience from a study of deeper learning funded by the William and Flora Hewlett Foundation as a guide to inform future studies about issues to anticipate, options for addressing those issues, and implications such issues can have for internal and external validity.

Improvement Initiative / Intervention / Program / Practice:
This paper is based on our experience conducting a proof-of-concept study to determine if

students attending high schools with a developed approach to promoting deeper learning experience greater deeper learning opportunities and outcomes than they would have had they not attended these schools. The deeper learning approach refers to instruction, supports, and structures that focus on: mastery of core academic content, critical thinking and complex problem-solving, effective communication, collaboration skills, learning how to learn, and development of an academic mindset (Trilling, 2010).

Setting:
The study uses data from high schools located in six districts across two states: California and New York.

Population / Participants / Subjects:
The study included students attending 20 treatment high schools associated with 10 school networks aligned with a deeper learning approach ("network schools"). Participating schools were considered moderate- to high-implementing deeper learning schools according to network staff. The study also included students attending 15 comparison schools selected to have similar student characteristics as the treatment schools ("non-network schools"). Due to our focus on the effects of school-wide reform on students from disadvantaged backgrounds, all of the schools included in the study had student populations in which 25 percent or more of students were eligible for free- or reduced-price lunch according to the Common Core of Data (CCD). Additionally, we focused on schools in operation since 2007-08, so that three cohorts of students would have had the opportunity to graduate by spring 2013.

Within these schools, the study included high school students who entered as a first time 9th grader in the 2007-08 (cohort 1) through 2011-12 (cohort 5) school years. Students included in the sample must have had 8th grade district administrative record data, so that we could match students on entry to high school and control for pre-high-school characteristics. Due to issues related to school recruitment and student consent (described below), student-level data collection occurred within 12 network schools and 10 non-network schools. A description of the treatment and comparison school students is provided in Appendix Table 1.

Research Design:
The study used a quasi-experimental matched design, in which each network school was matched to a non-network school within the same (or neighboring) school district. We then used inverse probability of treatment weighting (IPTW) to adjusting for pre-existing differences between network school students and non-network school students, in which weights were assigned to estimate the average treatment effect on the treated (ATT). School-level matching was based on school characteristics from the CCD, and student-level propensity scores were estimated separately for each 9th grade cohort, based on the student characteristics listed in Table 1.

Data Collection and Analysis:
The study collected data from multiple sources, though we focus on student-level data in this paper.[1] Student-level characteristics and outcomes were measured through extant district data,

[1] Data collection also included teacher-level survey data and school case studies.

student surveys, the OECD PISA-based Test for Schools, and postsecondary data from the National Student Clearinghouse (NSC). Details about the student-level data sources and coverage are provided in Table 2.

The analysis focused on six types of student outcomes: opportunities to learn (student survey), intra- and inter-personal competencies and dispositions (student survey), mastery of core content and complex problem-solving skills (OECD PISA-based Test for Schools scores), academic achievement (state test scores), high school graduation (district record data), and college enrollment (NSC data). For each outcome, the ATT was estimated for each matched pair based on a doubly-robust weighted regression model including student characteristics and 8th grade test scores as covariates (Funk et al., 2011). In addition to IPTW weights, the models included weights that accounted for sampling, attrition/non-consent, and non-response (discussed below).

A precision-weighted, fixed effects meta-analysis was used to calculate overall average effects across the matched pair-specific effect estimates. For this paper, we use the Shadish, Cook, and Campbell (2002) validity framework to describe the implications of different complications and decisions during the data collection and analysis phases of the study.

Findings / Outcomes:

In carrying out the study, we faced a series of trade-offs between the preferred research design and practical feasibility. Specific complications and implications are outlined in Table 3. We discuss some of the key points below.

Complications in the sample selection process when treatment conditions are not well defined. We were interested in the impact of attending schools focused on deeper learning, but there are multiple approaches to deeper learning and considerable variation from school to school in curricula, instruction, and organization. We dealt with this challenge by asking the 10 deeper learning network organizations to identify potential candidate schools for the study. Narrowing our analysis to a specific set of schools that networks identified as moderate- to high-implementers of the deeper learning model improved construct validity by providing a clearer and defensible definition of the "treatment" practices being tested. But this approach also required us to rely on network judgments of the meaning of "moderate and high implementers," and it also limits the external validity of the study by focusing the work on the impact of a particular set of schools. In addition, we sought to recruit non-network schools that operated within the same district as the network schools, thus limiting our sample to schools that operated in large districts that had at least one network high school and one non-network high school. Selecting districts with multiple network and non-network schools saved resources in terms of collecting district administrative data as well.

Complications in the data collection process when studying schools across multiple school districts (and states). Differences in district extant data availability limited our ability to conduct identical analyses for schools in different districts. Restricting propensity score models and outcome models to include only those covariates measured in common across districts had the potential to limit internal validity for some pairs. By matching schools within district and conducting separate within-pair propensity score models and outcome models, we could utilize the full breadth of covariates available. This approach does, however, complicate transparency and interpretation because some pair-specific estimates are based on a richer set of covariates than others.

Complications when the study requires administering study measures in addition to relying on administrative record data. Most propensity matching studies in the literature rely primarily or entirely on administrative record data. But many of the 21st century skills that researchers would like to study are not contained within extant data sources. Thus, we needed to conduct extensive data collection, placing significant burden on schools that were already impacted by external conditions, including fiscal retrenchment and high-stakes accountability. As a result, recruiting schools for the study was challenging and meant that we could not necessarily include the schools that were best matched in a statistical sense. In addition, parent consent requirements for data collection in some jurisdictions reduced the pool of students with outcome data.

Complications that arise from missing data due to attrition, non-consent, and non-response, when students are followed longitudinally. Since we administered surveys and tests within the school setting, we were only able to actively collect data from students who were still attending the school they entered in the 9th grade. In addition to the attrition that normally occurs within schools, several of the schools that participated in our study required active parental consent (rather than passive consent) for students to participate in data collection activities, further reducing the number of cohort students who were eligible to be included in our data collection efforts. In order to adjust for attrition, non-consent, and then survey and test non-response, we calculated inverse-probability weights. Applying these attrition, non-consent, and non-response weights, along with the IPTW weights, in our analysis improved both internal and external validity to the extent that weights were able to appropriately adjust for the factors related to students' likelihood of missing outcome data.

Implications for interpreting and communicating findings to decision-makers and practitioners. Interpreting and communicating research findings becomes more difficult, yet more important, as design and analysis complications mount. As discussed above, the varied research decisions required to complete the study have implications for the validity of findings. As a result, it is important to find ways to communicate the appropriate level of confidence in the findings for decision-makers. In our case, we stressed that the results are limited to the set of schools in the proof-of-concept study. In addition, by presenting evidence of the heterogeneity of findings across sites, researchers can temper the conclusions that should be drawn from a study in which treatment effects were not observed consistently across sites.

Conclusions:

The *Study of Deeper Learning* illustrates the tensions between the realities associated with conducting research based on matching methods and the simplicity of textbook study design and analysis. In this paper, we examined different areas in which the nature of the intervention being studied and the constraints and challenges of working with districts and schools affected design and analysis decisions and the potential threats to the validity of research findings. Our experience suggests that more attention should be given to the kinds of methodological adaptations needed to take these constraints and challenges into account. In addition, we must strengthen ways in which methodological compromises and complexities are appropriately communicated to decision-makers so they can place the research findings in proper context.

Appendices

Appendix A. References

Funk, M. J., D. Westreich, C. Wiesen, T. Sturmer, M. A. Brookhart, and M. Davidian. (2011). "Doubly robust estimation of causal effects." *American Journal of Epidemiology* 173(7): 761-767.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and Quasi-Experimental Designs for Generalized Causal Inference (2nd ed.). Boston, MA: Houghton-Mifflin.

Trilling, B. (2010). *Defining Competence in Deeper Learning* (draft report to the Hewlett Foundation). Menlo Park, CA.

Appendix B. Tables and Figures

Table 1. Student characteristics for treatment and comparison schools[1]

	Treatment	Comparison
Achievement Test Scores		
8th grade math (standardized)	-0.518	-0.501
8th grade ELA (standardized)	-0.437	-0.347
ELL Status	9.8%	8.6%
IEP status	3.5%	2.9%
Female	53.5%	50.9%
Race/Ethnicity		
black	4.6%	3.9%
Hispanic	71.6%	69.4%
white	3.1%	3.1%
FRPL status*	64.9%	62.7%
Parents' Education*		
Less than HS	22.3%	18.0%
High School	19.6%	22.4%
Some College	12.1%	16.7%
College Education	10.1%	12.7%
Declined	15.4%	6.5%

[1]Data are presented for students who were eligible to take the survey and OECD PISA-based Test for Schools, who entered grade 9 in 2009-10 and 2010-11. Other covariates included in propensity score models that are not listed here include 8th grade science test scores, 7th grade math and ELA test scores, attendance rates in the 8th grade, and student age at entry into the 9th grade.
*Data are not available within all districts included in the study.

Table 2. Description of study data sources

Data Source	Sample	Number of Schools	Number of Students	Response rate (of sampled students)	Description
Student Survey	Entering 9th grade students from 2009-10 to 2010-11	22	1,267	76%	Measures students' exposure to deeper learning practices as well as dispositional outcomes such as self-efficacy and academic engagement.
OECD PISA-based Test for Schools	Entering 9th grade students from 2009-10 to 2010-11	20	1,762	61%	Measures students' higher-order skills in reading, mathematics, and science.
District extant data	Entering 9th grade students from 2007-08 to 2011-12	28	22,635	100%[1]	Includes student demographics, middle school state test scores, high school state test scores, and graduation.
National Student Clearinghouse (NSC)	Entering 9th grade students from 2007-08 to 2009-10	28	16,968	100%	Measures students' college enrollment and institution type from 2011 to 2013.

[1]Students' graduation and achievement test scores are measured within district data. Students who leave the district prior to high school graduation are classified as "not graduated within the same district" within four years.

Table 3. Summary of study complications, decisions and validity implications

Study Phase	Complication	Study Decision	Validity Implications
Sample Selection	Treatment condition not well defined	Use network leaders to identify model deeper learning schools	• External validity limited to similar schools • Construct validity improved by better defined treatment condition
Sample Selection	Heterogeneous treatment/comparison conditions across schools	Conduct within-pair analysis & use meta-analysis for average effect estimates	• Treatment definition clarified by estimating pair-specific effects • External validity limited because average effects may not generalize to a specific version of treatment
Sample Selection	Interest in postsecondary outcomes for entering high school cohorts, but only a 3-year study	Study established deeper learning schools with at least 3 graduating cohorts at start of study	• External validity limited to established schools • Construct validity improved by expanding analysis to additional outcomes
Sample Selection	Interest in multiple outcomes based on multiple data sources raised data collection burden and limited pool of interested comparison schools	Allow "sub-optimal" school matches when best available comparison school declined participation	• Internal validity limited by quality of school-level match • Statistical conclusion validity improved by increased power
Sample Selection	Limited comparison school availability resulted in imperfect school-level matches	Use IPTW to weight comparison schools students to be representative of treatment school students	• Internal validity improved by more appropriate comparison group for treatment students, but study still relies on assumption of selection on observables
Extant Data Collection	District extant data not available prior to comparison school selection process	Use school-level CCD to identify comparison schools	• Internal validity limited because school-level matching did not include potentially important covariates captured in district record data but not the CCD
Extant Data Collection	Available extant district record data differed across schools in different districts & states	Match schools within district and conduct propensity score estimation & analysis separately within-pairs; use meta-analysis to combine results across pairs	• Internal validity improved over analysis restricted to "common" covariates, because within-pair covariate adjustment utilized more covariates • Internal validity assumptions more difficult to communicate because some pair-estimates based on richer set of data than others • External validity improved over study limited to a single district or state

Study	Complication	Study Decision	Validity Implications
Missing Data	Student attrition, non-consent, & non-response resulted in missing outcome data for a significant portion of the targeted first time 9th grade population	Use attrition and non-response weights so measured sample is representative of target population	• Internal validity improved if weights properly account for missing data, but increased reliance on missing data mechanism assumption • External validity improved by better representing target population
Communication of Findings	Heterogeneity of treatment effect needs to be communicated in addition to the average treatment effect	Present pair-specific results in addition to average treatment effect from meta-analysis	• Threat for perception of external validity: practitioners may conclude that results are generalizable to all schools implementing deeper learning reforms. With a larger number of sites, a random effects meta-analysis would be possible.
Communication of Findings	Given complex sampling and analysis strategies, results need appropriate caveats and a straightforward interpretation of methods	Devote study resources to explication of methods in friendly language	• Threats for perception of internal and external validity: studies that use jargon to describe research methods are likely to make practitioners question if they can trust what the study says and to which types of schools the results should be ascribed