4-27-2020

Using Nonexperimental Methods to Address Noncompliance

Daniel Litwok
Abt Associates, Social & Economic Policy Division

Upjohn Institute working paper ; 20-324

Citation

Litwok, Daniel. 2020. "Using Nonexperimental Methods to Address Noncompliance." Upjohn Institute Working Paper 20-324. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research. https://doi.org/10.17848/wp20-324

W.E. UPJOHN INSTITUTE
FOR EMPLOYMENT RESEARCH
Researching the causes and consequences of unemployment

Using Nonexperimental Methods to Address Noncompliance

Upjohn Institute Working Paper 20-324

Daniel Litwok
Abt Associates, Social & Economic Policy Division
email: dan_litwok@abtassoc.com

April 2020

ABSTRACT

Background: There is widespread interest in understanding which components of a job training program generate favorable impacts. Given constraints such as time, money, and logistics associated with experimental evaluation, program evaluators might choose a nonexperimental approach to answering such a question. This study tests the performance of such methods against an experimental benchmark. The analysis uses methodological guidance from the within-study comparison literature to compare experimental to nonexperimental estimates in the context of the Health Profession Opportunity Grants (HPOG) program, which experimentally tested the incremental impact of three specific program enhancements.

Methods: The analysis compares estimates of the incremental impact for those who receive HPOG with a program enhancement to the standard HPOG program. The experimental benchmark for the incremental impact comes from two-stage least squares with random assignment as an instrumental variable for enhancement take-up. Then, ignoring the randomly assigned conditions, the analysis estimates the counterfactual for those who "take up" the enhancement using ordinary least squares and inverse propensity weighting. The analysis also tests whether adding information that is only available due to the experiment—who complied with their randomization status and who did not—improves the nonexperimental estimates. The analysis compares these estimates using statistical tests recommended by the within-study comparison literature.

Results: Despite little statistical power, the nonexperimental approaches conclusively fail to replicate the incremental impacts from the experiment for two of three enhancements. Furthermore, adding information about compliance status does not meaningfully change the findings.

JEL Classification Codes: C310, J240

Key Words: Treatment effects, Experimental methods, Nonexperimental methods, Within-study comparison

Acknowledgments: The author gratefully acknowledges support from the Abt Associates Work in Progress Seminar and the W.E. Upjohn Institute for Employment Research through Early Career Research Grant No. 18-156-09. The Office of Planning, Research, and Evaluation (U.S. Department of Health and Human Services, Administration for Children and Families) provided access to study data. This work benefited from substantive discussions with Andrew Clarkwest, David Judkins, Jacob Klerman, Trang Nguyen, Austin Nichols, and Laura Peck.

Upjohn Institute working papers are meant to stimulate discussion and criticism among the policy research community. Content and opinions are the sole responsibility of the author.

Job training initiatives in the United States use a variety of approaches to support participants, provide training, and connect job candidates with employers. Program models differ substantially on many dimensions, such as the mix of products and services, sources of funding, and target population. When a program generates a favorable impact, administrators might aim to scale up the program model. Others may try to replicate the findings. This often leaves program administrators, policymakers, and researchers asking, "What works?"

The first step is to assess overall impact—did the program improve outcomes for individuals above what they would have experienced in the absence of the program? As is well known, an experimental approach provides the strongest internal validity for estimating overall impact. Many job training initiatives in the United States estimate their overall impact using an experimental approach.[1]

A common critique of experimental evaluation is that it does not get inside the "black box" of what specifically is generating impacts. Key stakeholders, practitioners, and the broader field want to understand which component of a program's bundle of training and supports leads to impact. Answering such questions with comparable internal validity is often more difficult than the question of overall impact. This usually requires a separate experiment. Samples may be even smaller than for the overall evaluation, limiting statistical power. Obstacles such as cost and logistical burden of implementation may make an additional experiment infeasible. Instead, researchers often turn to nonexperimental methods to estimate the causal impact of programs or their components—whether or not they use an experiment to estimate a program's overall impact.

[1] Summaries of many of these studies are available at the Department of Labor's Clearinghouse for Labor Evaluation and Research (CLEAR) at https://clear.dol.gov/.

This paper compares experimental and nonexperimental estimates of the impact of the treatment on the treated (TOT) in a health care–focused job training program in the United States. The Health Profession Opportunity Grants (HPOG) program offered health care–sector training to TANF recipients and other low-income individuals across the United States. The HPOG program experimentally tested three program enhancements: facilitated peer support, noncash incentives, and emergency assistance.

The analysis treats the three enhancements as separate case studies. The narrative discusses the meaning of the TOT in each case and reports the incremental impact of the take-up of each program component as estimated using a variety of methods. The analysis also tests whether adding information on compliance with a randomly assigned status improves the performance of nonexperimental methods.

Comparing the performance of nonexperimental methods to an experimental benchmark situates this work in the "within-study comparison" (WSC) literature (Fraker and Maynard 1987; Lalonde 1986; Wong, Steiner, and Anglin 2018). The analysis applies methods from the WSC literature to draw conclusions about equality of the experimental and nonexperimental impact estimates.

To preview the results, in one case there is no difference between the TOT as estimated using experimental and nonexperimental methods. In the other two cases, despite having only moderate sample sizes, the difference between the experimental and nonexperimental estimates of the TOT is so large that equality of the estimates can be rejected. In all three cases adding information on compliers to the analysis does not change the impact estimate, but in two cases it moves the difference across the threshold of statistical significance. Evaluators of job training programs, their audiences, and their funders should be aware that these methods could lead to incorrect inference and poor policy.

The remainder of the paper proceeds as follows. The next section provides additional details on the HPOG program, estimating the TOT, and the WSC literature. The third section discusses the data, and the fourth section discusses methods in further detail. The fifth section reports results separately for each of the three enhancements, and the sixth and final section discusses the findings and offers concluding thoughts.

BACKGROUND AND RELATED LITERATURE

The HPOG Program

The HPOG program was authorized by Congress to offer training opportunities for disadvantaged adults while also fulfilling the growing demand for a skilled workforce in the health care sector. The Administration for Children and Families (ACF) awarded the first round of HPOG grants to 32 grantees in 2010.[2] The ACF-funded impact study of this first round of grantees, or the HPOG Impact Study, is an experimental evaluation of 23 of these grantees who operated 42 programs. ACF offered broad programmatic guidelines (e.g., each of the programs follows a career pathways framework), but each of the individual programs was distinct, with notable variation across programs in their particular bundle of training and services (Werner et al. 2018).

The experimental evaluation used a hybrid design, where 23 programs randomly assigned individuals to either treatment or control and 19 programs randomly assigned individuals to

[2] A subsequent round of grants was awarded in 2015. The first round of grants came to be known as HPOG 1.0 and the second round as HPOG 2.0. This paper focuses only on HPOG 1.0.

enhanced treatment, standard treatment, or control. Peck et al. (2018) report on the short-term impacts of access to HPOG programs by pooling these designs together and comparing those in the treatment group (either enhanced or standard) to those in the control group across all 42 programs. To summarize, the authors find that those offered access to HPOG programs make more educational progress, are more likely to be employed in a health care job, and have slightly larger earnings five quarters after random assignment.[3]

In addition to learning the overall impact of the bundle of training and services offered under HPOG, the evaluation used three-armed randomization to isolate the incremental impact of three specific program "enhancements": access to facilitated peer support, noncash incentives, and emergency assistance. Only programs that did not include these characteristics as part of their standard bundle of programming were eligible to participate in this part of the evaluation. Within these programs, the evaluation offered these program enhancements to a random subset of individuals assigned to the treatment group.[4]

Peck et al. (2018) report estimates of the impact of offering these three enhancements—that is, estimates of the intention to treat (ITT)—on engagement with training. These three particular HPOG enhancements were selected ex ante as program characteristics that were likely to improve HPOG's impact on engagement with training. However, no enhancement increased HPOG's impact more than the standard bundle of training and services offered at HPOG programs. The lack of a favorable impact for these program components was a surprising and important finding, both for the research community and for HPOG program staff.

[3] Peck et al. (2019) report continued educational progress as of three years after random assignment, but no detectable impact on earnings.

[4] Among the remaining grantees who did not offer the randomized "enhancements," there was natural variation in whether these characteristics were included in the standard programming.

Peck et al. (2018) use findings from the implementation study to propose hypotheses for why the enhancements, which were selected to improve educational progress, wound up causing more harm than good. Specifically, they posit that noncash incentives are less likely to be effective for a motivated population such as those participating in HPOG, emergency assistance may not have been available for participants when it would have been most effective, and lack of participation in peer support resulted in programs mandating that students participate. The authors argue that mandatory participation in peer support may have crowded out other useful training experiences for participants with limited time to devote to education and training.

Experimental TOT

The focus of the analysis is necessarily on the TOT because the nonexperimental methods available are only able to estimate the TOT. The primary distinction between the ITT and the TOT in an experimental evaluation is noncompliance—the existence of those in the treatment group who fail to "take up" the enhancement. This is often referred to as one "side" of noncompliance, with the existence of those in the control group who manage to gain access to the enhancement being the second "side" of noncompliance. The HPOG application has individuals randomly assigned to receive the enhancement who choose not to take it up; however, by design, those in the control group could not gain access to the enhancements through HPOG.[5] As a result, this application has only one-sided noncompliance.

To understand the comparison of the experimental and nonexperimental approaches, it is helpful to describe the experimental TOT further. Consider the four classifications of individuals as defined in Angrist, Imbens, and Rubin (1996) in the context of this analysis:

[5] This ignores the possibility that study participants (whether in the treatment or control group) could have accessed these enhancements through other providers in the community. As such, the results imply the impact of the enhancements *from HPOG* as opposed to the impact of the enhancements from any source.

- *Always-takers* are those who always take up the enhancement through HPOG, regardless of treatment status;
- *Never-takers* are those who never take up the enhancement through HPOG, regardless of treatment status;
- *Compliers* are those who take up the enhancement when randomly assigned to the enhancement and do not take up the enhancement when they are not randomly assigned to the enhancement; and
- *Defiers* are those who do not take up the enhancement when randomly assigned to the enhancement but do take it up when not randomly assigned to it.

Because the HPOG evaluation has only one-sided noncompliance, there cannot be always-takers or defiers—there is no one assigned to the control group who is able to gain access to the treatment.[6] This implies that the analysis sample consists of never-takers and compliers.

In an experimental evaluation with one-sided noncompliance, the TOT is the same as the local average treatment effect (LATE) or complier average causal effect (CACE) identified by using random assignment as an instrumental variable for take-up (Angrist and Pischke 2009). Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996) demonstrate the LATE to be the treatment effect among the compliers for a given instrumental variable. The authors show that with relatively simple regularity assumptions, instrumental variables returns an estimate of the LATE with strong internal validity.[7]

Nonexperimental TOT

The manner in which a nonexperimental approach estimates the TOT differs by approach. In the context of the HPOG evaluation, ignoring the randomly assigned access to the

[6] As Gill et al. (2016) argue, if there are always-takers in the sample then it would not be possible to separately identify always-takers from compliers.

[7] The "regularity assumptions" are not the focus of this work and are not discussed in detail here. The specific "regularity assumptions" from Angrist, Imbens, and Rubin (1996) that are nontrivial to this application are the stable unit treatment value assumption (SUTVA) and monotonicity. SUTVA implies that the potential outcomes for one individual are not related to the treatment status of other individuals. Monotonicity implies that there are no "defiers"—individuals who do the opposite of their assigned treatment status (in any case).

enhancements implies few nonexperimental approaches are available. The analysis tests three reasonable approaches an evaluator might use: 1) ordinary least squares, 2) inverse propensity weighting where weights are assigned to everyone who did not take up the enhancement, and 3) inverse propensity weighting where weights are only assigned to those who did not take up the enhancement and did not have access to it. Each approach estimates a counterfactual for the treated group and takes the difference between the two to estimate the TOT.

Applying nonexperimental methods to an experimental evaluation creates a unique opportunity. Given the structure described above, the analysis is able to create an environment where compliers are known (within the enhancement arm of the experiment), and a counterfactual for compliers can be estimated (within the standard treatment arm of the experiment).

These observations result in two research questions for the analysis. First, in an evaluation of program components with one-sided noncompliance, can nonexperimental methods replicate an experimental benchmark? And second, does using information about compliance in the treatment group improve the performance of the nonexperimental methods?

Within-Study Comparison

This paper tests the relative performance of nonexperimental methods to an experimental benchmark. As such, it is related to the WSC literature. Starting with the work of LaLonde (1986) and Fraker and Maynard (1987), a growing body of studies in this literature aims to replicate the findings of experimental impacts using nonexperimental analytic techniques (Cook, Shadish, and Wong 2008; Glazerman, Levy, and Myers 2003; Michalopoulos, Bloom, and Hill 2004; Wong and Steiner 2018; Wong, Steiner, and Anglin 2018).

More specifically, this work is related to the WSC findings for job training evaluations. As a brief summary, this literature finds that nonexperimental approaches to estimating the overall impact of job training programs generally fail to reproduce those estimated using experimental methods (see Wong, Steiner, and Anglin [2018] for a complete summary) (Cook, Shadish, and Wong 2008; Dehejia and Wahba 1999; Glazerman, Levy, and Myers 2003; Fraker and Maynard 1987; Heckman and Hotz 1989; Heckman et al. 1998; Hetck, Ichimura, and Todd 1997; Lalonde 1986; Michalopoulos, Bloom, and Hill 2004; Smith and Todd 2005; Wong, Steiner, and Anglin 2018). Some possible explanations for this failure include complex mechanisms for selection into treatment, weak available covariates, and weak nonexperimental designs. Findings from within-study comparisons for other types of interventions are not as uniformly disparate (Cook, Shadish, and Wong 2008; Wong, Steiner, and Anglin 2018).

This paper's methodological approach is a variant of the "synthetic dependent design" defined by Wong and Steiner (2018). By Wong and Steiner's definition, a synthetic dependent design creates the nonexperiment by deleting some portion of the experimental sample to generate a nonequivalent comparison group. For instance, Gleason, Resch, and Berk (2018) delete observations on either side of a cutoff to simulate a regression discontinuity design. This analysis does not delete any portion of the experimental sample, but instead "deletes" the existence of a strong instrumental variable (random assignment) to motivate the use of other nonexperimental methods.

DATA

This study uses two sources of data for analysis:

1) HPOG's administrative data system, called the Performance Reporting System (PRS), contains administrative data on baseline characteristics and program information such as services received; and

2) HPOG's short-term follow-up survey, initiated 15–18 months after random assignment, contains self-reported information on participation in and completion of training (whether through HPOG or some other program), labor market outcomes such as employment and earnings, and other life outcomes such as receipt of public assistance.

The PRS has variables identifying random assignment status, HPOG program, take-up of services, and baseline characteristics.[8] The short-term follow-up survey measures the primary outcome measure: educational progress. This outcome, which is the same measure used as a confirmatory outcome by Peck et al. (2018), reflects completion of training or current enrollment in training. It is the first step in HPOG's logic model and directly connects training to subsequent success in the labor market.

While Peck et al. (2018) address missing covariate data using multiple imputation and nonresponse weighting, this analysis includes only complete cases. Litwok et al. (2019) note that subsequent testing revealed no advantage of multiple imputation over simpler approaches such as dummy-variable imputation. Further, rates of item-level missing data are low for the variables in this analysis, generally less than 1 percent per item (Harvill et al. 2018). This analysis is able to reproduce ITT impacts nearly identical to those reported in Peck et al. (2018) without imputation.

[8] The study has the capability to identify take-up of enhancements with either administrative data from the PRS or survey data (the short-term follow-up survey explicitly asked about receipt of the three enhancements). The analysis uses administrative records because they eliminate the possibility of recall bias and also explicitly reflect enhancements offered through HPOG (a survey respondent may have received emergency assistance from some source other than HPOG and so would respond affirmatively when asked).

ANALYSIS METHODS

This paper asks whether the incremental TOT impacts of program components could have been estimated reliably using mainstream nonexperimental methods in the absence of the three-armed experiment. To answer these questions, the analysis focuses on the programs where enhancements were experimentally tested.[9] Within those programs, the analysis considers two alternative hypothetical states: one that makes use of the experiment to estimate the impact of the enhancements and a second that ignores the third arm of the experiment, treats variation in take-up as if it were naturally occurring, and tries to reproduce the findings using nonexperimental methods.

The remainder of this section describes the methods in more detail. Specifically, a two-stage least squares approach when using the experiment, and ordinary least squares and two inverse propensity weighting approaches when ignoring the experiment.

Using the Experiment

The experimental analysis estimates the TOT by two-stage least squares using random assignment to enhanced services as an instrument for take-up of enhancements. This implies estimating the following model by two-stage least squares:

$$y_{ig} = \beta_0 + \beta_1 E_{ig} + \varepsilon_{ig} \tag{1}$$

where:

y_{ig} = educational progress for individual *i* in grantee *g*;

β_0 = the average level of educational progress for those who do not take up the enhanced HPOG bundle of services;

[9] See Appendix Table A1 for a list of all programs where enhancements were experimentally tested.

β_1 = the impact of take-up of the enhanced HPOG bundle of services on educational progress;

E_{ig} = an indicator for take-up of enhancement *E* by individual i in grantee *g*; and

ε_{ig} = an idiosyncratic error term for individual i in grantee *g*.

These analyses are conducted and reported separately by enhancement. Standard errors are robust to heteroskedasticity.

Of note, Equation (1) deliberately excludes covariates. The intuition for this analysis and the resulting estimates are more clearly conveyed with a comparison of various unadjusted means. While adjustment improves precision, it also raises other methodological concerns that are beyond the scope of this paper (see, for instance, Freedman [2008] and Lin [2013]). Appendix Tables A4 and A5 repeat the analysis with baseline covariates, and the findings remain unchanged.

Ignoring the Experiment

The second set of approaches ignores the experimental test of enhancements and tries to emulate the behavior of a researcher who observes naturally occurring variation in take-up of various program components. This approach is nonstandard and creates a unique environment for analysis. On the one hand, such an approach gets closer to the ideal counterfactual—people in the comparison group who would certainly take up the enhancement if they had the opportunity but were not allowed to do so. However, the trade-off to such an approach is that the results come from an environment that does not fully reflect actual participant experiences.

Each of the three nonexperimental methods estimates a different counterfactual. A naïve approach is to assume the average outcome among everyone who did not take up the enhancement is a reasonable counterfactual for the compliers. This approach estimates model (1) using OLS without any adjustment to E_{ig}. Of course, this completely disregards the known endogeneity in compliance that is observed among those assigned to the enhancement. However, Abdulkadiroglu et al. (2011) argue that if observational estimates (calculated by OLS) are similar to experimental estimates, they are also likely to be informative for nonexperimental samples. OLS might also be useful to generate bounds on the true impact if one can hypothesize the likely direction of the bias.

Given compliers are known in the treatment group, another way to approach this problem with nonexperimental methods is to treat compliance as a nonrandomly assigned "treatment." Researchers often use matching or reweighting strategies to correct for nonrandom selection of individuals into a treatment condition. After applying such strategies, outcomes are typically assumed to be independent of treatment status conditional on observables. Put differently, the comparison group more closely resembles a counterfactual for the treatment group.

Researchers often reduce the dimensionality of this problem using a propensity score (Rosenbaum and Rubin 1983). The score is calculated using logistic regression (where treatment is expressed as a function of a slew of observable characteristics).[10] Estimating a weighted treatment effect with weights a function of the inverse of the propensity score (also known as inverse propensity weighting, or IPW) is a particularly attractive way to incorporate the propensity score in analysis (Busso, DiNardo, and McCrary 2014).

The analysis applies IPW to estimate the TOT in two separate ways. The first approach, labeled "IPW – Take-up," treats take-up of the enhancement as a nonrandomly assigned treatment and estimates a propensity score using the entire analytic sample. These propensity

[10] The baseline observable characteristics include age, sex, presence of dependent children, race/ethnicity, indicator for born outside the United States, educational attainment, receipt of welfare, receipt of WIC/SNAP, work expectations, and number of reported barriers to employment/education.

scores are transformed into weights as follows: those who are observed to take up the enhancement all receive a weight of one; those who are not observed to take up the enhancement receive a weight of $\frac{\widehat{p_1}}{1-\widehat{p_1}}$, where $\widehat{p_1}$ is the estimated propensity score (Stuart 2010). This approach is intended to mimic the behavior of an evaluator who wants to estimate the impact of the enhancement without an experiment—so all they can observe is whether individuals take-up the enhancement or not.

The second approach, labeled "IPW – Compliers," treats take-up of the enhancement as a nonrandomly assigned treatment *within the third experimental arm*. It uses only those who were randomly assigned to the enhancement to estimate a model relating take-up to baseline characteristics, and then uses that model to predict a propensity score for the entire analytic sample (including those in the standard HPOG experimental arm). In this case the propensity scores are transformed into weights as follows: those who are observed to take up the enhancement all receive a weight of one, those randomly assigned to the enhanced group who do not take up the enhancement receive a weight of zero, and those who are assigned to the standard HPOG arm receive a weight of $\frac{\widehat{p_2}}{1-\widehat{p_2}}$, where $\widehat{p_2}$ is the estimated propensity score. This approach incorporates information about compliance into the nonexperimental analysis, and testing whether this estimate differs from the first IPW approach implicitly tests whether adding this information improves the reliability of the resulting estimate.

Comparing the Approaches

Ultimately, this study aims to understand whether the impacts that ignore the experiment are "close enough" to those that use the experimental data. Comparing the magnitudes of the impact estimates offers a simple eyeball test. Are the estimates the same sign? Are they the same magnitude? Would they lead to the same conclusion? While answering these questions is one way to compare the estimates, statistical tests are the appropriate way to incorporate all pertinent uncertainty into the comparison.

The complication with statistical testing is that the impact estimates do not come from independent samples. Resampling methods solve this problem by approximating the variance of the difference between the experimental TOT estimate and each of the nonexperimental TOT estimates (Steiner and Wong 2018). The analysis uses 1,000 bootstraps and repeats all of the estimation procedures within each bootstrapped sample. The resulting distribution of differences allows for calculation of a standard error.

RESULTS

Analysis Diagnostics

The focus of the experimental analysis—the TOT impact estimated via two-stage least squares—will be a function of the ITT impact and take-up of the enhancements (Bloom 1984). Table 1 reports take-up rates for each of the three enhancements—that is, the fraction of those assigned to the enhanced treatment arm who actually received the enhancement. Less than half of those individuals randomly assigned to enhancements took up emergency assistance and noncash incentives, which is much lower than the 86 percent take-up rate for those assigned to facilitated peer support. Although rates of take-up vary by enhancement, random assignment is a very strong instrument for take-up of each of the enhancements.[11]

[11] Appendix Table A2 reports first-stage results for this analysis.

Table 1 Sample Sizes and Take-Up Rates by Enhancement

	Facilitated peer support	Noncash incentives	Emergency assistance
Total N	833	1,194	1,465
N assigned to enhancement	250	269	437
N who took up enhancement	214	126	179
Take-up rate (%)	85.6	46.8	41.0

SOURCE*:* PRS.

Causal interpretation of the nonexperimental approaches relies on a conditional independence assumption. That is, conditional on observable baseline characteristics, one needs to be willing to assume that assignment to the treatment is independent of outcomes. This assumption is inherently untestable, but researchers support this assumption in two ways: 1) demonstrating balance of observable characteristics prior to the intervention,[12] and 2) arguing that there is minimal potential for bias due to unobservable characteristics. Clearinghouse standards, such as those used by the Department of Labor's Clearinghouse for Labor Evaluation and Research (CLEAR) often specify the particular covariates for which baseline balance must be demonstrated. In the case of employment and training studies, CLEAR guidelines require balance to be demonstrated on age, race, gender, and a preintervention measure of the outcome of interest (CLEAR 2019).

Tables 2–4 report standardized difference in baseline characteristics between the treatment and comparison groups. The tables report raw and weighted differences with both sets of weights for the IPW analyses.[13] They also differentiate between the covariates identified by the CLEAR guidelines and other baseline characteristics.

[12] If balance cannot be achieved, the solution is often to include the characteristic as a covariate in the analysis. As noted in the fourth section, the results in the main body of the paper do not include covariates (despite the fact that balance is weak for some covariates). Appendix Tables A4 and A5 repeat the main analyses in the paper with covariates, and the findings do not change.

[13] The raw differences for the "Take-up" group are the relevant differences for the ordinary least squares analysis.

Table 2 Balance of Observable Characteristics Before and After Weighting: Facilitated Peer Support

	Take-up		Compliers	
Characteristic	Raw	Weighted	Raw	Weighted
Age	0.056	−0.031	0.051	−0.114
Hispanic	0.113	0.058	0.105	0.138
Black	−0.092	0.000	−0.080	0.236
Female	0.022	0.028	−0.005	−0.133
Degree	0.034	0.067	0.030	0.138
License	−0.072	−0.045	−0.084	−0.202
Attend adult basic education	−0.073	0.035	−0.075	−0.170
Attend classes to succeed in school	−0.057	−0.025	−0.057	−0.058
Attend vocational/technical school	−.048	−.084	−.048	−.123
Attend classes to succeed at work	−0.005	−0.014	−0.010	−0.280
Dependent children	0.028	−0.014	−0.004	−0.267
Born outside U.S.	0.171	00.041	0.148	0.018
Welfare	0.096	0.033	0.070	−0.346
WIC/SNAP	−0.163	−0.059	−0.175	−0.070
Expect to be working	−0.061	−0.019	−0.025	0.526
Number of barriers	−0.016	−0.035	−0.029	0.040

NOTE: Table reports standardized effect size differences between groups as defined in the first two rows.
SOURCE: PRS.

Table 3 Balance of Observable Characteristics Before and After Weighting: Noncash Incentives

	Take-up		Compliers	
Characteristic	Raw	Weighted	Raw	Weighted
Age	0.189	0.023	0.142	−0.273
Hispanic	0.416	0.035	0.379	−0.136
Black	−0.602	−0.096	−0.530	0.123
Female	−0.098	0.049	−0.113	0.101
Degree	0.028	−0.015	0.006	−0.146
License	−0.131	−0.058	−0.119	−0.017
Attend adult basic education	0.086	0.004	0.079	−0.011
Attend classes to succeed in school	0.182	−0.022	0.194	0.040
Attend vocational/technical school	−0.060	−0.032	−0.028	0.193
Attend classes to succeed at work	0.184	0.014	0.152	−0.204
Dependent children	0.019	−0.037	0.037	−0.073
Born outside U.S.	−0.054	−0.021	−0.036	0.094
Welfare	−0.062	0.041	−0.065	0.072
WIC/SNAP	−0.125	−0.066	−0.078	0.003
Expect to be working	−0.198	−0.090	−0.141	0.250
Number of barriers	0.008	−0.058	−0.008	−0.160

NOTE: Table reports standardized effect size differences between groups as defined in the first two rows.
SOURCE: PRS.

Table 4 Balance of Observable Characteristics Before and After Weighting: Emergency Assistance

	Take-up		Compliers	
Characteristic	Raw	Weighted	Raw	Weighted
Age	−0.061	−0.021	−0.074	−0.235
Hispanic	0.131	0.047	0.141	0.122
Black	0.245	−0.016	0.178	−0.246
Female	−0.056	0.019	−0.071	−0.130
Degree	0.022	0.024	0.022	−0.045
License	0.089	0.040	0.088	0.048
Attend adult basic education	0.163	0.064	0.154	0.015
Attend classes to succeed in school	0.034	0.023	0.035	−0.030
Attend vocational/technical school	−0.086	−0.045	−0.094	−0.165
Attend classes to succeed at work	−0.018	−0.008	−0.027	−0.083
Dependent children	0.250	0.012	0.241	−0.017
Born outside U.S.	0.038	0.028	0.044	−0.084
Welfare	0.267	−0.032	0.258	0.011
WIC/SNAP	0.271	0.024	0.235	−0.087
Expect to be working	0.096	0.047	0.152	0.211
Number of barriers	0.050	−0.025	0.080	0.059

NOTE: Table reports standardized effect size differences between groups as defined in the first two rows.
SOURCE: PRS.

The results of Tables 2–4 can be summarized as follows. The raw differences show imbalances (sometimes substantial imbalances) for baseline characteristics. The reweighting procedure that uses all participants who did not take up the enhancement shrinks these differences to all be smaller than 0.1 standard deviations. However, in the IPW analysis that focuses on identifying compliers, the reweighting procedure exacerbates some of the differences to be quite large—in some cases magnitudes that are larger than 0.25 standard deviations. These differences have implications for interpreting the findings below.[14]

Case 1: Facilitated Peer Support

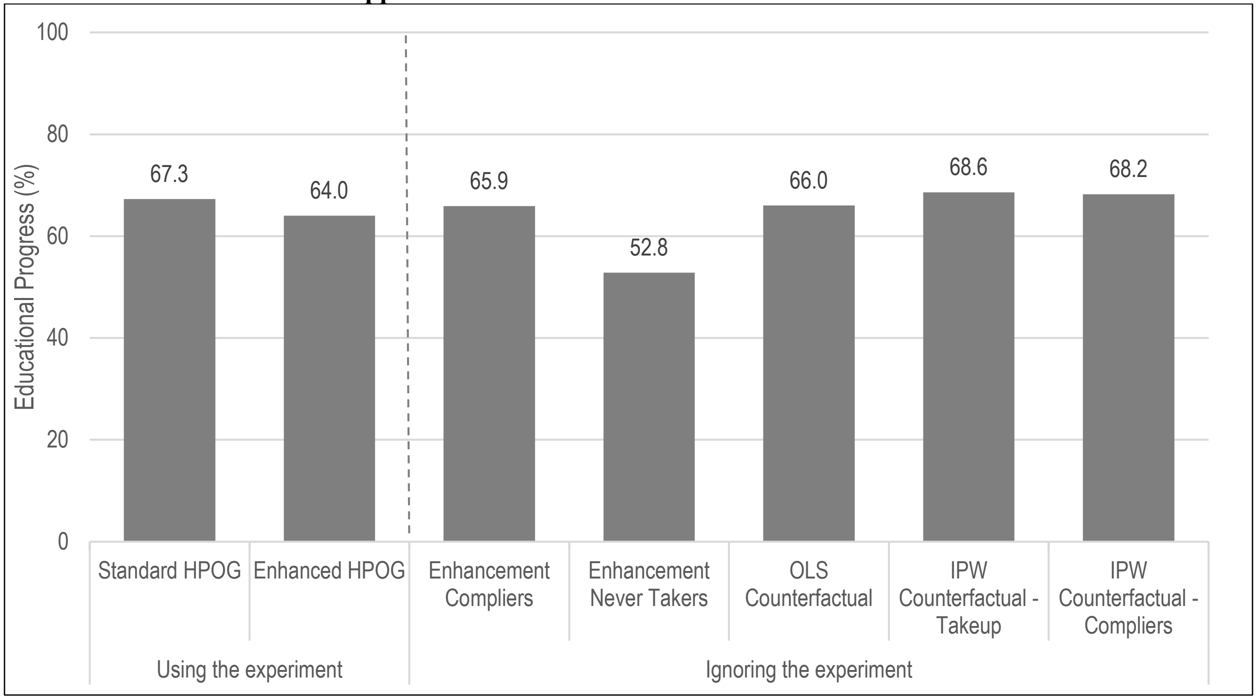
As noted by Peck et al. (2018), facilitated peer support was designed to develop meaningful connections between students, faculty, and staff, with the hope that these connections would translate to improved program outcomes. The structure of the support varied across the programs that tested the enhancement, but a common theme across the programs was challenges

[14] Another standard diagnostic in the literature on matching is to test for overlap of the estimated propensity scores. Visual inspection of the propensity scores show strong overlap in all cases (not reported).

to attendance. In some cases, programs responded by making the peer support mandatory, which likely explains the relatively higher take-up rate for this enhancement.

The program transition from voluntary participation in facilitated peer support to a requirement implies program administrators expected a meaningful TOT impact. Administrators making such a decision would need to have an estimate of TOT impact. Figure 1 reports means for various treatment and control/comparison groups that might be used to calculate that TOT—either by using the experiment or in the absence of an experiment.

Figure 1 Average Educational Progress for Various Treatment and Comparison Groups Defined by Receipt of Facilitated Peer Support



NOTE: N = 605 individuals randomly assigned to either standard HPOG or enhanced HPOG at programs that experimentally tested facilitated peer support. See Appendix Table A3 for standard errors and results of statistical tests for differences in means.
SOURCE: PRS and short-term follow-up survey.

An attractive feature of using the experiment is the simplicity with which the TOT can be calculated. The means in Figure 1 can be used to calculate the ITT impact—64.0 minus 67.3 implies an ITT impact of negative 3.3 percentage points. To help with interpretation of that impact estimate, Peck et al. (2018) report that about 60 percent of the control group made educational progress as of the short-term follow-up survey. This implies that an incremental impact of 3 percentage points can be roughly interpreted as a 5 percent effect relative to the entire control group for the study. To convert this to the TOT, one simply scales the ITT estimate by the take-up rate (see Table 1)—negative 3.3 percentage points divided by 0.856 implies a TOT of negative 3.9 percentage points. Although not reported in Figure 1, the standard error associated with this estimate implies it is not statistically different from zero.[15]

Moving to the right side of Figure 1 implies ignoring the experiment—while the experiment itself induced variation in take-up of the enhancement, the right side of Figure 1 behaves as if that variation is naturally occurring. The first two columns on the right side of Figure 1 show the decomposition of those in the "Enhanced HPOG" group: the 85.6 percent who took up the enhancement made educational progress at a rate of 65.9 percent; and the 14.4 percent who did not take up the enhancement made educational progress at a rate of 52.8 percent. The weighted average of those two is the 64.0 percent rate of educational progress reported for the "Enhanced HPOG" group on the left side of Figure 1.

The nonexperimental approaches aim to estimate the TOT by generating a plausible counterfactual for the rate of educational progress among the compliers. The bar labeled "OLS Counterfactual" reports the average rate of educational progress for everyone in the sample who did not take up the enhancement: those assigned to the "Standard HPOG" group and those assigned to the "Enhanced HPOG" group who did not take it up. That is, the rate of 66.0 is a weighted average of the 67.3 percent in the "Standard HPOG" group and the 52.8 percent in the

[15] See Appendix Table A3 for all estimates and standard errors reported in Table 5.

"Enhancement Never Takers" group. The resulting estimate from OLS is 65.9 minus 66.0, or negative .1 percentage points.

The last two bars in Figure 1 report the counterfactual using the two different IPW procedures. The first of the two bars is based on the propensity score that is calculated for the entire analytic sample—without incorporating any information about who is a complier. It is a different weighted average of "Standard HPOG" and "Enhancement Never Takers" where the weights are now a function of baseline characteristics (such that those with baseline characteristics that are more similar to the compliers have larger weights). The impact estimate implied by this counterfactual is 65.9 minus 68.6, or negative 2.7 percentage points. The IPW procedure moves the impact estimate closer to the experimental estimate.

The last bar in Figure 1 estimates the counterfactual by reweighting the "Standard HPOG" group alone to be more representative of the "Enhancement Compliers" group (in terms of baseline characteristics). The impact estimate implied with this counterfactual is 65.9 minus 68.2, or negative 2.3 percentage points. In this case limiting the comparison to those randomly assigned to the standard enhancement arm moves the impact estimate further from the experimental benchmark—perhaps because this analysis *increases* baseline imbalance between the two groups. It is worth noting, however, that none of the estimated impacts are different from zero.

Of greater interest than the statistical significance of the individual impact estimates is whether the impact estimates using the various procedures differ from each other. As noted in Section 4, those tests are conducted using resampling methods to appropriately estimate the variance. Table 5 reports the findings.

Table 5 Differences in TOT Impacts on Educational Progress for Facilitated Peer Support Using Experimental and Nonexperimental Variation

2SLS Compared to…	Difference in impacts (standard error)
OLS	−3.8
	(2.4)
IPW – Take-up	−1.2
	(3.0)
IPW – Compliers	−1.6
	(11.7)

NOTE: Standard errors from 1,000 bootstraps. *p < 0.10; **p < 0.05; ***p < 0.01.
SOURCE: PRS and short-term follow-up survey.

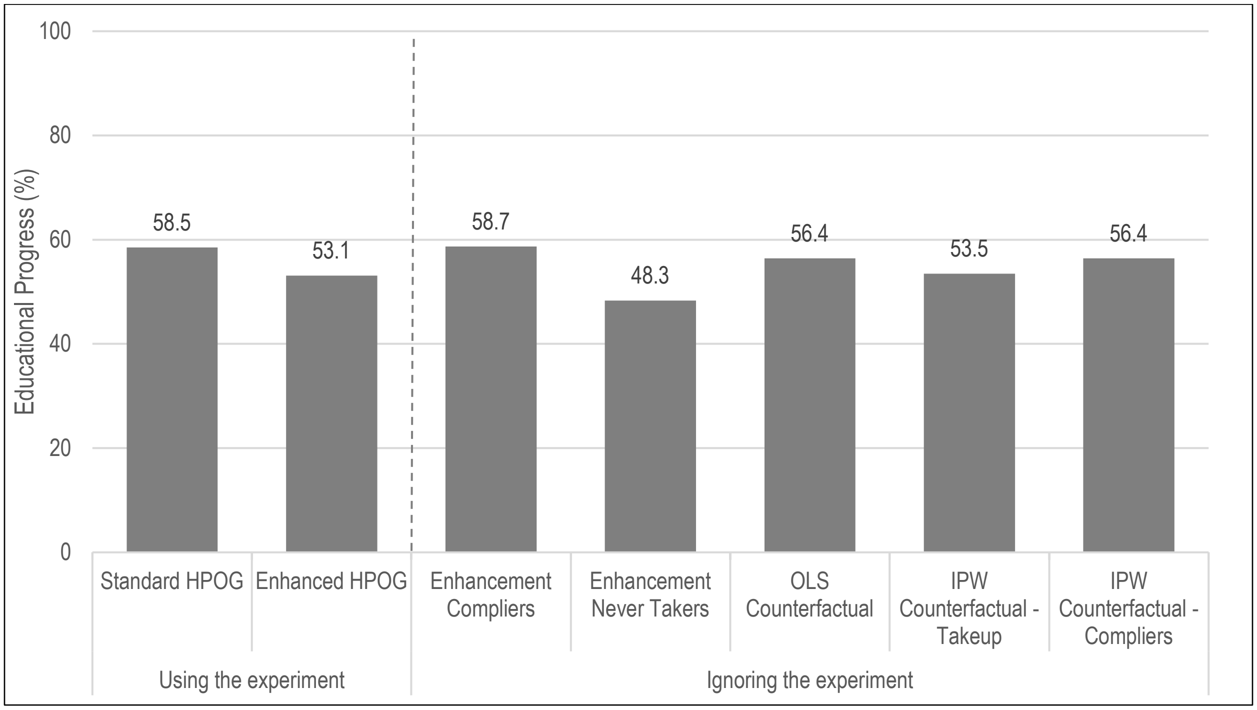
The results in Table 5 imply that none of the impact estimates differ from the experimental estimate in a statistical sense. In addition to tests for differences in the impact estimates, Steiner and Wong (2018) recommend WSCs report the results of a test for within-study comparison correspondence. Following Steiner and Wong, the test for correspondence combines two one-sided tests for whether the difference in the estimates is greater than 0.1 standard deviations of the outcome measure and less than the opposite of that threshold. The analysis of the facilitated peer support enhancement fails to reject the null of equivalence. As a result, in the terminology of Steiner and Wong, the analysis concludes that the differences between the experimental and nonexperimental estimates for facilitated peer support are indeterminate.

Case 2: Noncash Incentives

Noncash incentives allowed participants to earn points for achieving program milestones and convert those points to tangible rewards. Whether participants chose to take up the noncash incentive enhancement was a function of many factors, including the desirability of the incentives and implementation of the enhancement (Peck et al. 2018). The argument for the TOT impact is less compelling in this case than in the case of facilitated peer support—it is harder to imagine a scenario where a researcher is interested in only the impact of noncash incentives among those who received them as opposed to the impact of the offer of noncash incentives. One possibility is if the program is attempting to refine the incentives it is offering and so asks about the impact of particular noncash incentives among those who received them.

As in Case 1, Figure 2 reports means for various treatment and control/comparison groups that a researcher can use to calculate the TOT either by using the experiment or in the absence of an experiment.

Figure 2 Average Educational Progress for Various Treatment and Comparison Groups Defined by Receipt of Noncash Incentives



NOTE: N = 1,026 individuals randomly assigned to either standard HPOG or enhanced HPOG at programs that experimentally tested noncash incentives. See Appendix Table A3 for standard errors and results of statistical tests for differences in means.
SOURCE: PRS and short-term follow-up survey.

The means in Figure 2 imply an ITT impact of negative 5.3 percentage points. Using the take-up rate from Table 1 implies a TOT of negative 11.3 percentage points. Despite the large magnitude of this estimate, this estimate is also not statistically different from zero.[16]

[16] See Appendix Table A3 for all estimates and standard errors reported in Figure 2.

The first two columns on the right side of Figure 2 show that the compliers with noncash incentives made educational progress at a rate similar to the standard HPOG group (58.7 percent). Those who did not take up noncash incentives fared much more poorly, with only 48.3 percent making educational progress. The "OLS Counterfactual," which is a weighted average of "Standard HPOG" and the "Enhancement Never Takers" is 56.4 percent educational progress. This counterfactual implies a TOT impact of *positive* 2.3 percentage points, which is substantially different from the TOT impact estimated using the experiment.

The last two bars in Figure 2 aim to improve the counterfactual using an IPW approach. The counterfactual estimate based on a propensity score that is calculated for the entire analytic sample is a rate of 53.5 percent making educational progress, which implies a TOT impact estimate of 5.2 percentage points. That is, the IPW procedure has moved the impact estimate further away from the experimental estimate.

The counterfactual implied by the IPW procedure that makes use of information on compliers is 56.4 percentage points. While this moves the TOT impact in the right direction (the estimate moves back to 2.3 percentage points), the TOT estimate using a nonexperimental approach remains substantially different from the approach that used the experiment. In all three cases the nonexperimental estimate of the TOT impact is not statistically different from zero.

As in Case 1, however, the question of interest is whether the impact estimates using the various procedures differ from each other. Table 6 reports the results of those tests.

Table 6 Differences in TOT Impacts on Educational Progress for Noncash Incentives Using Experimental and Nonexperimental Variation

2SLS Compared to…	Difference in impacts (standard error)
OLS	−13.7** (6.5)
IPW – Take-up	−16.5** (6.9)
IPW – Compliers	−13.7 (8.3)

NOTE: Standard errors from 1,000 bootstraps. *p < 0.10; **p < 0.05; ***p < 0.01.
SOURCE: PRS and short-term follow-up survey.

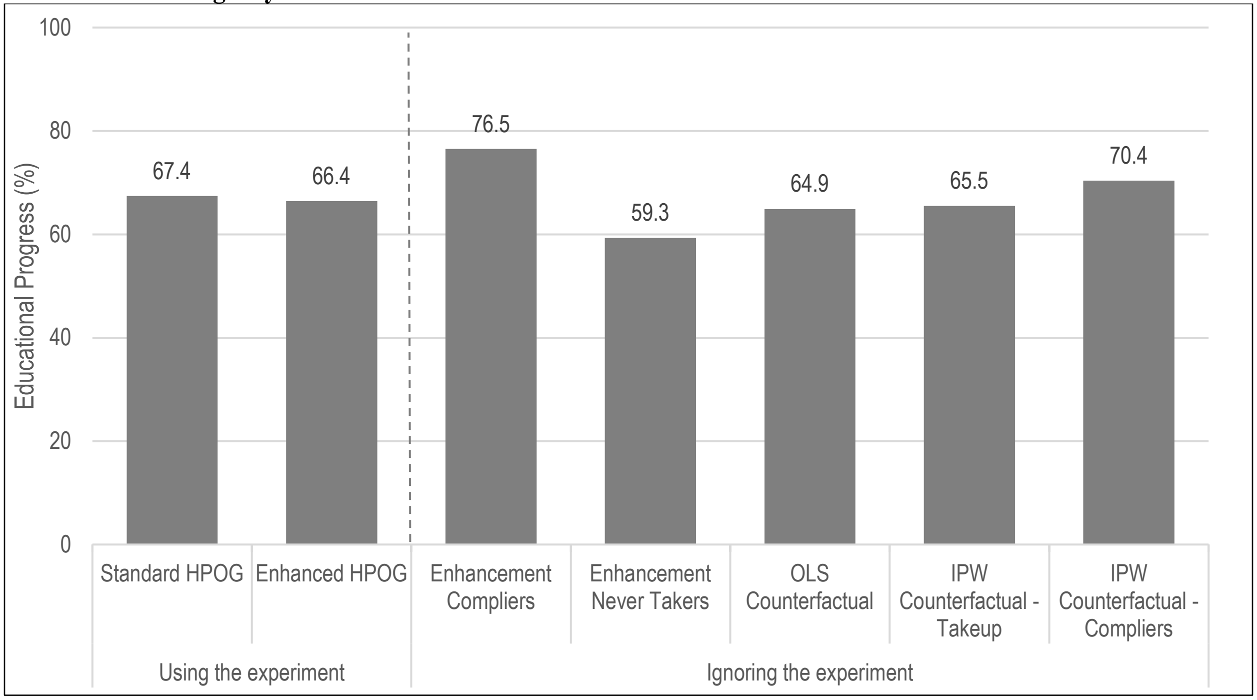
The results in Table 6 imply that two of the three nonexperimental impact estimates differ from the experimental estimate in a statistical sense. That is, if a researcher used OLS or IPW–Take-up to estimate the TOT, the resulting estimate contains substantial bias. In fact, despite the relatively small sample size, the disparity in impacts are so large that the Steiner and Wong (2018) test concludes that the estimates are different. While incorporating the information on compliers is not statistically different from the experimental benchmark, the lack of statistical significance appears to be due to the increase in variance that comes from weighting the analysis. The nonexperimental estimates are not statistically different from each other. In summary, all three nonexperimental approaches perform quite poorly for noncash incentives.

Case 3: Emergency Assistance

Emergency assistance provided financial support to HPOG participants who experienced sudden financial needs that threatened their ability to continue in the program. In essence, this enhancement works as an insurance policy for program participants. Peck et al. (2018) describe the various needs the emergency assistance was intended to cover, such child care, transportation, or utilities. Peck et al. (2018) also describe problems with implementation of the enhancement that may explain its lack of impact. From the perspective of this paper, emergency assistance is the weakest enhancement for this exercise because the impact of interest is very clearly the ITT. The TOT estimates the impact among those who took up emergency assistance, but the coverage of this insurance policy could have had an impact in its own right. It is hard to imagine that a program administrator or case manager would want to know the impact of take-up of emergency assistance. Despite the theoretical shortcomings of this concept of impact, this section estimates the TOT of emergency assistance to provide a complete picture of the TOT for all three HPOG enhancements.

Figure 3 reports the means needed to calculate the TOT for emergency assistance either by using the experiment or in the absence of an experiment. The ITT impact in Figure 3 is negative and very small in magnitude—only 1.0 percentage point. Scaling this by the take-up rate from Table 1 implies a TOT of negative 2.5 percentage points, which is not statistically different from zero.[17]

Figure 3 Average Educational Progress for Various Treatment and Comparison Groups Defined by Receipt of Emergency Assistance



NOTE: N = 818 individuals randomly assigned to either standard HPOG or enhanced HPOG at programs that experimentally tested emergency assistance. See Appendix Table A3 for standard errors and results of statistical tests for differences in means.
SOURCE: PRS and short-term follow-up survey.

[17] See Appendix Table A3 for all estimates and standard errors reported in Figure 3.

In this case the nonexperimental estimate produces very different results. The compliers with emergency assistance made substantial educational progress (76.5 percent). This finding is particularly noteworthy because the need for emergency assistance is triggered by a negative shock; but those who took up the emergency assistance fared better than even the standard HPOG group. The average rate of educational progress among those who did not take up emergency assistance was 59.3 percent.

The OLS counterfactual implies a rate of educational progress just under 65 percent. Taken at face value, the OLS estimate of impact would lead to the conclusion that impacts were *larger* for those who took up emergency assistance (a sizeable impact of 11.6 percentage points, which is statistically different from zero); when in fact the experiment returned an impact estimate indicating impacts were smaller, though not significantly different from zero.

In the case of emergency assistance, the IPW approaches have mixed performance. Using the counterfactual based on a propensity score that is calculated for the entire analytic sample, the impact estimate is barely changed to 11.0 percentage points and remains statistically different from zero. The counterfactual implied by the IPW procedure that makes use of information on compliers moves the impact to 6.1 percentage points, an estimate that is no longer statistically different from zero and has moved closer to the experimental benchmark. However, while using information on compliers moves the TOT impact in the right direction, the TOT estimate using a nonexperimental approach remains substantially different from the approach that used the experiment.

Once again, Table 7 reports the results of statistical tests for differences across the impact estimates.

Table 7 Differences in TOT Impacts on Educational Progress for Emergency Assistance Using Experimental and Nonexperimental Variation

2SLS Compared to…	Difference in impacts (standard error)
OLS	−14.1**
	(6.3)
IPW – Take-up	−13.6**
	(6.4)
IPW – Compliers	−8.7
	(6.1)

NOTE: Standard errors from 1,000 bootstraps. *p < 0.10; **p < 0.05; *** p<0.01.
SOURCE: PRS and short-term follow-up survey.

The results in Table 7 are similar to the results in Table 6 in terms of statistical significance—the OLS and IPW Take-up impact estimates are statistically different from the experimental approach, while the IPW Compliers estimate is not different from the experimental benchmark. The first two differences between the impacts are also large enough for the Steiner and Wong (2018) test to conclude that the estimates do indeed differ from each other. In this case adding information on compliers moved the impact estimate in the right direction and resulted in a difference that was not statistically different from the experiment; but again, the nonexperimental impact estimates are not statistically different from each other.

DISCUSSION AND CONCLUSION

This paper presents three separate cases of attempting to reproduce experimental TOT results using nonexperimental methods in an experimental evaluation setting with one-sided noncompliance. The analysis asks two questions: 1) Do the nonexperimental approaches reproduce the experimental benchmark? 2) Does adding information about compliance with random assignment improve the performance of the nonexperimental analysis?

In terms of the first research question, the results of the exercise varied across the three cases. In one of the cases—facilitated peer support—the estimates using nonexperimental methods were not different from the experimental estimates in a statistical sense. In the other two cases the estimates using nonexperimental methods differed substantially from the experimental estimates—in both a statistical and a practical sense. For instance, in the case of emergency assistance the nonexperimental approach gave the incorrect impression that take-up of the enhancement had a favorable impact. Such a conclusion could ultimately lead to poor policy.

The second question asks about the relative contribution of information on compliers, which is only available because of the experimental approach. While adding this information moved impact estimates across the threshold of statistical significance in two cases, the estimates themselves were not statistically different from the other nonexperimental estimates.

This paper also discusses the distinction between the ITT and TOT for all three cases. The three cases each have their own complexities and nuance, and the arguments lay out the conditions where one might be interested in an ITT analysis or a TOT analysis. Studies that aim to estimate the TOT should make similar arguments for why this is the impact of policy interest.

As noted in the methods section, the analysis throughout the main body of this paper excludes covariates. Appendix Tables A4 and A5 reproduce the main findings of the paper with the addition of covariates. The primary findings of the study remain unchanged when regression analyses include covariates.

The general failure of nonexperimental methods to reproduce the experimental estimate is consistent with prior findings on within-study comparison in evaluations of job training programs (Cook, Shadish, and Wong 2008; Michalopoulos, Bloom, and Hill 2004). This particular application is similar to examples presented in those prior works where only a weak set of covariates is available or the details of the selection process are unknown. Those prior works also found that nonexperimental methods performed poorly in estimating the overall program effect relative to an experimental benchmark. This study finds that the same result holds for estimating the incremental impact of a program component. Had a stronger nonexperimental approach been possible, such as a cutoff that lent itself to a regression discontinuity design, the performance of the nonexperimental approach might have improved.

Nonexperimental methods performed best in the case of facilitated peer support. It is no coincidence that take-up rates were substantially higher for peer support than for the other two enhancements (see Table 1). This observation suggests that the performance of nonexperimental methods might improve with stronger compliance.

Relatedly, the fundamentally untestable selection on observables argument is key to the IPW analyses. It is possible that key variables for predicting compliance vary by enhancement and may have been omitted from the propensity score calculation. For instance, it could be the case that conditioning on prior engagement with educational classes and receipt of a degree or credential minimizes bias due to unobserved characteristics in selection into facilitated peer support. At the same time, it could be the case that take-up of emergency assistance and noncash incentives is much more idiosyncratic, and therefore subject to more significant threats due to unobserved characteristics.

To summarize, all three of the enhancements have no detectable impact on educational progress for those who complied with random assignment to enhancements (although the point estimates are all negative). More importantly, even given information on compliance with random assignment, nonexperimental methods are not a good replacement for identifying the incremental impact of the enhancement among compliers for two of the three enhancements. The bias in these estimates is so large that there is strong statistical evidence of differences in the estimates, despite the relatively small samples. As a result, practitioners, researchers, policymakers, and research sponsors should be wary of using these approaches for informing policy and practice.

References

Abdulkadiroglu, Atila, Joshua D. Angrist, Susan M. Dynarski, Thomas J. Kane, and Parag A. Pathak. 2011.Accountability and flexibility in public schools: Evidence from Boston's charters and pilots. *Quarterly Journal of Economics* 126(2): 699–748.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91(434): 444–455.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Bloom, Howard S. 1984. "Accounting for No-Shows in Experimental Evaluation Designs." *Evaluation Review* 8(2): 225–246.

Busso, Matias, John DiNardo, and Justin McCrary. 2014. "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators." *Review of Economics and Statistics* 96(5): 885–897.

Clearinghouse for Labor Evaluation and Research. 2019. Review Protocol for the Employment and Training Topic Area. Version 2.0. https://clear.dol.gov/reference-documents/employment-and-training-topic-area-review-protocol.

Cook, Thomas D., William R. Shadish, and Vivian C. Wong. 2008. "Three Conditions under Which Experiments and Observational Studies Often Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons." *Journal of Policy Analysis and Management* 27(4): 724–750.

Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94(448): 1053–1062.

Fraker, Thomas, and Rebecca Maynard. 1987. "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs." *Journal of Human Resources* 22(2): 194–227.

Freedman, David A. 2008. "On Regression Adjustments to Experimental Data." *Advances in Applied Mathematics* 40(2): 180–193.

Gill, Brian, Joshua Furgeson, Hanley Chiang, Bing-ru Teh, Joshus Haimson, and Natalya Verbitsky Savitz. 2016. "Replicating Experimental Impact Estimates with Nonexperimental Methods in the Context of Control-Group Noncompliance." *Statistics and Public Policy* 3(1): 1–11.

Glazerman, Steven, Dan M. Levy, and David Myers. 2003. "Nonexperimental versus Experimental Estimates of Earnings Impacts." *The Annals of the American Academy of Political and Social Science* 589(1): 63–93.

Gleason, Philip, Alexandra Resch, and Jillian Berk. 2018. "RD or Not RD: Using Experimental Studies to Assess the Performance of the Regression Discontinuity Approach." *Evaluation Review* 42(1): 3–33.

Harvill, Eleanor, Daniel Litwok, Shawn Moulton, Alyssa Rulf Fountain, and Laura R. Peck. 2018. *Technical Supplement to the Health Profession Opportunity Grants (HPOG) Impact Study Interim Report: Report Appendices*. OPRE Report No. 2018-16b. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Heckman, James J., and V. Joesph Hotz. 1989. "Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association* 84(408): 862–874.

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66(5): 1017–1098.

Heckman, James J., Hidehiko Ichimura, and Petra Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies* 64(4): 605–654.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2): 467–475.

LaLonde, Robert J. 1986. Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *American Economic Review* 76(4): 604–620.

Lin, Winston. 2013. "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique." *The Annals of Applied Statistics* 7(1): 295–318.

Litwok, Daniel, D. Walton, Laura R. Peck, and Eleanor Harvill. 2019. *Health Profession Opportunity Grants (HPOG) Impact Study's Three-Year Follow-Up Analysis Plan*. OPRE Report No. 2018-124, Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Michalopoulos, Charles, Howard S. Bloom, and Carolyn J. Hill. 2004. "Can Propensity-Score Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?" *Review of Economics and Statistics* 86(1): 156–179.

Peck, Laura R., Daniel Litwok, D. Walton, Eleanor Harvill, and Alan Werner. 2019. *Health Profession Opportunity Grants (HPOG 1.0) Impact Study: Three-Year Impacts Report*.

OPRE Report 2019-144. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Peck, Laura R., Alan Werner, Eleanor Harvill, Daniel Litwok, S. Moulton, A. Rulf Fountain, and G. Locke. 2018. *Health Profession Opportunity Grants (HPOG 1.0) Impact Study Interim Report: Program Implementation and Short-Term Impacts*. OPRE Report 2018-16a. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Smith, Jeffrey A., and Petra E. Todd. 2005. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125(1-2): 305–353.

Steiner, Peter M., and Vivian C. Wong. 2018. "Assessing Correspondence between Experimental and Nonexperimental Estimates in Within-Study Comparisons." *Evaluation Review* 42(2): 214–247.

Stuart, Elizabeth A. 2010. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* 25(1): 1.

Werner, Alan, Pamela Loprest, Schwartz, Deena, Robin Koralek, and Nathan Sick. 2018. *National Implementation Evaluation of the First Round Health Profession Opportunity Grants (HPOG 1.0)*. OPRE Report 2018-09. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Wong, Vivian C., and Peter M. Steiner. 2018. "Designs of Empirical Evaluations of Nonexperimental Methods in Field Settings." *Evaluation Review* 42(2): 176–213.

Wong, Vivian C., Peter M. Steiner, and Kylie L. Anglin. 2018. "What Can Be Learned from EmpiricalEvaluations of Nonexperimental Methods?" *Evaluation Review* 42(2): 147–175.

APPENDIX A

Appendix Table A1 summarizes this information across all 42 programs in the HPOG Impact Study. For each of the three enhancements, the table identifies whether the enhancement was “offered,” meaning the enhancement was already part of the standard bundle of services; “tested,” meaning the enhancement was tested experimentally using the three-armed design; or neither, meaning the enhancement was not part of the HPOG bundle at all. The breakdown by programs is as follows: 19 programs offered and 11 programs tested emergency assistance, 4 programs offered and 5 programs tested noncash incentives, and 2 programs offered and 3 programs tested facilitated peer support.

Table A1 HPOG Enhancement Sites

State	Grantee—program operator	Emergency assistance	Noncash incentives	Facilitated peer support
AZ	Pima County Community College District			
CA	San Diego Workforce Partnership - MAAC South	OFFER		
CA	San Diego Workforce Partnership - Metro CTS	OFFER		
CA	San Diego Workforce Partnership - North County Lifeline	OFFER		
CT	The WorkPlace	OFFER		TEST
FL	Pensacola State College	OFFER		
IL	Will County WIB - Central States SER	OFFER		
IL	Will County WIB - College of Lake	OFFER		
IL	Will County WIB - Instituto del Progreso Latino	OFFER		
IL	Will County WIB - Jewish Vocational Services	OFFER		
IL	Will County WIB - Joliet Junior College	OFFER		
KS	Kansas Dept of Commerce - Heartland Works, Inc.			
KS	Kansas Dept of Commerce - Southeast KANSASWORKS, Inc.			
KS	Kansas Dept of Commerce - Workforce Alliance of South Central Kansas			
KS	Kansas Dept of Commerce - Workforce Partnership			
KS	Kansas Dept of Commerce - WorkforceOne			
KY	Gateway Community and Technical College	OFFER	TEST	
LA	WIB SDA-83 Inc.			
MO	Full Employment Council	TEST	OFFER	OFFER
NE	Central Community College		OFFER	
NH	New Hampshire Office of Minority Health	OFFER		TEST
NJ	Bergen Community College - Bergen Community College	TEST		
NJ	Bergen Community College - Brookdale Community College	TEST		
NJ	Bergen Community College - Community College of Morris	TEST		
NJ	Bergen Community College - Essex County College		TEST	
NJ	Bergen Community College - Hudson County Community College	TEST		
NJ	Bergen Community College - Middlesex County College	TEST		
NJ	Bergen Community College - Passaic County Community College	TEST		
NJ	Bergen Community College - Sussex County Community College	TEST		
NJ	Bergen Community College - Union County College	TEST		
NJ	Bergen Community College - Warren County Community College	TEST		
NY	Research Foundation of CUNY-Hostos Community College	TEST		
NY	Buffalo and Erie County WDC	OFFER		TEST
NY	Schenectady County Community College	OFFER	OFFER	
NY	Suffolk County Department of Labor	OFFER	TEST	
OH	Eastern Gateway Community College	OFFER	OFFER	OFFER
PA	Central Susquehanna Intermediate Unit	OFFER		
SC	South Carolina Department of Social Services		TEST	
TX	Alamo Community College District and University Health System		TEST	
WA	Edmonds Community College	OFFER		
WA	Workforce Development Council of Seattle-King County	OFFER		
WI	Milwaukee Area WIB			

SOURCE: HPOG Evaluation Design Implementation Plans.

Table A2 First-Stage Performance

	Facilitated peer support	Noncash incentives	Emergency assistance
Estimate	0.87***	0.47***	0.41***
	(0.02)	(0.02)	(0.02)
Weak 2SLS F-statistic	2,103	483	408

SOURCE: PRS and short-term follow-up survey.

Table A3 TOT Impacts on Educational Progress among Compliers

	Facilitated peer support (N = 605)	Noncash incentives (N = 818)	Emergency assistance (N = 1,026)
2SLS	−3.9	−11.3	−2.5
	(4.6)	(7.9)	(7.3)
OLS	−0.1	2.4	11.6***
	(4.0)	(4.8)	(3.6)
IPW – Take-up	−2.7	5.2	11.1***
	(3.8)	(3.5)	(2.8)
IPW – Compliers	−2.3	2.3	6.2
	(10.2)	(5.8)	(3.9)

NOTE: Standard errors appear in parentheses. *p < 0.10; **p < 0.05; ***p < 0.01.
SOURCE: PRS and short-term follow-up survey.

Table A4 TOT impacts on educational progress among compliers (including covariates)

	Facilitated Peer Support (N = 605)	Non-cash Incentives (N = 818)	Emergency Assistance (N = 1,026)
2SLS	-6.3	-9.5	-5.2
	(4.5)	(7.5)	(7.1)
OLS	-3.1	7.0	10.3***
	(4.0)	(4.8)	(3.6)
IPW – Take-up	-1.8	5.3	10.5**
	(3.7)	(3.2)	(2.7)
IPW – Compliers	-2.4	4.2	7.2
	(9.6)	(5.3)	(3.9)

NOTE: Standard errors appear in parentheses. * p < 0.10 ** p < 0.05 *** p<0.01.
SOURCE: PRS and short-term follow-up survey.

Table A5 Differences in TOT impacts on educational progress using experimental and nonexperimental variation (including covariates

2SLS Compared to…	Facilitated Peer Support	Non-cash Incentives	Emergency Assistance
OLS	-3.2	-16.5***	-15.6**
	(2.5)	(6.2)	(6.2)
IPW - Take-up	-4.4	-14.8**	-15.8**
	(2.6)	(6.2)	(6.3)
IPW – Compliers	-0.8	-13.9**	-12.3**
	(9.7)	(6.6)	(5.6)

NOTE: Standard errors from 1,000 bootstraps appear in parentheses. * p < 0.10 ** p < 0.05 *** p<0.01.
SOURCE: PRS and short-term follow-up survey.