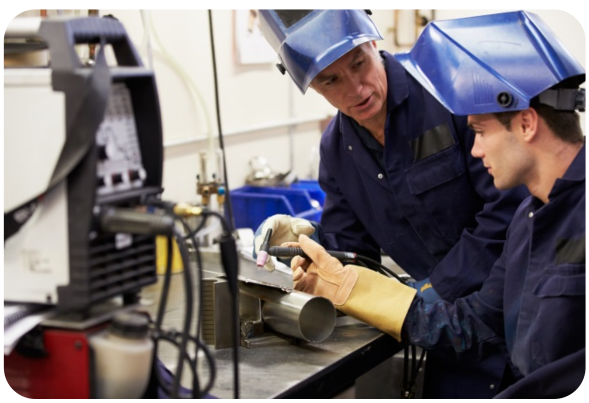
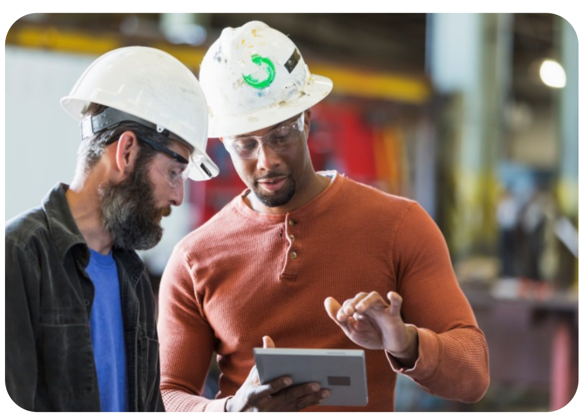
Des Moines Area Community College's Workforce Training Academy Connect Program

Appendices for Three-Year Impact Report

OPRE Report 2021-98



May 2021



PACE
Pathways for Advancing Careers and Education

Des Moines Area Community College's Workforce Training Academy Connect Program: Appendices for Three-Year Impact Report

A Pathways for Advancing Careers and Education (PACE) / Career Pathways Intermediate Outcomes Study Publication

OPRE Report 2021-98

May 2021

David Judkins, Douglas Walton, Gabriel Durham, and Daniel Litwok, Abt Associates

Submitted to:
Nicole Constance and Amelia Popham, Project Officers
Office of Planning, Research, and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services

Contract Number: HHSP23320095624WC, Task Order HHSP23337019T

Project Director: Larry Buron
Principal Investigator: David Fein; Director of Analysis: David Judkins
Abt Associates
6130 Executive Boulevard
Rockville, MD 20852

Contents

List of Exhibits

Appendix A: Baseline Characteristics and Adjustments

This appendix starts with a description of the specification for baseline characteristics, including our approach to handling missing values (Section A.1). The next section compares distributions for treatment and control group members on these and other baseline measures (Section A.2), and the last section explains how the analyses control for these covariates in estimating impacts (Section A.3). It should be noted that Sections A.1 and A.2 are nearly unchanged from parallel appendices in the initial report on this program (Hamadyk and Zeidenberg 2018). In contrast, the approach to covariate control in Section A.3 describes some important procedural changes from those used in that short-term report.

A.1 Details on Baseline Covariates

Exhibit A-1 shows the specifications and data sources for baseline covariates. Item nonresponse rates on these covariates were generally low. Across all nine PACE sites, item nonresponse rates were less than 4 percent except for parental college attendance (6.0 percent), typical high school grades (7.2 percent), family income (9.5 percent), and expected near-term future work hours (6.0 percent).

We imputed values for missing covariates using SUDAAN®/IMPUTE, a weighted hotdeck imputation procedure (Research Triangle Institute 2012). This imputation step entailed a single computer run on the combined sample from all nine PACE sites.[1] With this process, we replaced each missing value with an observed response from a similar case. Within specified strata, we random-matched cases with missing values to cases with reported values; we then copied over the reported value to the case where the value was missing. The strata represented a cross-classification of treatment-control status, site, National Student Clearinghouse (NSC)-reported enrollment status (*some* or *none*),[2] NSC-reported credential award (*some* or *none*), and number of months of NSC-reported enrollment.[3]

---

1 Using the combined dataset better controlled for school enrollment status, as measured in NSC, in the smaller sites.

2 NSC has information on monthly enrollment and many credentials for 96 percent of college students. https://nscresearchcenter.org/workingwithourdata/.

3 In instances where this level of matching was too restrictive because we found no matched case with a reported value, we re-ran the procedure matching only on treatment status and NSC-reported enrollment status. In this second pass imputation, matches were allowed across sites.

Exhibit A-1: Operationalization of Baseline Measures Used as Covariates in Regression-Adjusted Impact Estimates

Variable Description	Operationalization Details	Data Source(s) (Survey Instrument: Survey Item Number)
Demographic Background		
Age	Categorical measure: Under 21 21-24 25-34 35+ [a]	BIF: B2_dob RABIT: R_RA_Date_Assigned
Female	Binary variable: 1 if female 0 if male	BIF: B7
Race/ethnicity	Categorical measure: Hispanic, any race Black, non-Hispanic White, non-Hispanic[a] Another race, non-Hispanic	BIF: B9
Family structure	Categorical measure: Spouse/partner, with children Spouse/partner, without children Single, with children[a] Single, without children (Only biological and adopted children of randomized participant considered here. Stepchildren, grandchildren, younger siblings, and other children not considered.)	BIF: B13
Living with own parents	Binary variable: 1 if living with own parent(s) 0 otherwise (Presence of parents of spouse not considered.)	BIF: B13
Educational Background		
Parent attended college	Binary variable: 1 if either parent attended college 0 otherwise	BIF: B21
Usual high school grades	Categorical measure: Mostly A's Mostly B's Mostly C's or below[a]	BIF: B23
Highest level of education completed	Categorical measure: No college[a] Less than one year of college credit One or more years of college credit Associate degree or above	BIF: B17
Career Knowledge		
Career Knowledge Index (average of items)	Proportion of responses to seven questions about career orientation and knowledge to which respondent answered "strongly agree." Missing if four or more of seven responses blank.	SAQ: S13

Variable Description	Operationalization Details	Data Source(s) (Survey Instrument: Survey Item Number)
Psycho-Social Indices		
Academic discipline[b]	Average of 10 items (scale ranging 1=strongly disagree to 6=strongly agree) after reversing responses to negatively phrased items. Missing if seven or more of 10 responses blank.	SAQ: S11a
Training commitment[c]	Average of 10 items (scale ranging 1=strongly disagree to 6=strongly agree) after reversing responses to negatively phrased items. Missing if seven or more of 10 responses blank.	SAQ: S11b
Academic confidence[d]	Average of 12 items (scale ranging 1=strongly disagree to 6=strongly agree) after reversing responses to negatively phrased items. Missing if nine or more of 12 responses blank.	SAQ: S11d
Emotional stability[e]	Average of 12 items (scale ranging 1=strongly disagree to 6=strongly agree) after reversing responses to negatively phrased items. Missing if nine or more of 12 responses blank.	SAQ: S11e
Social support[f]	Average of 10 items (scale ranging 1=strongly disagree to 4=strongly agree). Missing if seven or more of 10 responses blank.	SAQ: S12
Resource Constraints (Financial)		
Family income in past 12 months	Categorical measure: Less than $15,000 $15,000-$29,999 $30,000+ [a]	BIF: B27
Received food assistance (WIC/SNAP) in past 12 months	Binary variable: 1 if yes 0 if no	BIF: B26b
Received public assistance or welfare in past 12 months	Binary variable: 1 if yes 0 if no	BIF: B26c
Financial hardship in past 12 months	Binary variable: 1 if yes to ever missed rent/mortgage payment in prior 12 months or reported generally not having enough money left at the end of the month to make ends meet over the last 12 months. 0 if otherwise	SAQ: S8, S9
Resource Constraints (Time)		
Current work hours	Categorical measure: 0-19[a] 20-34 35+	BIF: B24
Expected work hours in next few months	Categorical measure for covariate: 0-19[a] 20-34 35+	SAQ: S2
Expecting to attend school part-time if accepted	Binary variable: 1 if yes 0 if no	SAQ: S1

Variable Description	Operationalization Details	Data Source(s) (Survey Instrument: Survey Item Number)
Life Challenges		
Frequency of situations interfering with school, work, job search, or family responsibilities	Average of six items of frequency of problems in past 12 months (childcare, transportation, alcohol or drug use, health, family arguments, physical threats). Scale ranges from 1=never to 5=very often. Missing if four or more of six responses blank.	SAQ: S15
Stress[g]	Average of four items about feeling in control of important things and able to handle personal problems (scale 1=never to 5=very often over the past month) after reversing responses to negatively phrased items. Missing if three or more of four responses blank.	SAQ: S14

Key: BIF = Basic Information Form. RABIT = Random Assignment and Baseline Information Tool. SAQ = Self-Administered Questionnaire. SNAP = Supplemental Nutrition Assistance Program. WIC = Special Supplemental Nutrition Program for Women, Infants, and Children.

[a] Category omitted in creating binary (dummy) variables for regression-adjustment models.

[b] Modified version of the Academic Discipline scale in the Student Readiness Index (SRI), a proprietary product of ACT, Inc.; Le et al. (2005). Further validation in Peterson et al. (2006).

[c] Modified version of Commitment to College scale in the Student Readiness Index (SRI), a proprietary product of ACT, Inc.; Le et al. (2005). Further validation in Peterson et al. (2006).

[d] Modified version of the Academic Self-Confidence scale in the Student Readiness Index (SRI), a proprietary product of ACT, Inc.; Le et al. (2005). Further validation in Peterson et al. (2006).

[e] Modified version of the Emotional Control scale in the Student Readiness Index (SRI), a proprietary product of ACT, Inc.; Le et al. (2005). Further validation in Peterson et al. (2006).

[f] Modified version of the Social Provisions Scale; Cutrona and Russell (1987). Original scale has 24 items. This short version developed by Hoven (2012).

[g] Cohen et al. (1983).

A.2 Comparing Treatment and Control Groups at Baseline

Exhibit A-2 shows tests for similarity in characteristics of treatment and control group members at baseline. If the means in the two columns are congruent, then “baseline balance” was achieved.

The last column contains *p*-values for tests of hypotheses of no systematic differences between the treatment and control groups. On average, one would expect that out of 28 tests, three will fall outside a 90 percent confidence interval due to chance. In this case, there were six statistically significant differences (in red font). The research team found no evidence that there were any problems with random assignment. It is likely that these are simply chance results. Furthermore, as described in the next section, regression adjustment helps to control for any effects such chance differences might have on the impact estimates.

Exhibit A-2: Baseline Balance

Characteristic	All Participants	Treatment Group	Control Group	*p*-Value
Age (%)				.236
20 or under	14.3	14.7	14.0	
21-24	16.4	17.2	15.6	
25-34	27.7	24.7	30.7	
35+	41.6	43.4	39.8	
Female (%)	62.6	65.0	60.3	.130
Race/Ethnicity (%)				.414
Hispanic, any race	15.3	13.9	16.7	
Black, non-Hispanic	47.4	50.2	44.7	
White, non-Hispanic	33.8	33.3	34.3	
Another race, non-Hispanic	7.1	6.3	7.8	
Family Structure (%)				.019
Not living with spouse/partner and not living with children	49.5	48.6	50.4	
Not living with spouse/partner but living with children	20.1	24.1	16.2	
Living with spouse/partner and not living with children	19.9	18.3	21.5	
Living with spouse/partner and children	10.5	9.1	11.8	
Living with parents (%)	16.9	16.8	17.1	.907
One parent has at least some college (%)	30.2	30.6	29.8	.812
Usual High School Grades (%)				.814
Mostly A's	8.3	8.5	8.1	
Mostly B's	36.9	35.8	38.0	
Mostly C's or below	54.8	55.7	53.9	
Highest Level of Education (%)				.054
Less than a high school diploma	40.1	39.1	41.1	
High school diploma or equivalent	36.8	35.7	37.9	
Less than one year of college	10.8	13.1	8.4	
One or more years of college	8.2	9.0	7.3	
Associate degree or higher	4.2	3.0	5.4	
Received vocational or technical certificate or diploma (%)	20.9	20.7	21.0	.896
Career Knowledge Index (mean)	0.36	0.37	0.35	.312
Psycho-Social Indices (means)	4.92	4.95	4.89	.259
Academic Discipline Index	5.34	5.40	5.28	.007
Training Commitment Index	4.32	4.35	4.29	.218
Academic Self-Confidence Index	4.85	4.87	4.82	.341
Emotional Stability Index	3.09	3.12	3.06	.030
Social Support Index	2.58	2.56	2.60	.397
Stress Index	1.82	1.81	1.84	.431
Depression Index	4.92	4.95	4.89	.259
Family Income in Past 12 Months (%)				.571
Less than $15,000	56.0	56.8	55.1	
$15,000-$29,999	26.1	26.6	25.5	
$30,000+	18.0	16.6	19.4	
Family income (mean)	$16,364	$15,783	$16,966	.289

Characteristic	All Participants	Treatment Group	Control Group	*p*-Value
Public Assistance/Hardship Past 12 Months				
Received WIC or SNAP (%)	65.8	68.8	62.8	.057
Received public assistance or welfare (%)	14.4	14.6	14.2	.878
Reported financial hardship (%)	62.7	62.4	62.9	.825
Current Work Hours (%)				.674
0	62.2	61.7	62.7	
1-19	5.1	6.0	4.2	
20-34	13.3	12.8	13.8	
35+	19.5	19.6	19.3	
Expected Work Hours in Next Few Months (%)				.394
0	22.4	20.1	24.6	
1-19	4.7	5.3	4.1	
20-34	27.9	29.0	27.0	
35+	45.1	45.7	44.4	
Life Challenges Index (mean)	1.77	1.79	1.75	.355
Owns a car (%)	59.2	62.0	56.3	.084
Has both computer and internet at home (%)	50.6	49.6	51.5	.545
Ever arrested (%)	41.1	41.3	40.9	.870
Sample sizes	943	470	473	

Source: PACE Basic Information Form and Self-Administered Questionnaire.

Note: Tests for statistically significant imbalance were based on SAS®/FREQ procedure for categorical outcomes and on the SAS®/TTEST procedure for other outcomes. Significant imbalances are highlighted in red, using a threshold for statistical significance of 10 percent. All values are based on baseline balance prior to imputation of missing values.

A.3 Regression Adjustment

This section describes the regression adjustment approach used to improve precision and minimize effects of sampling error on impact point estimates.

In a rigorous evaluation, random assignment ensures that if the sample size is large enough, differences in average potential outcomes between the treatment and control groups will become vanishingly small so that any observed differences in average outcomes across the two groups must almost certainly be the result of treatment.[4] Even when sample sizes are modest, random assignment ensures that differences in average potential outcomes between the treatment and control groups arise from chance rather than biases of program operators or program evaluators. This means that unbiased estimates of the effects of treatment can be obtained by simply comparing average outcomes across the treatment and control groups. Moreover, it is easy to run formal tests of the hypothesis that the program has no effect (and

4 Potential outcomes are a central concept in the Neyman-Rubin causal model (Holland 1986). In this model, each person has an innate pair of possible outcomes: one if treated and the other if not treated. Only one of the two possible outcomes is ever observed for each person. The average difference in possible outcomes across a specific population is said to be the *local average treatment effect* (LATE)—or more simply, just the "effect of treatment," with the context making clear the population to which it applies and supplemental analyses exploring whether the effect is homogenous within that population.

that therefore the observed difference in mean outcomes is the result of those accidental imbalances in potential outcomes across the two groups).

Despite these favorable properties of analysis based on simple comparisons of observed means, use of regression adjustment can reduce the impact of accidental imbalances in potential outcomes across the groups, thereby increasing power to detect small program impacts (Lin 2013). To achieve this benefit, the variables used in the regression adjustment must be predictive of potential outcomes. Including other variables will increase the variance on the estimated program impact rather than decreasing it.

Opinions and practice differ on how strong the evidence for correlation between a baseline variable and the outcome must be before it makes sense to include the baseline variable in the regression adjustment.[5] Some favor a lean approach, including just those baseline variables that have a demonstrated strong relationship to the outcome. Others favor a more comprehensive approach including all baseline variables that have a plausible theoretical relationship to outcomes of interest, believing that doing so generally bolsters confidence in study findings (Tukey 1991).

Given demands to minimize burden on participants, all measured PACE baseline variables have at least plausible relationships to PACE outcomes, but some baseline variables have been discovered to have only weak empirical relationships with PACE outcomes. Moreover, one could combine the directly measured characteristics into a limited number of interactions. So some judgment must be exercised about which covariates to include in regression adjustments and which to exclude.

Opinions and practice also differ on how much to customize decisions about covariate inclusion across outcomes in evaluations (such as this PACE evaluation of WTA Connect) with multiple outcomes. A single uniform set of decisions promotes transparency, making it easier for readers to understand the procedure, whereas a more customized approach is likely to improve variances for at least some outcomes given that the correlation between a covariate and an outcome will vary by outcome.

In preliminary analyses for the first round of PACE reports, the team planned to use a fairly comprehensive approach with a uniform set of decisions but discovered that this approach was causing the variances on adjusted impacts to be larger than the variances on unadjusted impacts. The discovery prompted a switch to a different approach for the first round of reports, which ultimately proved not to work as well as hoped (Judkins 2019). In response, the team developed a new approach for the current round of PACE reports. This new approach emphasizes transparency and control on imbalanced covariates, while still trying to maximize precision as far as possible given those priorities. Details follow.

Equation (A.1) below shows the conventional regression-adjustment model:

$$Y_i = X_i\beta + \delta T_i + e_i \tag{A.1}$$

5 For a current review of practice, see Ciolino et al. (2019).

where Y_i is the outcome; X_i is a row vector of baseline characteristics (hereafter referred to as covariates); β is the vector of parameters indicating the influence of each covariate on the outcome; δ is the effect of treatment; T_i is a 0/1 dummy variable indicating treatment group membership; and e_i is an error term. We fit models of this sort using SAS®/SurveyReg, a procedure that uses a robust estimator of the variance of $\hat{\delta}$ and can accommodate the required nonresponse-adjustment weights for survey-measured outcomes. (See Appendix Section B.3 for a discussion of nonresponse-adjustment weights.)

This method is known as ordinary least squares (OLS) and has excellent properties when the sample size is many times larger than the number of baseline characteristics used as covariates (Lin 2013), even when the outcomes are not normally distributed (Judkins and Porter 2016). Estimates of the treatment effect are "asymptotically unbiased," and under most conditions, the variance of the estimated treatment effect declines from the simple difference-in-mean-outcomes estimator of impact in proportion to the amount of outcome variation explained by the covariates.

Specifically, the relationship between the variance of the estimated treatment effect from the OLS estimation of Equation (A.1) and the explanatory power of the covariates is $var\left(\hat{\delta}\right) \approx (1 - R^2)var(\bar{y}_t - \bar{y}_c)$, where R^2 is the proportion of the variance in Y_i explained by the baseline characteristics (X_i) in OLS estimation of Equation (A.2) below:

$$Y_i = X_i\beta + e_i \tag{A.2}$$

However, as mentioned above, when there are a large number of potential covariates, not all of which are useful in predicting every outcome of interest, the effect of adjustment can be the opposite of the intended effect: variances are increased rather than decreased.[6] To avoid unnecessary variance inflation, the analyst needs to drop or otherwise reduce the influence of extraneous covariates that do not have a strong influence on the outcome of interest.

Simulation research (Judkins 2019) showed that dropping (with "backward selection") or downweighting covariates[7] based on simple analyses of the same data used in the evaluation yields slightly biased estimates of the variance of the estimated treatment effects (but still unbiased estimates of the treatment effect itself). This bias is negative, meaning that the variance estimates are slightly too small, making confidence intervals for impact estimates

[6] Mathematically, the presence of extraneous variables causes the coefficients of the true determinants of the outcome to be less accurately estimated. For example, if the best prediction model is $Y = 2X$ but the model is fit with many extraneous covariates, the fit prediction formula could easily end up having coefficients of 1.9 or 2.1 for X instead of the best value of 2. If the wrong slope is used to correct for a treatment-control imbalance in X, the adjusted estimate of impact can be worse than an unadjusted estimate of impact.

[7] An example of a method that downweights covariates is the "modified Koch method" developed for and used in the first round of PACE reports (Judkins et al. 2018; Koch et al. 1998).

misleadingly narrow and hypothesis tests too likely to conclude that a nonzero impact has occurred when the true impact is zero or negative. [8]

To select covariates in a manner that does not compromise variance estimation, we use the relatively recently developed technique "least absolute shrinkage and selection operator" (LASSO) with "10-fold cross-validation."[9] With the LASSO, the sum of absolute values of the estimated regression coefficients in Equation (A.2) is constrained to be less than a preselected value (the "constraint"). If the value for this constraint is small enough, many coefficients in Equation (A.2) will be forced to zero to fit within the cap on the sum of absolute coefficient values and thus can be removed from the list of baseline covariates. The 10-fold cross-validation is used to optimize the value of the constraint, rather than just relying on an arbitrary choice for it.

Details of the procedure are as follows:

(1) With 10-fold cross-validation, the sample (both treatment and control group members) is divided into 10 equal and mutually exclusive random subsamples.

(2) For each of a range of candidate values of the constraint, the LASSO procedure is run to select covariates on a sample in which one of the 10 subsamples has been dropped.

(3) The model in Equation (A.2) is fit on the same sample using just the variables selected in the second step for each of the candidate values of the constraint.

(4) The model is used to create out-of-sample predictions of the outcome for everyone in the dropped piece of the sample, and the prediction error is $\hat{Y}_i - Y_i$ measured for each of the candidate values of the constraint.

(5) Steps 2 through 4 are repeated 10 times for each candidate value of the constraint. On each iteration, a different one of the 10 subsamples is dropped. In this manner, out-of-sample prediction errors are obtained for the entire sample.

(6) Mean squared prediction errors across all 10 replicates are then calculated for each of the candidate values of the constraint.

(7) The value of the constraint that minimizes this cross-validated mean squared prediction error and thus captures most of the variation reduction possible with the available

8 If the sample size is very large, the estimated variance of the estimated effect of treatment will be nearly unbiased even if the evaluation data are used to cull or downweight extraneous covariates. However, simulations clearly show that PACE sample sizes are not large enough to avoid biased variance estimates if "backward selection" on local data is used to prune covariates or if the modified Koch method is used to downweight extraneous covariates. Accordingly, impact analyses at three years for WTA Connect and all other PACE local programs are not using the modified Koch method used in the initial round of reports covering the first 18 months of follow-up.

9 See Bühlmann and van de Geer (2011) for a full explanation of these techniques.

covariates is selected as the optimal constraint.[10] Whichever variables have nonzero coefficients in the model for that optimal constraint are used as covariates in the impact regressions. All other baseline characteristics are discarded. All of this is done automatically in SAS®/GLMSELECT. Simulations carried out under PACE-like conditions (in terms of sample sizes and the numbers of covariates) when developing the analysis plan for the entire suite of PACE three-year reports (Judkins et al. 2018) demonstrate that this technique reduces the true variances without biasing variance estimates.[11]

In principle, we could repeat the LASSO with 10-fold cross-validation independently for every outcome for each of the nine PACE programs. But such an approach would produce a different final covariate list for each outcome and program, leading to some loss in transparency and making it harder for outside researchers to replicate the PACE results. At the other extreme, we could run the LASSO just once for each program for the most important confirmatory outcome and then use the resulting set of selected covariates for all impact estimates for the program. But we believe that this would result in more precision loss than can be justified for the sake of transparency.

As a compromise between these extremes, we selected one set of covariates for each of three domains and customized them for each of the nine PACE programs. The three domains are (1) analyses of *employment and earnings* outcomes that are conducted on the dataset of merged data from the three-year follow-up survey and the National Directory of New Hires (NDNH); (2) analyses of *education* outcomes (whether based on the survey, NSC, or local or state college records); and (3) analyses of all *other* outcomes (most of which concern personal and family well-being and economic independence). The pool of potential covariates was the same for all three domains—with one important exception: indicators of pre-baseline earnings based on NDNH data are only allowed in analyses of NDNH-based outcomes.[12]

To identify covariates for this report, we ran the LASSO procedure for the most salient outcome within each of the three domains (*earnings and employment, educational progress, other*) at each of the nine PACE programs.[13] For NDNH analyses, the confirmatory outcome is average quarterly earnings for the 12th and 13th quarters after randomization (Q12, Q13), so that is a natural choice for the outcome around which to optimize covariate selection. In the *educational progress* domain, the most important outcome varies by PACE program. As discussed in the

---

10 One could simply use the LASSO to select covariates with a pre-specified value of the constraint, but the 10-fold cross-validation provides a principled method for selecting the constraint.

11 See Judkins (2019) for additional detail.

12 This is because we analyzed survey outcomes on Abt's secure server rather than on the ACF secure server. Though both systems have very high security procedures, agreements with the Office of Child Support Enforcement (OCSE) permit the NDNH data to reside only on the ACF secure server. It would have been possible to analyze all survey outcomes on the ACF secure server, but doing so would have significantly burdened the study's analytic operations without any commensurate benefit. It would also prevent us from analyzing survey data for people whose names and Social Security numbers do not match properly according to OCSE.

13 Selection started with the set of baseline covariates used in the analyses of follow-up data at 18 months after random assignment for the short-term impact report (shown in Exhibit A-3).

body of this report, for Workforce Training Academy Connect (WTA Connect) the most salient education outcome is receipt of exam-based certifications and licenses. As the most salient outcome for the third domain, we selected whether anyone in the household draws means-tested public benefits. We made this last decision because of the centrality of the concept of self-sufficiency in the rationale for creating the PACE project.[14] We made these choices prior to reviewing any impact estimates.

In addition to covariates based on the above procedures, regression models included covariates for which baseline distributions differ for treatment and control group members at the 5 percent level.[15]

Exhibit A-3 below shows the covariates that we selected with the LASSO procedure or by virtue of their being out of balance (OOB) at baseline. For categorical variables, the LASSO procedure worked on dummy variables for the individual levels; so for a variable with four levels, it was possible for just one dummy variable to be selected. In contrast, the out-of-balance test selected all or none of the levels of a categorical variable. The exhibit shows all possible levels of categorical variables and indicates which specific categories we selected as covariates. So, for example, LASSO selected two age levels as educational covariates, but all Family Structure levels were flagged as OOB and included as covariates for every outcome.

Exhibit A-3: Covariates Selected by Outcome Domain

Baseline Covariate	NDNH-Based Employment and Earnings Domain	Educational Progress Domain	Other Domains
Age			
20 or under		LASSO	
21-24		LASSO	
25-34			
35+			
Sex			
Female			LASSO
Male			
Race/Ethnicity			
Hispanic, any race			
Black, non-Hispanic			
White, non-Hispanic			
Another race, non-Hispanic		LASSO	

14 The original name for PACE was "Innovative Strategies for Increasing Self-Sufficiency." The promotion of self-sufficiency is also central to the goals of the career pathways framework as articulated by Fein (2012).

15 Baseline balance was assessed prior to imputation of missing data. See Exhibit A-2.

Baseline Covariate	NDNH-Based Employment and Earnings Domain	Educational Progress Domain	Other Domains
Family Structure	OOB	OOB	OOB
Not living with spouse/partner and not living with children			
Not living with spouse/partner but living with children			
Living with spouse/partner and not living with children			
Living with spouse/partner and children			
Living with parents		LASSO	
One parent has at least some college			
High School Grades			
Mostly A's			
Mostly B's			
Mostly C's or below			
Current Education			
High school diploma equivalent or less			
Less than one year of college			
One or more years of college			
Associate degree or higher			
Career Knowledge Index			
Family Income in Past 12 Months			
Less than $15,000	LASSO		LASSO
$15,000-$29,999		LASSO	
$30,000+			
Pre-Randomization Quarterly Earnings (NDNH)		Not available	Not available
4 quarters prior to randomization	LASSO		
3 quarters prior to randomization	LASSO		
2 quarters prior to randomization	LASSO		
1 quarter prior to randomization	LASSO		
Pre-Randomization Quarterly Employment (NDNH)		Not available	Not available
4 quarters prior to randomization			
3 quarters prior to randomization			
2 quarters prior to randomization			
1 quarter prior to randomization			
Psycho-Social Indices			
Academic Discipline Index	OOB	OOB	OOB
Training Commitment Index			
Academic Self-Confidence Index			
Emotional Stability Index	OOB	OOB	OOB
Stress Index			
Life Challenges Index			LASSO
Public Assistance/Hardship Past 12 Months			
Received WIC or SNAP			LASSO
Received public assistance or welfare			
Reported financial hardship			

Baseline Covariate	NDNH-Based Employment and Earnings Domain	Educational Progress Domain	Other Domains
Current Work Hours			
0-19			
20-34			
35+			
Expected Work Hours in Next Few Months			
0-19			
20-34			
35+			
Plan to attend school only part-time if admitted to WTA Connect			

Key: SNAP = Supplemental Nutrition Assistance Program. WIC = Special Supplemental Nutrition Program for Women, Infants, and Children.

Source: PACE Basic Information Form and PACE Self-Administered Questionnaire.

Note: "LASSO" flags that the covariate was selected by the LASSO for variance reduction. "OOB" flags that the covariate was selected because it was significantly out of balance.

Exhibit A-4 below shows impacts on selected confirmatory and secondary outcomes before and after regression adjustment without weights.[16] The two sets of estimates lead to similar impact estimates, and most (seven out of 12) of the regression-adjusted standard errors are equal to or smaller than the unadjusted standard errors.

16 We did not use the weights in the preparation of this table because they are not required for the first panel (Full Sample) and because in this section we want the focus to be on the role of covariates. See Appendix Exhibit B-11 for the impact of nonresponse-adjustment weights on these estimates.

Exhibit A-4: Comparison of Confirmatory and Secondary Impact Estimates Unadjusted and Adjusted for Baseline Imbalances

Outcome	Impact (Unadjusted Estimate)	Standard Error	Impact (Adjusted Estimate)	Standard Error
Confirmatory Outcome: Employment (NDNH)	**Full Sample**			
Average quarterly earnings Q12-Q13 after randomization ($)	175	261	86	224
Secondary Outcome: Employment (Survey)	**Survey Respondents without Weights**			
Employed at survey follow-up (%)	0.7	3.6	0.8	3.6
Employed at $13 per hour or above (%)	−1.3	3.4	−0.7	3.2
Employed in a job requiring at least mid-level skills (%)	1.8	2.3	2.3	2.3
Secondary Outcome: Education (Survey)	**Survey Respondents without Weights**			
Full-time-equivalent months enrolled at any school (months)	−0.1	0.4	−0.1	0.4
Receipt of an exam-based certification or license (%)[a]	6.7***	2.8	5.9**	2.8
Secondary Outcome: Other (Survey)	**Survey Respondents without Weights**			
Indicators of Independence and Well-Being				
Has health insurance coverage (%)	−3.1	2.2	−3.7	2.2
Receives means-tested public benefits (%)	6.3	3.1	4.0	2.9
Personal student debt ($)	−294	363	−457	384
Any signs of financial distress (%)	−1.8	3.7	−3.5	3.6
Indices of Self-Assessed Career Progress (average)				
Confidence in career knowledge[b]	0.09**	0.05	0.08*	0.05
Access to career supports[c]	0.01	0.02	0.00	0.02
Sample sizes (across treatment and control groups): NDNH 920; Survey 698				

Source: PACE 18-month follow-up survey; PACE three-year follow-up survey; National Directory of New Hires.

[a] Blended 18-month and three-year survey results.

[b] Seven-item scale tapping self-assessed career knowledge; response categories range from 1=strongly disagree to 4=strongly agree.

[c] Six-item scale tapping self-assessed access to career supports; response categories range from 1=no to 2=yes.

Statistical significance levels, based on one-tailed *t*-tests of positive differences between research groups for positive outcomes and negative differences for negative outcomes (such as student debt), are summarized as follows: *** 1 percent level; ** 5 percent level; * 10 percent level.

Appendix B: Three-Year Survey Data

This appendix documents key technical detail underlying analyses of the three-year follow-up survey data for Workforce Training Academy Connect (WTA Connect).[17] Section B.1 documents coding for scales based on follow-up survey data. Section B.2 describes the imputation process for some missing survey data elements in the construction of outcomes. Section B.3 analyzes survey nonresponse and documents the process we used to build the nonresponse weights used in the impact analysis. Sections B.4 and B.5 present evidence about the quality and completeness of survey responses. Before getting into those details, we provide an overview of the measurement goals and structure of the survey instrument.

The survey sought to collect a complete history of jobs and periods of schooling since randomization (including the progression and interleaving of these spells), credits and credentials earned, earnings growth, and self-employment. In addition, the survey measured 21st century skills, family formation and growth, income and material well-being, and child outcomes.

The goal of the Integrated Training and Employment History module of the three-year survey was to collect a complete history of training and employment between randomization and the day of interview three years later. Given data collection plans, the approach needed to work over the phone. The instrument development team reviewed several past efforts to collect such histories, but only one of the past approaches seemed likely to be workable over the phone—an approach developed for a German survey instrument that studies the training and work histories of German youth.[18] This was the first time that the German approach had been attempted in the United States.

Conceptually, a history could be built either forward from randomization or backward from the day of interview. The German study worked forward with apparent success, so we adopted that approach. One modification we made was to take each respondent through his or her training and employment history twice instead of just once. First, the survey collects the spell history (dates, whether work or school, and place names). This is the “scaffolding.” Once the scaffolding has been built, the interviewer takes the respondent back through the history a second time to systematically collect more information about each training spell.

---

17 The full instrument is available at http://www.career-pathways.org/career-pathways-pace-three-year-instrument/.

18 The 2011 BIBB Transitional Study was a retrospective longitudinal survey conducted by the *Bundesinstitut für Berufsbildung* (Federal Institute for Vocational Education and Training) on a representative basis that recorded in detail the whole of the educational and occupational biographies of persons born between 1987 and 1992 and resident in Germany. For full details, see Beicht and Friedrich (2008). For a brief English synopsis of one report using some of the survey data: http://www.bibb.de/en/64317.htm.

There are two motivations for this two-pass approach:

(1) By asking the respondent to focus on one type of information at a time, collection of date data may be more consistent across spells.

(2) This approach allows more-straightforward programming.

B.1 Measures Based on Follow-up Survey Data

Exhibits in this section detail the operationalization of survey-based outcomes used in impact analyses in the main report. These exhibits also reference the underlying survey questions. Exhibit B-1 provides details on outcomes in the education domain, as reported in the report's Chapter 3. Exhibit B-2 provides similar details on outcomes in the employment/earnings domain as reported in Chapter 4. Finally, Exhibits B-3, B-4, and B-5 do the same for intermediate outcome domains, other life outcomes domains, and child outcomes, respectively, as reported in Chapter 5.

Exhibit B-1: Details on Specifications for Survey-based Education Outcomes in Chapter 3

Outcome	Details on Derivation of Outcome	Follow-up Survey Question(s)
Secondary Outcomes		
Education		
Full-time-equivalent months enrolled at any school through 35 months after randomization	Students were asked for the dates of attendance of each school attended and their status while enrolled. If their status was "part-time," then the number of months was multiplied by 0.25 to estimate full-time-equivalent months. Similarly, if their status was "equal mix," then number of months was multiplied by 0.50 to estimate full-time-equivalent months. We developed this rule based loosely on guidance in NSC documents about how schools should classify less-than-full-time enrollment. Because the survey response categories were different from those used in the NSC and because students might have different understandings than schools did, this decision was fairly arbitrary. Alternate rules might have worked just as well.	C1, C2, C3, D2
Received an exam-based certification or license	Respondents were asked whether they had "received a professional, state, or industry certification, license, or credential from an authority other than a school." This measure uses the18-month survey for exam-based credentials reported through the time that survey was completed and uses the three-year survey for exam-based credentials that were reported to be earned after completion of the short-term survey.	3-year: I3d, I3di, I3h 18-month: A56, A56a

Outcome	Details on Derivation of Outcome	Follow-up Survey Question(s)
Exploratory Outcomes		
The survey asked about credentials in three different ways. (1) The survey first asked whether respondents had received "a diploma, certificate, or academic degree for completing any regular college classes." Among those answering the question affirmatively, the survey asked for a list of such credentials and more about each one, including issuing school, award date, and (for sub-degree credentials) the name and typical length of study required to earn it. (2) The survey then asked whether they had received "any diplomas or certificates from a school for completing any vocational training." Among those answering the question affirmatively, the survey asked for a list of such credentials with no further detail beyond issuing school and award date. (3) Finally, the survey asked whether they had "received a professional, state, or industry certification, license, or credential from an authority other than a school." Among those answering the question affirmatively, the survey asked for a list of such credentials, award dates, and the type of authority issuing it. In post-survey processing, PACE research team imputed the required length of study required to earn the credential based on the respondent-provided name of the credential. This imputation had three levels: less than a year, a year or more but not as much as for an associate degree, as much as an associate degree, or as much as a bachelor's degree. The research team also standardized the names of issuing schools and matched them to the IPEDS database to determine the school control (public, private nonprofit, and private for-profit) and school level (college, other Title IV school). The team imputed schools with web sites that clearly did not match to IPEDS to be non–Title IV schools. These reported and imputed data then served as the basis for the following recodes:		I2, I2a_2, I2c, I3, I3a_1, I3c, I3d, I3di, I3h
Received any type of credential from any school	Credential earned through colleges and non-college schools.	
Received any type of credential requiring less than a year of credits from any school	In addition to the procedures listed above, respondents were asked how long a period of study was required for their credential.	
Received any type of credential requiring at least a year of credits from any school	Ibid.	
Enrolled in training or education at survey follow-up	Determined based on reported dates of enrollment in education and training activities and date of interview.	Most of modules B, C, and E
College enrollment by quarter	Respondents were asked to list the periods following randomization for which they were enrolled in college. Respondents were also asked to classify enrollment as full-time or part-time. We used these responses to determine quarterly college enrollment	Most of module D

Key: IPEDS = Integrated Postsecondary Education Data System. NSC = National Student Clearinghouse.

Source: Three-year follow-up survey.

Exhibit B-2: Details on Specifications for Survey-based Employment/Earnings Outcomes in Chapter 4

Outcome	Details on Derivation of Outcome	Follow-Up Survey Question(s)
Secondary Outcomes		
Employment		
Employed at survey follow-up	Determined based on reported dates of jobs and date of interview.	Most of modules B, C, and E
Career Progress		
Earning $13 per hour or more	Analyzed response to survey question for control group. Selected $13 per hour as the threshold because it was close to the 60th percentile of hourly wages among employed control group members. This percentile was picked as being a reasonable goal for graduates of WTA Connect.	F5
Working in a job requiring at least mid-level skills	Three open-ended questions about the kind of work done, the usual activities completed, and the job title were coded into an SOC code. We then looked up the Job Zone[a] for each SOC code in the O*NET system.[b] Job Zone 3—occupations that need medium preparation—seemed a reasonable goal for graduates of WTA Connect.	G2a, G3, G4
Exploratory Outcomes		
Working at least 32 hours per week	Currently employed respondents were asked about their typical hours worked.	F6
Working straight day, evening, or night shifts	Currently employed respondents were asked about their typical work schedule. Answer possibilities included straight shifts, rotating shifts, split shifts, irregular schedules, and other.	G6, G6a
Working in job that offers health insurance	Currently employed respondents were asked whether health insurance was available through the employer as a fringe benefit.	G8a
Working in job with supportive working environment	Questions about job benefits and conditions were used to cluster jobs into three categories. Jobs in this category generally provided employees with flexibility to balance work and family, a supportive set of co-workers and supervisors, a rich set of benefits, and opportunities for advancement.	G7, G8a-G8e, G9, G10
Working in a job closely aligned with training	Respondents chose from three response options to a question about alignment of job to training: "Closely related," "Somewhat related, or "Not related." This question was asked of all employed persons even if they had no postsecondary training.	G11

Key: SOC = U.S. Department of Labor Standard Occupational Classification.

[a] https://www.onetonline.org/help/online/zones [accessed September 12, 2016].

[b] https://www.onetonline.org/ [accessed September 12, 2016]. There are five Job Zones. A Job Zone is a group of occupations that are similar in education needed to do the work, related experience needed to do the work, and amount of on-the-job training needed to do the work. Job Zone 3 is described in the O*NET system documentation as "Employees in these occupations usually need one or two years of training involving both on-the-job experience and informal training with experienced workers. A recognized apprenticeship program may be associated with these occupations."

[c] Being employed in a healthcare occupation is usually associated with employment in the healthcare industry, but this is not always true. School nurses are one example of a healthcare worker being employed in an industry other than healthcare. Conversely, many people employed in the healthcare industry are not healthcare workers. Hospital janitors are one example. The survey did not ask about industry of employer.

Exhibit B-3: Details on Specifications for the Career Knowledge, Availability of Career Supports, and Psycho-social Skills Outcomes in Chapter 5

Outcome	Details on Derivation of Outcome	Follow-Up Survey Question(s)
Secondary Outcomes		
Access to career supports	This was a new scale created for PACE at the 18-month follow-up. It is a six-item scale counting number of types of career-supportive relationships in workforce and education settings. The motivation for creating this scale was the theory that richer social networks are one of the benefits of higher education (e.g., Goldrick-Rab and Sorensen 2010). Say you need advice or help in taking a next step on a career pathway of interest to you. Please tell me if there is anyone you'd be comfortable turning to: • Who has a college degree? • Who is currently going to college? • Who works at a local college, either as a teacher or staff member providing help to applicants or students? • Who works for a local community organization helping people find education and training, work, and related supports? • Who works in an occupation of interest to you? • Who has a management job in a work setting matching your career interests?	K4
Confidence in career knowledge	This seven-item scale was based on a review of six survey instruments as well as literature. The first two scale items (a, b) were adapted from the Career Decision Self-Efficacy–Short Form (Betz and Taylor 2001). Three items (d, e, f) were adapted from the Career Exploration Survey (Stumpf, Colarelli, and Hartman 1983). Two items (c, g) were new and written specifically for the PACE Basic Information Form. Response categories ranged from 1=strongly disagree to 4=strongly agree. a. You know how to accurately assess your abilities and challenges? b. You know how to make a plan that will help achieve your goals for the next five years? c. You know how to get help from staff and teachers with any issues that might arise at school? d. You know the type of job that is best for you? e. You know the type of organization you want to work for? f. You know the occupation you want to enter? g. You know the kind of education and training program that is best for you?	K6
Exploratory Outcomes		
Perceived career progress	This was a new scale created for PACE at the 18-month follow-up. It is a three-item scale of self-assessed career progress. Response categories range from 1=strongly disagree to 4=strongly agree. It was designed specifically to measure a respondent's sense of progress in a career pathways program as described by Fein (2012). • I am making progress towards my long-range educational goals • I am making progress towards my long-range employment goals • I see myself on a career path	I5, I6
Grit	Existing scale from Duckworth et al. (2007). The eight-item scale captures persistence and determination. Response categories ranged from 1=strongly disagree to 4=strongly agree.	K1

Outcome	Details on Derivation of Outcome	Follow-Up Survey Question(s)
Core self-evaluation	Existing scale from Judge (2009). The 12-item scale's response categories ranged from 1=strongly disagree to 4=strongly agree. Core self-evaluations (CSEs) represent a stable personality trait that attempts to capture one's self-perception. A positive self-image will correspond to a higher CSE, whereas those who view themselves more negatively will score lower in this category. This trait involves four personality dimensions: locus of control, neuroticism, generalized self-efficacy, and self-esteem. Various studies have shown CSE scores to have predictive ability for work outcomes such as job satisfaction and job performance.[a]	K3
Index of Life Challenges	A new scale adapted for PACE from a longer instrument by Kessler et al. (1998). Average of five items of frequency of situations that interfered with school, work, job search, or family responsibilities. The response categories ranged from 1=never to 5=very often. Missing if four or more responses are blank. • Childcare arrangements • Transportation • Alcohol or drug use • An illness or health condition • Another situation	K7
Social Support Index	Existing scale from Hoven (2012). The 10-item scale response categories ranged from 1=strongly disagree to 4=strongly agree. It is a short-form version of the Social Provisions Scale of Cutrona and Russell (1987), a scale that has 24 items.	K5
Stress Index	Existing scale from Cohen, Kamarck, and Mermelstein (1983). This scale was first used in the PACE Basic Information Form, and has since then been included in both follow-up surveys. The response categories ranged from 1=never to 4=very often.	K8

[a] Judge and Bono (2001); Judge, Locke, and Durham (1997, 1998).

Exhibit B-4: Details on Specifications for Survey-based Family Economic Well-being Outcomes in Chapter 5

Outcome	Details on Derivation of Outcome	Follow-Up Survey Question(s)
Secondary Outcomes		
Personal student debt	Respondents were asked about personal borrowing to go to school since randomization. For those who had difficulty answering the question about the exact amount, a categorical response option was offered. These were then imputed to continuous levels.	M6, M6a
Has health insurance coverage	Includes the offer of healthcare by employer or actual receipt if not offered by employer.	G8a, M12
Receives any means-tested public benefits	Respondents were asked whether they or anyone else in their household received TANF, SNAP, WIC, Medicaid, subsidized childcare, Section 8 or Public Housing, LIHEAP, or FRPL.	M3a, M3b, M3c, M3e, M3f, M3g, M3h, M3i

Outcome	Details on Derivation of Outcome	Follow-Up Survey Question(s)
Any signs of financial distress	For the three-year follow-up, this scale is an expanded version of the *financial hardship measure* used in the 18-month follow-up survey. It flags any signs of financial distress: utility disconnects (gas/electric/oil, telephone), delayed healthcare, delayed dental care, delayed prescription drug procurement, not having enough to eat (sometimes or often), trouble paying bills (rent/mortgage, gas/oil/electricity), or not having enough money to make ends meet at the end of the month.	M9a-g, M10, M11
Exploratory Outcomes		
Average monthly personal income	Respondents were first asked to provide an open-ended amount for the prior month, specifically excluding income tax refunds. If no answer was given, the respondent was asked to choose one of seven bracketed amounts. Item nonresponse was multiply imputed. Exact amounts were also multiply imputed for people who chose a bracket.	M2, M2a
Average monthly household income	Respondents were first asked to provide an open-ended amount for the prior month, specifically excluding income tax refunds, where the household was clarified to include anyone who lived in the household for at least half of the prior month. If no answer was given, the respondent was asked to choose one of seven bracketed amounts. Item nonresponse was multiply imputed. Exact amounts were also multiply imputed for people who chose a bracket. People who lived alone were not asked this question. Instead, their personal income was assumed to equal the household income.	M4, M4a
Unsecured debt of $5,000 or more	Respondents were asked about debt other than student debt and secured debt (such as mortgages or title loans). Debts in the name of spouse or partner were included.	M8
Parental student debt	Respondents were asked about borrowing by parents on behalf of the student to go to school since randomization. For those who had difficulty answering the question about the exact amount, a categorical response option was offered. These were then imputed to continuous levels.	M7, M7a
Didn't experience food insecurity	Respondents were asked about adequacy of household food over prior six months. The possible responses were: 1=Enough of the kinds of food you want 2=Enough but not always the kinds of food you want 3=Sometimes not enough to eat 4=Often not enough to eat Response of 1 or 2 counts as not having experienced food insecurity.	M10
Individual receipt of SNAP	Respondents were asked about receipt in the prior month.	M1b
Individual receipt of Medicaid	Respondents were asked about receipt in the prior month.	M1e

Key: FRPL = free or reduced-price lunch. LIHEAP = Low Income Home Energy Assistance Program. SNAP = Supplemental Nutrition Assistance Program. TANF = Temporary Assistance for Needy Families. WIC = Special Supplemental Nutrition Program for Women, Infants, and Children.

Exhibit B-5: Details on Specifications for Survey-based Parental Engagement and Child Outcomes in Chapter 5

Outcome	Details on Derivation of Outcome	Follow-Up Survey Question(s)
Exploratory Outcomes		
Children of All Ages		
Parent believes child will graduate college	Parent asked how far child will go in school. Outcome equals 1 if parent reports child will finish college or if parent reports child will earn advanced degree after college; 0 otherwise.	P1
Highly engaged parent	This is a new scale developed for the three-year evaluations of PACE and HPOG 1.0. It was based on imputed average hours of time per day spent with the child in the typical week. The algorithm was different for preschoolers versus school-age children. Both thresholds were set at the 75th percentile for all children in the pooled evaluation samples for PACE and HPOG 1.0.[a] For preschoolers, parents were credited with 1 hour for each shared breakfast in the typical week; 1 hour for each shared dinner; 7 hours if they usually put the child to bed; and 1.5 hours if they read to the child once or twice a week, 4.5 hours if they read to the child three to six times a week, and 7 hours if they read to the child every day. These hours were summed and then divided by 7. The maximum value was 4 and the 75th percentile was 3.64. If the quotient was greater than this percentile, the parent was said to be highly engaged with the preschooler. For school-age children, parents were credited with 1 hour for each shared breakfast in the typical week, 1 hour for each shared dinner, 7 hours if they usually put the child to bed, 7 hours if they were usually present before the child leaves for school, 7 hours if they were usually present after the child comes home from school, 7 hours if they were usually present after dinner, and 7 hours if they were present with the child during the weekend. These hours were summed and then divided by 7. The maximum value was 8 and the 75th percentile was 7.28. If the quotient was greater than this percentile, the parent was said to be highly engaged with the school-age child.	O3a, O4a, O5a, O6a, O7a, O7b, O7c, P3, P6
Parent self-efficacy for helping child navigate school	Existing scale.[b] The seven-item scale captures parents' beliefs about their capability to help their child succeed in school. Response categories ranged from 1=disagree very strongly to 6=agree very strongly.	P9
Children Grades K-12		
Child repeated any grade	Yes/no question if child repeated any grade(s) in school.	Q10
Days child late for school last month	How many days child was late for school in last month (if in summer vacation, asked about last month child was enrolled in school).	Q12
Days child absent from school last month	How many days child was absent from school in last month (if in summer vacation, asked about last month child was enrolled in school).	Q11

[a] ACF's Health Profession Opportunity Grants (HPOG) Program, like the PACE project, provides training to low-income adults, but specifically for healthcare occupations. A first round of grants was awarded in 2010 (HPOG 1.0). Three of the nine programs studied in PACE were HPOG 1.0 grantees. For more: https://www.acf.hhs.gov/ofa/programs/hpog.

[b] Walker et al. (2005).

B.2 Imputation in the Three-Year Survey

As in any survey, some respondents did not answer every question. We employed a variety of approaches to allow us to use these cases despite their partial responses. Our approach varied across questions, depending on whether the question was embedded in a sequence of questions in which all questions needed to be answered to calculate the value of a scale, whether the question was embedded in a block of unanswered questions, and the frequency of nonresponse to the question.

The default rule was to exclude persons who failed to answer a question from any analysis involving that question, but to include them for all other analyses. Where this rule would result in a sharp drop in sample size—either for the question by itself or for a scale involving the question—then we imputed responses for people who failed to answer the question. Additionally, we imputed blocks of responses for two groups of people: those with large blocks of missing data and those who appear, based on administrative data, to have failed to report one or more education spells.

The goals of imputation were both variance and bias reduction.[19] Both goals are achievable with the rich set of parallel outcomes measured in the three-year survey. For example, indications of problems paying bills is valuable information for imputing missing income. We made decisions for all PACE sites on a global basis. Either we implemented an imputation procedure for a question in all nine PACE sites, or we left the question blank in all sites. Specifically, we imputed seven types of missing data:

(1) Number of college credits;

(2) Credential award dates;

(3) Income (personal and household);

(4) Early certifications and licenses (first 18 months after randomization);

(5) Skipouts (i.e., missing data on spells caused by trying to avoid respondents ending the survey);

(6) Spell start and end dates (job spells and school spells); and

(7) Survey data on school spells reported to the National Student Clearinghouse (NSC) but not by respondent.

This section briefly describes each of these imputations and their prevalence. We used a common methodology for the first four types of missing data. Section B.2.1 provides the detail

[19] Systematic nonresponse (e.g. those without college credentials are less likely to answer questions about credential attainment) can cause biased estimates. Effective imputation can reduce this bias. Making use of more data also increases sample size, thereby reducing the variance of impact estimates.

on these imputations. Section B.2.2 gives details on the imputation methodology for the other three types of missing data.

Types and Rates of Imputation. Exhibit B-6 below lists the seven types of imputation and shows the imputation rates for the survey respondents in the evaluation sample for WTA Connect. The instrument asked about credits spell by spell. It was fairly common for respondents to be unable to recall the number of credits they had earned during one or more training spells. They also had trouble recalling the dates on which they received credentials. Income was also frequently missing. The instrument prompted respondents to give a categorical answer ("bracketing") if they could not give an exact figure.

Exhibit B-6: Imputation Rates among Survey Respondents in Workforce Training Academy Connect

Type of Imputation	Job Spells (%)	School Spells (%)	Credentials (%)	People (%)
1. Number of college credits	n/a	n/a	n/a	10.2
2. Credential award dates	n/a	n/a	8.5	n/a
3. Income				
Personal (categorical)	n/a	n/a	n/a	7.4
Personal (exact)	n/a	n/a	n/a	12.9
Household (categorical)	n/a	n/a	n/a	14.2
Household (exact)	n/a	n/a	n/a	32.8
4. Early certifications and licenses	n/a	n/a	n/a	11.0
5. Skipouts	3.2	3.1	4.9	3.0
6. Spell start and/or end dates (job, school)	5.2	11.3	n/a	n/a
7. Survey data on school spells reported to NSC but not by respondent	n/a	8.6	21.2	5.7

Source: PACE three-year follow-up survey.

Note: Exact income was missing more often than categorical income because respondents unable or unwilling to provide an exact amount were encouraged to report a bracketed amount. n/a indicates not applicable.

The "Early Certifications and Licenses" row refers to our decision to impute this outcome for the 18-month follow-up survey for those study participants who were not interviewed at 18 months after randomization but who were interviewed at three years. This imputation involved creating a composite scale using the 18-month interview to measure receipt in the first 18 months and the second interview to measure receipt in the second 18 months. Section B.4 provides information about the rationale for this composite scale.

The "Skipouts" row refers to block missingness in the Integrated Training and Employment History module. The German survey upon which this module was modeled experienced a high level of breakoff (12 percent), meaning people discontinued the interview midstream and declined to restart it. To prevent similar problems for this three-year analysis, the PACE survey added a skipout feature in the module. If a person refused to answer any question in the module or gave a response of "don't know" to any of several critical flow-controlling questions in the module, the interview flow automatically skipped ahead to the next, less burdensome module (e.g., on 21st century skills, family structure, income and material well-being, or child outcomes).[20] With this approach, complete interview breakoffs were nearly eliminated, but a large block of missing data was created for about 7 percent of respondents (across the entire PACE three-year sample) and 3.0 percent of WTA Connect treatment and control group respondents specifically—much lower than the breakoff rate on the German study, but still high enough to require special attention.

Nonresponse was non-negligible for start and end dates of both job and school spells, particularly start dates. This is not surprising given that the reference period was up to three years long (and longer for people interviewed later in the survey period and for spells that started prior to randomization).

The final row of Exhibit B-6 refers to an adjustment for undercoverage of NSC-reported spells. This adjustment started with a match of survey reports with administrative data on college attendance from NSC. We flagged respondents who had spells of college attendance according to NSC but who did not themselves report any training (college or other type of school) since randomization. Although NSC is not error free, its enrollment coverage is generally high (see Appendix C). Accordingly, we imputed all the data from the matched NSC spells to survey respondents who did not report such spells.

B.2.1 College Credits, Credential Award Dates, Income, and Early Certifications and Licenses (Imputations 1-4)

As mentioned above, four of the seven types of imputation used a common imputation procedure: college credits, credential award dates, income, and certifications and licenses in the first 18 months. This section discusses the core procedures used and provides additional details for each of the four types of missing data.

Core Imputation Procedure. The core imputation methodology involved a number of steps. The first step entailed assembling a list of potential predictors and imputing any missing data in them.[21] The list of potential predictors included program, treatment status, the interaction of program with treatment status, baseline variables, parallel outcomes, and two-way and three-way interactions of both baseline variables and parallel outcomes with program and treatment status.

The second step entailed the use of a cross-validated LASSO procedure to fit a linear model for the target variable in terms of the assembled predictor list.[22] We did this on a pooled dataset that contained respondents from all nine PACE sites (*n*=6,773, of whom 5,910 responded to both follow-up surveys) and sometimes respondents from Health Profession Opportunity Grants

20 The original intent was not to skip past questions about credential attainment and current job conditions, but a mistake in the specifications caused these sections to also be skipped.

21 The only purpose of the imputation of potential predictors was to facilitate automated variable selection in the next step. After we used these imputed values of the predictors to predict new exam-based certifications and licenses as of the time of the 18-month survey, we discarded them. We carried out this imputation with SAS/MI/FCS.

22 See Appendix A.3 for details on the cross-validated LASSO.

(HPOG)-only programs, as well.[23] Note that though this procedure allowed program, treatment, their interaction with each other, and their interactions with many other variables to enter the model, it did not force any of them in. We discuss the implications of this decision after first finishing a description of the procedure.

The third step used predicted values from the final linear model to create a nested set of three partitions for each combination of site and treatment status.[24] The finest partition involved splitting the sample into 20 equal-sized groups based on the predicted probability of having reported an exam-based certificate or license if respondents had been interviewed at 18 months. The middle partition corresponded to deciles of this same probability, and the coarsest partition corresponded to quintiles of this same probability.

The fourth and final step used the hotdeck imputation procedure in SUDAAN to randomly match each nonrespondent with a respondent within cells defined by PACE program, PACE treatment status, and the nested partitions. Most cases were matched within cells defined by the 20-level partition. If there were no matches within those cells, then the procedure sought matches within the coarser partitions, first with the 10-level version and then with the five-level version if necessary. If even that did not permit a match, then the procedure randomly matched any unmatched nonrespondents with any respondent in the same PACE program with the same treatment status.

We ran the final hotdeck procedure five times with different random seeds to produce multiple imputations. We used these multiple imputations in the formal analysis runs to add between-imputation onto the naïve variance estimates on the full sample, using Rubin's classic formula.[25]

We now return to the implications of our decision not to force the interactions of site and treatment group with every other variable in the model. First, it is critical to note that we constrained matches to be from the same site and treatment group. This provided strong protection against imputation-caused bias in the estimated treatment impact. We used the models from the pooled dataset only to guide the matching of respondents and nonrespondents *with the same treatment status in the same site*. One way to think of this is that we used the pooled dataset to define a distance metric that we then applied within site and treatment group. An alternative procedure would have been to just randomly match respondents and nonrespondents within cells defined by site and treatment group. The point of using a distance metric rather than randomly matching is to reduce variance and the possibility of nonresponse bias. For a site with a large sample size, forcing in all the interactions of site and treatment

---

23 ACF's Health Profession Opportunity Grants (HPOG) Program, like PACE, provides training to low-income individuals, but only for healthcare occupations. The impact study of 32 first-round HPOG awardees (HPOG 1.0) included three awardees and one subgrantee (Carreras en Salud) also studied in PACE. For more: https://www.acf.hhs.gov/ofa/programs/hpog.

24 A "partition" of a sample is an exhaustive and mutually exclusive collection of subsets of the sample.

25 See for example, Rubin (1987).

group with other variables might not cause much deterioration in model quality, but in small sites forcing would almost certainly have made it more difficult to detect subtle main effects.[26]

Life Trajectory Clusters. The survey contained multiple measures of financial and social-emotional well-being. We theorized that these variables would be useful predictors of several types of missing data, particularly the missing data created by skipouts because none of these questions was involved in the bad skip pattern. However, interpretation of high-dimensional models is difficult. As a way of incorporating these rich data on well-being into imputation models while still keeping the models fairly easy to interpret, we condensed all these measures into a partition of the sample using cluster analysis. We were able to identify five clusters of respondents who vary clearly in terms of quality of life and core self-evaluation and family dependence. For shorthand, we refer to them as “life trajectory” clusters because one of the variables that they vary on most clearly is a sense of career progress. The five are:

- “Overextended”—above average income but also above average financial stress and low scores on psycho-social skills.
- “Family supported”—below average income but strong family supports that protect them from financial stress.
- “Strivers”—strong psycho-social skills and sense of career progress but low income (personal and household) and dependent on public support.
- “Down and out”—very low psycho-social skills, low sense of career progress, severe life challenges, low income (personal and household), and strong reliance on public support.
- “Winners”—strong psycho-social skills and sense of career progress, high income (personal and household), few bill problems, and little dependence on either family or public support.

Missing College Credits

For missing credits, we assembled a rich set of predictors from the baseline forms (the PACE Basic Information Form and the Self-Administered Questionnaire), NSC, the 18-month follow-up survey, person-level scales in the three-year survey, and spell-level data from the School Experiences module of the three-year survey. This was a spell-level file pooling data across the nine PACE sites, but not HPOG-only sites as no NSC data were available for the HPOG-only sample. We also added a large number of two- and three-way interactions with site and treatment group. After creating dummy variables for categorical variables, the total number of potential predictors was 1,584. The LASSO procedure working on this predictor set selected just six variables, yielding a model with an R-squared of 27 percent. Four of the six variables were significant predictors with standardized regression coefficients of at least 0.01. They were:

- Adjusted spell duration (adjusted for the longest break);
- Spell duration interacted with full/part-time student status;

[26] Algorithmically, the way to force in all interactions is to run the LASSO on a dataset restricted to just the cases in a particular site and treatment group. Even for the largest PACE site, this would not have provided nearly as much power to detect subtle main effects.

- Credits reported at 18 months; and
- NSC-reported full-time-equivalent months of enrollment through 35 months after randomization.

After controlling on the six factors, program and treatment were not important and nor were any of their interactions with each other or with other predictors. After imputing credits at the spell level, we summed to the person level for respondents with multiple school spells.

Missing Credential Award Dates

On the pooled PACE/HPOG credential sample, we modeled the lag between randomization and credential award date for those respondents with reported award dates (*n*=12,392, with 11,628 responses). The potential predictor list included site, treatment, the interaction of site with treatment, type of credential (10 categories), life trajectory cluster, 20 parallel outcomes at the person level, the lag between randomization and interview, 16 baseline variables, and a large set of two- and three-way interactions with site and treatment group. After creating dummy variables for categorical variables, the total number of potential predictors was 1,160. The LASSO procedure working on this predictor set selected 14 variables, yielding a model with an *R*-squared of 8.4 percent. The significant predictors with standardized regression coefficients of at least 0.01 were:

- HPOG versus PACE;
- Credential was awarded for regular college classes and typically takes less than a year to earn;
- Credential is an associate degree;
- Credential is a bachelor's degree;
- Self-assessed career progress;
- Student debt;
- Two interactions of HPOG with main effects;
- One interaction of treatment status with a main effect; and
- Two 3-way interactions of HPOG status with treatment status with main effects.

After matching nonrespondents with respondents, we adjusted for the difference in randomization dates between the two people, by adding the lag from the respondent to the randomization date for the nonrespondent. If this was past the interview date for the nonrespondent, we truncated the award date to equal the interview date.

Missing Income

The instrument yielded four related measures of income in the past month: (1) exact personal income; (2) categorical personal income; (3) exact household income; and (4) categorical household income. As could be seen in Exhibit B-6 above, missing data rates were considerably higher for the continuous variables than the categorical variables. This is because categorical income is only missing if both exact (which can be put in the appropriate income category) and categorical income are missing. For prediction purposes, we assembled a person-level file with program, treatment status, the interaction of program with treatment status, self-reported earnings by quarter, 10 variables about economic well-being, four variables about psycho-social skills, nine measures of educational progress, 12 baseline characteristics, and a large collection of two- and three-way interactions with site and treatment group. We used this list for modeling both personal and household income. We ran the LASSO on the pooled PACE/HPOG three-year dataset (n=14,467, with 12,782 exact personal income reports and 9,219 exact household income reports). After creating dummy variables for categorical variables, the total number of potential predictors was 1,414.

The LASSO procedure working on this predictor set selected 11 variables for personal income, yielding a model with an R-squared of 58 percent. The significant predictors with standardized regression coefficients of at least 0.01 were:

- Dummy variables for three of the five life trajectory clusters;
- Personal earnings for the 12th quarter after random assignment;
- A dummy variable for having earned an associate degree since randomization;
- A scale for being able to make ends meet at the end of the month; and
- An interaction of earnings with a dummy for receipt of any means-tested public benefits.

For household income, the LASSO procedure selected 26 variables, yielding a model with an R-squared of 52 percent. The significant predictors with standardized regression coefficients of at least 0.01 were:

- Dummy variables for three of the five life trajectory clusters;
- Personal earnings for the 12th quarter after random assignment;
- A dummy variable for being an Earned Income Tax Credit claimant;
- A dummy variable for living with a spouse;
- A dummy variable for living with parents;
- A dummy variable for living alone;
- Annual baseline family income below $15,000;
- Baseline SNAP (Supplemental Nutrition Assistance Program) or WIC (Special Supplemental Nutrition Program for Women, Infants, and Children) receipt;
- A dummy variable for having earned an associate degree since randomization;
- A scale for being able to make ends meet at the end of the month;
- An interaction of earnings with a dummy for receipt of any means-tested public benefits;
- An interaction of personal earnings with living arrangements; and
- Three 2- and 3-way interactions involving program.

Note that neither the model for personal income nor the model for household income involves three-way interactions of program with treatment status that are both statistically significant and substantively large. This does not mean that there are no program effects on income. Rather, it means that the measured parallel outcomes already capture whatever program effects might be present.

Certifications and Licenses in the First 18 Months

As mentioned earlier and as is discussed in detail in Section B.4 below, measures of ever-receipt of certifications and licenses blended reports from the 18-month and three-year surveys. This decision also required imputing what nonrespondents[27] to the 18-month survey would have reported if they had responded at that time. We used the core imputation described above for this imputation.

On the pooled PACE three-year survey respondent sample (*n*=6,773, of whom 5,906 responded to both the 18-month and three-year follow-up surveys and 867 responded to only the three-year survey), we modeled the receipt of such credentials among those who responded to the 18-month follow-up. The potential predictor list included program, treatment status, the interaction of program with treatment status, and about 40 baseline and three-year follow-up variables. After creating dummy variables for levels of categorical variables, this led to 80 potential predictors in total.

The LASSO selected 10 of the 80 predictors, yielding a model with an *R*-squared of 12.0 percent, a high value for a binary outcome. The selected variables included treatment status, dummy variables for two programs, one treatment-by-program interaction, five measures of educational progress and well-being at three years, and a dummy variable for employment in healthcare at three years. Of these, the predictors with standardized coefficients of at least 0.01 were:

- Treatment status;
- One dummy variable for site;
- One treatment-by-site interaction;
- Number of licenses obtained at three years;
- Report of a short-term college credential at three years;
- Report of a long-term college credential at three years; and
- Current employment in healthcare.

After imputing new exam-based certifications and licenses for 18-month survey nonrespondents, used the donor's interview date to we separate exam-based certifications and licenses reported in the three-year survey into two categories—early (would have been reported by the nonrespondent in the 18-month survey if the interview had taken place) versus late (would have been earned after the 18-month survey). We then created a blended flag for having earned an exam-based certification or license as of the three-year survey. The flag was set to

[27] Nonrespondents here were people who could not be located, refused to be interviewed, or were otherwise unavailable for an interview. The concept does not include people who skipped questions about credentials when interviewed at 18 months. We assumed that these respondents did not earn any credentials by the time of the 18-month interview.

yes if the 18-month nonrespondent had an imputed early exam-based certification or license or had reported a late exam-based certification or license in the three-year survey.

B.2.2 Skipout, Start and End Dates, and Unreported School Spells

The remaining three types of missing data required more customized procedures. This section provides details on our approach to each type.

Skipout

We considered several approaches to this type of missing data. One option we considered and rejected was to treat respondents with skipouts as nonrespondents and give them nonresponse-adjusted weights of zero. This simple option would have significantly boosted the overall nonresponse rate and wasted information collected after the skipout. A second rejected approach would have been to treat respondents with skipouts as nonrespondents only for analyses involving educational progress and employment. This option would have required the creation of a second set of nonresponse-adjusted weights and would have led to inconsistencies across analyses. A third rejected option was to impute each outcome and scale requiring any data from the Integrated Training and Employment History module. This option was more attractive but would not have supported estimation of career trajectories.

The approach we adopted was to use a block imputation approach that was initially used in medical expenditure surveys in the United States (Williams and Folsom 1981). The general method involves matching a nonrespondent to a respondent and then copying the entire block of missing data from the respondent to the nonrespondent. Our objective was to find a respondent whose training and employment history would align well with the nonrespondent's baseline characteristics and measures of well-being at three years. If the matched person had a missing response to a question within the Training and Employment History module, we copied this missing value over the skipout along with all the other variables.

We used sequential hotdecks as in the core imputation methodology, but we formed the partitions in a different manner. Rather than modeling a single variable and then forming a nested set of partitions based on model-based predictions of that single variable, we crossed the life trajectory clusters discussed above with other important measures. We used a sequence of four hotdecks, where the first had the most stringent criteria for matches, and each succeeding hotdeck had loosened criteria.

The first hotdeck matched nonrespondents to respondents within cells defined by program, treatment status, any schooling reported prior to skipout, any work reported prior to skipout, life trajectory cluster, and lag between randomization and interview in whole months. This was on the pooled PACE/HPOG sample (*n*=14,169, with 13,245 respondents who did not skip out).[28] This run found donors for 815 of the 924 skipouts on the pooled dataset.

The second hotdeck replaced program with site. This run found donors for 86 of the remaining 109 skipouts on the pooled dataset. The third hotdeck replaced the exact number of months in

[28] This excludes 302 three-year survey respondents that reported no training or employment between randomization and the survey interview.

the reference period with a dummy variable for whether the number was greater than 38 months. This run found donors for 22 of the remaining 23 skipouts on the pooled dataset. The fourth hotdeck used a collapsed version of self-assessed goal progress in place of life trajectory cluster and the binary recode of length of the reference period. This found a donor for the last remaining skipout.

Given the challenges in matching many of the nonrespondents to appropriate respondents, we did not carry out multiple imputation for skipouts. For the imputation of skipouts, our judgment was that the donor pools would be frequently small and that multiple random matches would, in fact, be the same match over and over. This lack of variation in the matched donors would have rendered variance estimates based on multiple imputations little better than variance estimates based on single imputation.

Because respondents with skipouts were missing a long stretch of data that are important to most of the secondary outcomes in this report, we prepared impact estimates with and without these cases, as displayed in Exhibit B-7 below. The two sets of impact estimates are very similar. The imputation allowed us to use as many as 21 more cases for WTA Connect (about 3 percent of the survey respondent sample), with the exact count depending on item nonresponse.

Imputation shifted the impact of the program most on self-reported employment at survey follow-up, but not enough to change the statistical significance of the impact.

Exhibit B-7: Comparison of Selected Impact Estimates of Workforce Training Academy Connect

Outcome and Sample	Impact Estimate	Standard Error	Sample Size	*p*-Value
Employed at Survey Follow-Up (%)				
Full sample	0.7	3.7	698	.429
Omitting skipouts	1.6	3.7	677	.332
Employed at $13 Per Hour or Above (%)				
Full sample	−1.2	3.2	686	.644
Omitting skipouts	−0.7	3.2	665	.590
Employed in a Job Requiring at Least Mid-Level Skills (%)				
Full sample	2.3	2.3	680	.162
Omitting skipouts	2.6	2.4	659	.136
Full-Time-Equivalent Months Enrolled in Any Type of School (months)				
Full sample	−0.1	0.5	697	.574
Omitting skipouts	0.0	0.5	676	.480
Receipt of an Exam-Based Credential (%) (blended 18-month and three-year surveys)				
Full sample	5.2**	2.9	698	.036
Omitting skipouts	6.7**	2.9	677	.011

Source: PACE 18-month and three-year follow-up surveys.

Note: "Full sample" rows include values imputed for skipouts. All estimates are regression-adjusted as discussed in Appendix Section A.3. Statistical significance levels, based on one-tailed *t*-tests tests of positive differences between research groups for positive outcomes and negative differences for negative outcomes (such as student debt), are summarized as follows: *** 1 percent level; ** 5 percent level; * 10 percent level.

Spell Start and End Dates

As mentioned earlier, respondents were frequently unable to remember dates. We decided to impute them to make the most use of the partial information in each respondent's reported history. Our primary objective was to create a high-quality measure of the duration of study over the entire reference period. Secondary objectives included the ability to estimate quarterly earnings over the entire reference period and supporting a broader set of exploratory analyses of career trajectories (transitions between school, work, and other activities).

For this imputation, we used a different approach from any of those discussed above. This decision was motivated by the complexity of partial information in the Training and Employment History module. Across the pooled PACE/HPOG sample, respondents had as many as six school spells and as many as 11 job spells. Even when respondents could not remember dates, we had many bounding conditions (e.g., spell #4 started after spell #3 ended). We devised a method that would respect these bounding conditions to create a coherent history while also supporting high-quality estimates of the site-specific impact of treatment on duration of study and quarterly earnings.

Before explaining the method, it will be useful to understand the bounding conditions:

- For every spell, we knew whether it ended before the three-year follow-up interview or was ongoing at that time.
- For all closed spells, we knew whether there was another spell that started after it but prior to the three-year interview.
- For most spells, we knew
  - whether it started before or after randomization;
  - whether it started in the middle of another spell or after some period during which the person was neither working for pay nor enrolled in school; and
  - whether a new spell started during it.
- For spells that followed other spells, we would most often know the end date of the prior spell.
- For spells that preceded other spells, we would most often know the start date of the succeeding spell.
- For spells that started during other spells, we would most often know the start and end dates of the "mother" spell.
- For spells that spanned the start of a new spell, we would most often know the start and end dates of the "daughter" spell.

Our general approach to imputing missing dates involved the following steps on the pooled PACE/HPOG sample:

(1) Express the date as a lag to some benchmark date. Specifically, we expressed start dates of main spells (those that did not start in the middle of any other spell) as the lag between randomization and the start of the spell, start dates of daughter spells as the lag from the

start of the mother spell to the start of the daughter spell, and end dates of all spells as the lag from spell start date to spell end date.

(2) Construct a statistical model for lag, and extract the predicted lag for spells with both known and unknown dates. (More details on this modeling process follow below. We constructed nine separate models.)

(3) Identify the nearest neighbor case in the pooled dataset in terms of the predicted lag. Copy the lag from the spell with the known relevant date (start or end) to the case with an unknown value for the relevant date.

(4) Add the imputed lag onto the benchmark date for the spell with an unknown date to obtain a preliminary date.

(5) If the preliminary imputation violates any of the constraints, truncate it to just barely satisfy the constraints. For example, if preliminary imputation of an end date placed the end date past the date of follow-up interview but the respondent had reported that the spell ended before the interview, then we truncated the lag so that the job ended the month before the interview.

Before providing details on the nine models constructed in step 2, we offer some general observations about this methodology. We considered conducting this process separately for each PACE site. We rejected that approach because of the complexity of the boundary constraints on dates and the rarity of patterns for respondents with multiple spells. Instead, we focused on constructing high-quality models and then finding the best match available.

The pooled sample size consisted of 27,939 job spells plus 13,093 school spells. After discarding spells reported by skipouts and spells that ended prior to randomization, the total number of spells was 40,672. Among these spells, either the start date or the end date was missing for 3,302, or 8 percent. Missing start dates was the more common problem, with 538 spells missing just the end date and 2,764 missing just the start date or both dates. Missing dates were slightly more common for school spells than for job spells (10 percent versus 7 percent). Missing dates for closed spells were much more common than for open spells (10 percent versus 4 percent). For WTA Connect, the overall missing data rate for spell dates was slightly lower than for the rest of the PACE pooled sample (7 percent versus 8 percent).

Exhibit B-8 below lists the models we created for each of nine types of lag and some features of each, including average imputed values for the various lags. Main spell #1 was always the ongoing spell at the time of randomization for those respondents working or going to school at the point of randomization, and so always has a negative lag. Main spell #2 was always the first spell after randomization for those not working or going to school at the point of randomization. Other main spells always followed main spell #1 or #2. Given this structure, we prepared separate models for the start date of main spell #1, main spell #2, and all other main spells (lag types 1, 5, and 6 below), and we modeled other features associated with the first spell separately, as well (lag types 2, 3, and 4). For other lag types, we modeled on a pooled dataset combining main spells #2 and higher (lag types 7, 8, and 9) and their associated subspells.

Exhibit B-8: Date Imputation for Three-Year Impact Study (Pooled PACE/HPOG Sample)

Lag Type	Modeled Variable	*R*-Squared (%)	Tested Variables	Selected Variables	Sample Size	Missing Data Rate (%)	Average Lag/Duration: Reported (months)	Average Lag/Duration: Imputed (months)
1	Lag from randomization date to start of main spell #1 (always negative because spell #1 was activity at time of randomization)	15	1,071	18	8,994	9.7	−18.8	−18.6
2	Duration of main spell #1 (closed only)	79	3,625	3	7,377	7.3	25.9	28.0
3	Lag from start of main spell #1 to start of subspell	78	2,989	3	5,459	8.8	23.2	16.9
4	Duration of subspells of main spell #1 (closed only)	0	3,103	2	4,563	8.8	16.2	15.7
5	Lag from randomization date to start of main spell #2	7	1,089	2	3,863	7.0	6.7	6.7
6	Lag from randomization date to start of main spells #3 and higher	38	5,113	33	18,082	4.9	18.9	17.4
7	Duration of main spells #2 and higher (closed only)	16	4,760	23	13,509	5.4	8.3	8.3
8	Lag from start of main spells #2 and higher to start of subspell	43	4,105	11	4,270	6.3	6.0	4.2
9	Duration of subspells for main spells #2 and higher (closed only)	14	3,383	9	2,546	6.8	7.3	7.1

Source: National Directory of New Hires; National Student Clearinghouse; PACE and HPOG 1.0 three-year follow-up surveys.

Note: Sample pooled across HPOG 1.0 and all nine PACE sites. Sample also pooled across treatment and control samples. A "main spell" is a spell that did not start in the middle of another spell. A "subspell" is a spell that did start in the middle of another spell.

The set of variables allowed into each model varied across the nine lag types. Tested variables included program, randomized treatment group, the interaction of program with treatment group, elapsed time between randomization and follow-up interview (and its square), job/school status, next activity (work, school, or other), school control (three levels, nested within job/school status), school level (three levels, nested within job/school status), open/closed status, life trajectory cluster (five levels), self-assessed goal progress, baseline covariates, two- and three-way interactions of these variables with program and treatment status, and other variables.

Model fit as measured by *R*-squared varied substantially across models, ranging from 0 percent to 79 percent. The reasons for this variation are not clear to us. Average imputed values were generally quite similar to average reported months.

Undercoverage of NSC-Reported Spells

As noted previously, we decided to supplement the histories of survey respondents who reported no training since randomization with any spells recorded for them in NSC and then to impute the spell attributes collected in the survey beyond the simple start and end dates for the spells. Across the nine PACE sites, this edit changed the training history for 7 percent of the sample, switching them from a status of no training to some. In the WTA Connect sample, there were 40 such respondents, accounting for 6 percent of the sample. We added these NSC-reported spells to the three-year follow-up survey history for those respondents and imputed the missing survey outcomes, such as earned credits and credentials.

This imputation proceeded by matching these 40 respondents to other WTA Connect study participants and copying over the donors' outcomes. This matching was structured, not random. We constrained matches to be from the same treatment group and to have a similar predicted profile of four survey-reported spell-level variables:

- Received a diploma or certificate typically requiring less than a full year's worth of credits during the spell;
- Received a diploma or certificate typically requiring a year or more's worth of credits, but less than an associate degree during the spell;
- Received an associate degree or higher during the spell; and
- Total credits earned during the spell.

We formed linear models for each of these survey-reported spell-level outcomes in terms of baseline variables and NSC-reported spell- and person-level variables on enrollment and credential attainment. We fit these models on the pooled (treatment plus control) sample for the WTA Connect program. Given that the matching was not random, we did not conduct multiple imputation. We instead conducted single imputation and have ignored the impact on variances.

Exhibit B-9 compares estimated program impacts with and without the addition of NSC-reported spells for several outcomes in the educational progress domain. Though the addition of NSC-reported spells does affect the impact estimate for each outcome, this effect is minor and does not affect the statistical significance of any of the outcomes below.

Exhibit B-9: Comparison of Selected Impact Estimates of Workforce Training Academy Connect with and without Imputation of NSC-Inferred Unreported Spells

Outcome and Sample	Impact Estimate	Standard Error	Sample Size	*p*-Value
Full-Time-Equivalent Months Enrolled in Any Type of School (months)				
Full sample	−0.1	0.5	697	.574
Omitting NSC-only spells	0.2	0.5	697	.341
Full-Time-Equivalent Months Enrolled in a College (months)				
Full sample	0.0	0.5	698	.464
Omitting NSC-only spells	0.3	0.4	698	.230
Receipt of a College Credential Typically Requiring Less Than a Year of Credits (%)				
Full sample	6.8***	3.0	698	.012
Omitting NSC-only spells	6.0***	2.6	698	.012
Receipt of a College Credential Typically Requiring a Year or More of Credits (%)				
Full sample	1.0	1.2	698	.218
Omitting NSC-only spells	−0.3	1.0	698	.627

Source: National Study Clearinghouse; PACE three-year follow-up survey.

Note: All estimates are regression-adjusted as discussed in Appendix Section A.3.

Statistical significance levels, based on one-tailed *t*-tests tests of positive differences between research groups for positive outcomes and negative differences for negative outcomes (such as student debt), are summarized as follows: *** 1 percent level; ** 5 percent level; * 10 percent level.

B.3 Survey Nonresponse Analysis

As in any survey, nonresponse can lead to bias if nonresponse propensity is correlated with outcomes. In the context of a randomized experiment such as this evaluation of WTA Connect, concern about nonresponse is heightened if the nonresponse rate is different in the treatment group than in the control group. Nonresponse can lead to biased impact estimates even without differential nonresponse rates across study groups, but it is widely accepted that differential rates heighten concerns about biased impact estimates.[29]

The three-year follow-up survey for WTA Connect obtained very similar response rates in the treatment and control groups, with both groups having a 74 percent response rate. Such similarity suggests that any differences in baseline characteristics and outcomes for respondents versus the full sample will tend to be similar in size for the two groups. We studied this matter further using administrative data and found weak evidence of nonresponse bias (illustrations of these biases are presented in Exhibit B-11). We developed a set of nonresponse adjustment weights that appears to remove most of this bias. This section first presents the evidence of nonresponse bias in unadjusted impact estimates and then documents the nonresponse adjustment weights that we created to mitigate this bias.

29 See for example, Deke and Chiang (2017). For a slightly contrarian view, see Hendra and Hill (2018).

B.3.1 Evidence of Nonresponse Bias in Unadjusted Impact Estimates

We gauged the likelihood of nonresponse bias through two types of analysis, one involving baseline data and one involving post-randomization administrative data.

The first analysis takes baseline equivalence as an indication of the potential for bias. If randomization is correctly implemented, there should be no systematic differences between the treatment group and the control group. We directly tested that using complete data from the Basic Information Form (see Section A.2). This insight also provides a proxy for nonresponse bias and the ability of our weighting scheme to correct for it. In the absence of nonresponse bias, appropriately weighted tabulations of that data *among survey respondents* should also show baseline equivalence.

The second type of analysis looks directly at estimated impacts. We know who responded to the survey and we have administrative data outcomes for both survey respondents and nonrespondents. We can thus compute two impact estimates from the administrative data: one estimate from the unweighted full sample, which we treat as truth; and a second estimate from the weighted survey sample. In the absence of nonresponse bias (and with large enough samples), we should get the same (up to sampling variability) estimates of impact on the full sample and on the weighted sample of survey respondents. Theoretically, it is possible to test whether estimated differences between these two impact estimates are statistically significant, but we did not do this, relying instead on impressions of consistency across a collection of administratively measured outcomes.

Exhibit B-10 below considers baseline equivalence among survey respondents. In the first three columns reflecting all study participants, there are six characteristics where we see statistically significant differences between the treatment and control groups.[30] The next three columns, which report statistics for survey respondents only, show statistically significant differences for only four characteristics (two of the same characteristics as before and two different ones), so nonresponse slightly improved baseline imbalance. The last set of three columns shows that weighting further improved baseline balance, having reduced the number of significant imbalances to three.

30 Note that the numbers in the first three columns of Exhibit B-10 reflect baseline balance for the full sample following imputation of missing values, whereas Appendix Exhibit A.2 presented pre-imputation figures.

Exhibit B-10: Baseline Balance on Full Sample, Unweighted Respondent Sample, and Weighted Respondent Sample

Characteristic	Treatment (All)	Control (All)	*p*-Value	Treatment (Unweighted Survey)	Control (Unweighted Survey)	*p*-Value	Treatment (Weighted Survey)	Control (Weighted Survey)	*p*-Value
Age (%)			.236			.210			.310
20 or under	14.7	14.0		14.1	13.4		14.0	13.9	
21-24	17.2	15.6		17.8	14.9		17.3	15.3	
25-34	24.7	30.7		24.4	31.4		25.1	31.5	
35+	43.4	39.8		43.7	40.3		43.7	39.3	
Sex (%)			.123			.283			.331
Female	65.1	60.3		69.0	65.1		65.7	62.0	
Male	34.9	39.8		31.0	34.9		34.3	38.0	
Race/Ethnicity			.209			.320			.099
Hispanic, any race	13.8	17.3		15.5	16.9		15.1	16.3	
Black, non-Hispanic	50.0	44.6		50.9	43.1		51.9	43.3	
White, non-Hispanic	34.3	34.9		32.8	37.7		31.9	37.6	
Another race, non-Hispanic	5.7	7.8		5.2	6.9		4.9	7.2	
Family Structure (%)			.007			.004			.003
Not living with spouse/partner and not living with children	47.9	50.3		46.6	47.1		48.2	48.7	
Not living with spouse/partner but living with children	24.5	15.9		25.3	15.4		24.9	14.9	
Living with spouse/partner and not living with children	18.7	22.0		19.3	24.0		18.3	23.5	
Living with spouse/partner and children	8.9	11.8		8.9	13.4		8.5	13.0	
Living with parents (%)	17.2	16.7	.828	17.0	15.4	.585	16.4	15.8	.823
One parent has at least some college (%)	32.3	30.2	.486	34.5	29.7	.178	34.2	30.1	.248
High School Grades (%)			.394			.462			.122
Mostly A's	8.1	7.2		6.6	7.4		8.1	6.9	
Mostly B's	34.5	38.7		32.5	36.3		31.8	39.5	
Mostly C's or below	57.5	54.1		60.9	56.3		60.1	53.6	

Characteristic	Treatment (All)	Control (All)	*p*-Value	Treatment (Unweighted Survey)	Control (Unweighted Survey)	*p*-Value	Treatment (Weighted Survey)	Control (Weighted Survey)	*p*-Value
Current Education (%)			.060			.075			.168
Less than high school diploma	38.9	41.0		35.3	39.7		36.3	39.6	
High school diploma or equivalent	36.0	38.1		37.6	37.7		37.2	37.9	
Less than one year of college	13.2	8.7		14.4	8.9		14.1	8.9	
One or more years of college	9.2	7.2		9.5	7.7		9.4	8.4	
Associate degree or higher	3.0	5.3		3.5	6.0		3.3	5.3	
Received vocational or technical certificate or diploma (%)	20.6	20.7	.976	19.0	22.3	.279	18.7	20.6	.516
Career Knowledge Index (average of items)	0.37	0.35	.354	0.36	0.34	.395	0.37	0.35	.384
Psycho-Social Indices									
Academic Discipline Index	4.95	4.89	.213	4.94	4.88	.365	4.92	4.90	.693
Training Commitment Index	5.41	5.29	.011	5.38	5.26	.025	5.37	5.30	.195
Academic Self-Confidence Index	4.36	4.29	.205	4.32	4.27	.393	4.32	4.29	.569
Emotional Stability Index	4.88	4.83	.434	4.87	4.82	.419	4.86	4.83	.647
Social Support Index	3.13	3.07	.035	3.14	3.06	.015	3.14	3.06	.011
Stress Index	2.55	2.59	.507	2.54	2.58	.451	2.54	2.59	.362
Depression Index	1.81	1.83	.666	1.80	1.81	.867	1.81	1.81	.990
Income (%)			.147			.179			.314
Less than $15,000	57.2	54.6		55.5	52.0		58.2	54.5	
$15,000-29,999	26.6	25.0		27.9	25.4		26.9	24.3	
$30,000+	15.3	19.9		16.7	21.7		15.0	20.1	
Mean ($)	15,373	17,022	.104	16,218	17,879	.176	15,239	16,934	.152
Public Assistance / Hardship Past 12 Months (%)									
Received WIC or SNAP	68.9	62.4	.034	66.4	61.7	.200	67.7	62.7	.169
Received public assistance or welfare	15.7	13.7	.386	14.7	13.4	.641	15.4	13.3	.433
Reported financial hardship	62.3	62.8	.887	60.1	62.9	.448	61.8	63.5	.642

Characteristic	Treatment (All)	Control (All)	*p*-Value	Treatment (Unweighted Survey)	Control (Unweighted Survey)	*p*-Value	Treatment (Weighted Survey)	Control (Weighted Survey)	*p*-Value
Current Work Hours (%)			.590			.331			.440
0	62.1	63.0		61.2	64.3		62.5	64.8	
1-19	5.7	4.0		6.9	4.3		7.0	4.4	
20-34	12.6	14.0		12.9	10.6		12.0	10.5	
35+	19.8	18.8		19.5	20.9		18.8	20.4	
Expected Work Hours in Next Few Months (%)			.134			.101			.164
0	19.6	25.4		19.8	26.6		19.8	26.9	
1-19	4.9	3.6		5.2	4.0		5.3	4.0	
20-34	30.0	26.4		29.3	23.4		28.9	25.6	
35+	45.5	44.6		45.7	46.0		45.9	43.5	
Life Challenges Index (average in original units 1-5)	1.78	1.74	.297	1.77	1.71	.214	1.78	1.74	.421
Owns a car (%)	61.3	55.8	.089	63.5	58.9	.208	61.9	55.8	.104
Has both computer and internet at home (%)	48.9	50.5	.625	52.0	52.3	.942	48.4	50.8	.526
Ever arrested (%)	41.3	40.8	.883	38.5	38.3	.952	39.7	39.9	.949
Sample sizes	470	473		348	350		348	350	

Key: SNAP = Supplemental Nutrition Assistance Program. WIC = Special Supplemental Nutrition Program for Women, Infants, and Children.

Source: PACE Basic Information Form; PACE Self-Administered Questionnaire); response status to the PACE three-year follow-up survey. Sample restricted to those randomized for the evaluation of Workforce Training Academy Connect.

Note: SAS/SURVEYFREQ used to test for significant imbalances for categorical variables. SAS/TTEST used to test for significant imbalances for other variables. Weights are based on the dual raking system explained in Appendix Section B.3.2 below. Significant imbalances are highlighted in red, using a threshold for statistical significance of 10 percent.

Exhibit B-11 presents evidence about the level of nonresponse bias with and without adjustment weights. The first three panels of Exhibit B-11 compare three sets of regression-adjusted impacts on earnings outcomes from NDNH records (panels 1 and 2) and on college outcomes from NSC records (panel 3).[31] The first set of impact estimates (column 1) is based on the full sample. The second set of impact estimates (column 3) excludes survey nonrespondents. Differences between the first and second set of impacts signal nonresponse bias. The third set of impact estimates (column 5) also excludes survey nonrespondents but weights survey respondents with nonresponse adjustment weights, which are explained in Section B.3.2 below. If the weights are good, then the differences between the first and fifth columns will be smaller than those between the first and third columns.

Note that all three sets of impact estimates are regression-adjusted with the covariates discussed in Section A.3.

Exhibit B-11: Comparison of Selected Estimates of the Impact of Workforce Training Academy Connect for the Unweighted and Weighted Survey Samples

Outcome (Data Source)	Impact Estimate (All)	Std. Error	Impact Estimate (Unweighted Survey)	Std. Error	Impact Estimate (Weighted Survey)	Std. Error
Confirmatory Outcome (NDNH)						
Average quarterly earnings Q12-Q13 ($)	93	224	193	252	190	254
Exploratory Outcomes (NDNH)						
Q5 earnings ($)	−82	178	80	209	45	211
Q9 earnings ($)	164	213	369*	241	349*	248
Q13 earnings ($)	6	246	54	277	78	279
Q17 earnings ($)	−38	267	199	295	222	295
Any earnings Q5 (%)	8.8***	3.0	12.9***	3.4	12.8***	3.5
Any earnings Q9 (%)	5.4**	3.0	7.1**	3.4	7.2**	3.5
Any earnings Q13 (%)	5.2**	3.0	4.1	3.4	4.7*	3.4
Any earnings Q17 (%)	6.0**	3.1	8.6***	3.5	9.3***	3.5
Auxiliary Education Outcomes (NSC)						
Number of months of any enrollment through 35 months	0.3	0.3	0.6	0.4	0.3	0.4
Number of months of FTE enrollment through 35 months	0.1	0.2	0.2	0.2	0.0	0.3
Any enrollment through 35 months (%)	3.8	2.8	5.3	3.2	3.5	3.4
Any credentials through 35 months (%)	0.7	0.8	0.2	0.9	0.6	1.0
Number of months of FTE enrollment through September 2018	0.1	0.2	0.3	0.3	0.1	0.3
Any credentials through September 2018 (%)	0.8	0.9	0.5	1.0	1.1	1.1

31 NSC outcomes in this table are not formal outcomes for the evaluation of WTA Connect. We decided not to use them for the formal evaluation because the colleges attended by these students frequently are not reporting their credentials to NSC, as discussed in Section C.3. Nonetheless, these outcomes are observed for the full sample and thus are useful for assessing the contribution of the weights to inference.

Outcome (Data Source)	Impact Estimate (All)	Std. Error	Impact Estimate (Unweighted Survey)	Std. Error	Impact Estimate (Weighted Survey)	Std. Error
Secondary Employment Outcomes (Survey)						
Employed at survey follow-up (%)			0.8	3.6	0.7	3.7
Earning $13 per hour or more (%)			−0.7	3.2	−1.2	3.2
Working in a job requiring a least mid-level skills (%)			2.3	2.3	2.3	2.3
Secondary Education Outcomes (Survey)						
Full-time-equivalent months enrolled at any school (#)			−0.1	0.4	−0.1	0.5
Received an exam-based certification or license (%)			5.9**	2.8	5.2**	2.9
Other Secondary Outcomes (Survey)						
Indicators of Independence and Well-Being						
Has health insurance coverage (%)			−3.7	2.2	−3.6	2.2
Receipt of means-tested public benefits (%)			4.0	2.9	3.6	2.9
Personal student debt ($)			−457	384	−311	405
Any signs of financial distress (%)			−3.5	3.6	−3.6	3.7
Indices of Self-Assessed Career Progress (average)						
Confidence in career knowledge			0.08*	0.05	0.08*	0.05
Access to career supports			0.00	0.02	0.01	0.02
Sample size (treatment + control)	920		698		698	

Source: National Directory of New Hires; National Student Clearinghouse; PACE three-year follow-up survey, except “Received an exam-based certification or license” is a blended variable based on PACE 18-month and three-year follow-up surveys.

Note: All estimates are regression-adjusted as discussed in Appendix Section A.3. The Full Sample columns are blank for survey-measured outcomes as they are not available for the full sample.

Statistical significance levels for secondary outcomes are based on one-tailed *t*-tests tests of positive differences between research groups for positive outcomes and negative differences for negative outcomes (such as student debt). Statistical significance levels for exploratory outcomes are based on two-tailed *t*-tests of differences. They are summarized as follows: *** 1 percent level; ** 5 percent level; * 10 percent level.

We did not formally test the differences between the alternative estimates, but given that the survey respondents constitute a very large subset (74 percent) of all study participants, many of the differences would be statistically significant. For several follow-up administrative variables, there is weak evidence of bias in estimated impacts based on the unweighted respondent sample. Generally, that bias does not change substantive conclusions, but there is one case that crosses the threshold of statistical significance. The estimated impact on earnings in Q9 for the full sample is positive but not statistically significant ($164), whereas for the unweighted respondent sample the estimated impact is larger and significant ($369); the survey nonresponse weights reduce the impact estimate slightly ($349).

If we were evaluating WTA Connect as a stand-alone program (rather than one of nine in the PACE project), we might have decided not to use weights. However, there was strong positive bias in the estimated program impacts on earnings and educational progress at another PACE site (Judkins et al. 2020) that led us to conclude that current earnings and educational progress are related to nonresponse propensity in different ways on the treatment and control groups at that site. Given the centrality of earnings and educational progress in the logic models for how PACE programs would affect a wide variety of life outcomes measured in the survey, this relationship clearly implies some survey nonresponse adjustment was required for that site. Out of an abundance of caution, we then applied nonresponse adjustment at all sites.

The final pair of columns shows that the nonresponse weights generally bring impact estimates based only on survey respondents back into good alignment with impact estimates on the full study sample. For example, the impact on earnings in Q5 for the full sample is −$82. The estimated impact for the weighted survey sample is +$45, which is closer to the full sample estimate than is the unweighted estimate (+$80). Neither of these impact estimates is statistically significant. This illustrates how the nonresponse weights removed much of the bias in the unweighted survey sample. The weighted impacts do not agree exactly with the full-sample impacts, but that would be an unreasonable goal for an adjustment procedure. Altogether, the weights reduced nonresponse bias for about half of the NDNH outcomes shown in Exhibit B-11. Furthermore, as shown in that exhibit, the weights strongly reduced bias in estimated impacts on NSC-reported college enrollment.

We implemented this solution across all nine PACE sites. Nonresponse bias was modest at WTA Connect; still, the procedure appears to do no harm even when not strictly required.

For the survey-based outcomes, the fourth, fifth, and sixth panels of Exhibit B-11 compare the unweighted and weighted impact estimates. There are only minor differences between the estimates. This is consistent with the NDNH and NSC outcome findings, which found few differences at the WTA Connect site between the full sample, unweighted, and weighted impact estimates.

B.3.2 Construction of Nonresponse Adjustment Weights

Construction of weights to reduce the biases just discussed was more complex than anticipated. At first, we tried a standard propensity scoring approach,[32] as was used in the short-term report on WTA Connect (see Hamadyk and Zeidenberg 2018). However, that approach was not successful in removing the biases in estimated impacts based on administrative data for survey respondents at that other PACE site. Data storage arrangements posed a further challenge in developing a set of nonresponse adjustment weights. Contractual arrangements permitted the merging of survey data with either NDNH data or NSC data, but they did not permit the merging of NDNH and NSC data. In response to this challenge, we developed a new approach that we call dual-system raking.

“Raking” is the name for iterative procedures that create weights for a sample in such a manner that marginal tabulations of the sample agree exactly with pre-specified “control” totals in multiple dimensions. For example, raking can be used to create weights that will cause

32 In the standard approach, a logistic model for response status is fit in terms of universally available covariates (baseline and administrative). The model is used to generate a predicted response propensity for each person (respondent and nonrespondent), then people are sorted on this prediction into strata. The empirical response rate is calculated for each stratum, and finally the inverse of this rate is applied to respondents as a nonresponse-adjustment weight.

tabulations by sex, tabulations by race, and tabulations by age all to agree with pre-specified totals for sex, race, and age. In this example, sex, race, and age are dimensions.

In the context of nonresponse, if tabulations are prepared from the full sample and raking is used on the respondents, then weighted tabulations of the respondent sample will be in perfect agreement with parallel tabulations of the full sample. This exact multi-dimensional agreement is referred to as “hyperbalance.” In the context of an experiment, if this procedure is run separately for the treatment and control groups, then hyperbalance between respondents and nonrespondents means that the weighted balance between the treatment and control groups on the respondent sample should be just as good as on the full sample.

This hyperbalance by arm means that if we estimated treatment impact on just the respondent sample with these weights but without regression adjustment, the estimated program impact on each of these hyperbalanced variables would agree exactly with corresponding program impacts estimated on the full sample. The use of regression adjustment to estimate program impacts (rather than simple mean difference between arms) means that this agreement will not be exact, but agreement should still be very good for hyperbalanced variables. Theoretically, it should also improve agreement (between impact estimates based on the full sample and impact estimates based on just the respondent sample) for a variety of related parallel outcomes.

Key raking variables include both categorical variables (e.g., any NSC-reported enrollment) and interval-valued variables (e.g., number of months enrolled in college according to NSC records). Including these interval-valued variables seems particularly important because many educational outcomes are associated with the length of study.

The need to include continuous variables in the raking is challenging because traditional raking algorithms work only with categorical variables. In contrast, the generalized raking we propose and use here can handle a mix of categorical and continuous variables.[33] For categorical variables, the procedure guarantees perfect correspondence between the respondent sample and full sample by arm on the distribution of the sample across the categories of each variable; for continuous variables, the procedure induces perfect agreement on the marginal means of each of them.

The generalized raking procedure of Folsom and associates is available in the WTADJUST procedure of SUDAAN. A similar procedure that only works for categorical covariates is the SAS raking macro of Izrael, Hoaglin, and Battaglia (2000). It was necessary to use both software packages because the analyses had to be run on two servers, one that had SUDAAN installed (at Abt) and one that did not (at ACF). We refer to our system as “dual-system” raking because it permits raking both to NDNH information and to NSC information though the two types of data reside on two different systems.

---

33 Generalized raking is most fully developed by Folsom and Singh (2000), who in turn draw on methods originally proposed by Folsom (1991), Deville and Särndal (1992), and Folsom and Witt (1994). Dual raking is similar to the approach of Judkins et al. (2007) that involves the use of raking to construct weights in quasi-experimental designs.

The details of the dual-system raking procedure are as follows:

(1) We used SUDAAN/WTADJUST to develop survey weights on the Abt server that induced hyperbalance by arm for the means of four NSC variables. Two of these NSC variables were counts on months: months with any enrollment and months of full-time-equivalent enrollment. Two of the NSC variables were binary flags: any enrollment and any completions (credentials). All four of these variables were constrained to enrollment and completions within 35 months of randomization.

(2) We merged the weights from step 1 with baseline data and follow-up survey data on the Abt server. We then passed these merged data through to a secure ACF server, where third-party ACF contractors merged our data with NDNH earnings data, removing personal identifiers from the merged dataset. We had verified that this set of NSC-adjusted weights provides nearly unbiased impact estimates for survey-based education outcomes, but after merging the weights with NDNH data, we discovered that these NSC-adjusted weights did not remove bias in survey-based impact estimates for earnings outcomes.

(3) To remedy this, we used the Izrael-Hoaglin-Battaglia macro on the ACF server to rake the weights from step 1 in such a manner as to attain hyperbalance by arm on three categorized versions of NDNH earnings. Specifically, we obtained hyperbalance for a six-level categorization of earnings at Q12 and Q13, a five-level categorization of earnings at Q9, and a five-level categorization of cumulative earnings from Q1 through Q12.[34] We verified that these weights removed most of the nonresponse bias on estimates of program impacts on NDNH earnings at the other PACE site when estimated from nonrespondents instead of from the full sample. This sensitivity analysis included the continuous versions of the variables used in the raking, as well as continuous earnings at Q5 and Q17 and binary indicators for any employment at Q5, Q9, Q13, and Q17.

(4) We used the weights from step 3 on the ACF server to estimate (by arm) the distributions of survey-reported earnings. Specifically, we split Q12 earnings at \$0, \$6,000, and \$9,000; Q9 earnings at \$0, \$6,000, and \$9,000; and average quarterly earnings for Q1 through Q12 at \$3,000 and \$6,000. (The breaks for survey-reported earnings needed to be coarser than the breaks for NDNH earnings because of the smaller sample sizes in the respondent survey sample.)

(5) We again used the Izrael-Hoaglin-Battaglia macro on the ACF server to rake the weights from step 1, but for this step we used the control totals from step 4 rather than the NDNH totals used in step 3. We then verified that these weights removed most of the nonresponse bias on estimates of program impacts on NDNH earnings when estimated from nonrespondents instead of from the full sample at the other PACE site. These weights

[34] This process is also referred to as “binning.” We used more bins for the confirmatory outcome than for the exploratory outcomes. Reducing the number of bins generally speeds convergence and reduces the frequency of extreme adjustments.

did not perform as well as the weights from step 3 in reducing nonresponse bias on the respondent sample, but the deterioration (not shown) was not very large.

(6) We exported the 11 estimated totals from step 4 for each arm from the ACF server to the Abt server. (The data use agreement permitted the transfer of tabulations; only the export of microdata was prohibited.)

(7) We again used the Izrael-Hoaglin-Battaglia macro to rake the weights from step 1 to the control totals from step 4, but this time we did the raking on the Abt server rather than on the ACF server. We then merged these with NSC data on the Abt server and verified that these weights removed most of the nonresponse bias on estimates of program impacts on NSC outcomes when estimated from nonrespondents instead of from the full sample at the other PACE site.

B.4 Quality and Completeness of Exam-Based Credentials Reported in the Survey

Earlier analyses for another PACE site had identified a potential quality issue in reporting on receipt of exam-based credentials in the PACE three-year follow-up survey. Specifically, estimates of exam-based certifications and licenses for the San Diego Workforce Partnership's Bridge to Employment in the Healthcare Industry program at three years (see Judkins et al. 2020) were much lower than those based on the 18-month survey (see Farrell and Martinson 2017). This points to a clear problem, because the percentage of participants who ever received such credentials cannot diminish over time.

A review of the survey's skip patterns and wording identified three features in the design of the three-year instrument for the PACE project that might have led to fewer credentials of this type being reported than were reported in the 18-month survey:

- First, the three-year instrument allowed only respondents with some formal schooling since randomization to report exam-based certifications and licenses. However, people who learn skills on the job or through independent online study (such as YouTube tutorials) can sit for the exams for many certifications and licenses.
- Second, the wording for the three-year instrument strongly emphasized that "school-issued certificates" were not the same thing as "exam-based certifications and licenses." We had introduced this language to ease confusion about the difference between credentials issued by schools and credentials issued by other authorities. However, because some schools serve as proxy administrators of exams for credentials that are actually issued by other authorities, it is possible that this wording led some respondents to report exam-based credentials as school-based credentials or to not report them at all.
- The third feature is just the greater passage of time. Respondents may not have renewed exam-based certifications and licenses or they might have discovered that the credentials are less useful than anticipated, either of which could have reduced respondents' inclination to report older exam-based credentials.

Given this review, the PACE research team decided that the 18-month follow-up survey reporting on early exam-based credentials earned are probably more accurate than the three-year survey reporting. Accordingly, we decided to *combine* data on exam-based credentials for the two time periods for PACE three-year impact reports. This new composite measure of receipt of any exam-based credential since randomization was set to yes if the respondent reported receiving such a credential either in the 18-month survey or in the three-year survey at a time point after the date of the 18-month survey interview.

For the 15 percent of the study sample that did not respond to the 18-month survey, we imputed a response on receipt of exam-based credentials. When receipt dates were not reported in the three-year survey, we imputed them, as well. Both of these imputations were discussed in Section B.3.

B.5 Quality and Completeness of School-Issued Credentials Reported in the Survey

The problem with the reporting on exam-based credentials discovered at Bridge to Employment just discussed in Section B.4 raised a question: Had similar data problems occurred for *school-issued* credentials that would justify similarly combining data on receipt of those credentials from the two surveys?

For the Bridge to Employment report, the research team had decided to not to combine the three-year survey data with the 18-month survey data for *other* types of credentials (i.e., those that were *not* exam-based). Further, we decided the same for all other PACE three-year impact reports using survey data. This decision we based on analyses of data for yet another PACE site: Pima Community College (PCC)'s Pathways to Healthcare. We chose this site because the PCC study offered college records to support the analysis, making it a good choice for investigating these survey outcomes, and because the evaluation's processing of those records was further along at that time than the other PACE sites.

Analysis of PCC's records showed that for *school-issued* credentials, the three-year survey was more accurate than the 18-month survey. We focused on Pathways to Healthcare respondents who reported a school-issued credential in only one of the two surveys, checking the PCC records to see whether that survey-reported credential had actually been issued. Among respondents who reported a school-issued credential at 18 months but not at three years, PCC records confirmed just 35 percent of those reports. In contrast, among respondents who reported a school-issued credential at three years but not at 18 months, PCC records confirmed fully 81 percent. For some reason, the 18-month survey instrument generated many more unverifiable school-issued credential claims than the three-year survey did.

As a result, the research team decided that in all PACE sites where we used the three-year survey rather than college records to measure educational progress, we would rely on the three-year survey *uncombined* for data on school-issued credentials.

Appendix C: National Student Clearinghouse Data

The National Student Clearinghouse (NSC) is a national database of college enrollment records designed to aid the administration of student loan programs, but it can be a useful tool for education researchers. In this report, we used NSC records for imputation of missing data and to prepare alternate estimates of the impacts of Workforce Training Academy Connect (WTA Connect). Section C.1 summarizes statistics on NSC coverage. Section C.2 provides details on how raw data from NSC were recoded to make them more relevant to the evaluation of WTA Connect. Finally, Section C.3 presents estimates of WTA Connect impacts based on NSC data and contrasts them with the estimates presented in Chapter 3 of this report.

C.1 Coverage

Given the focus on loan administration, NSC does not cover schools that are not Title IV schools, the set of schools approved for federal student loans by the U.S. Department of Education. Moreover, although NSC does include a few schools that are not colleges in the sense used elsewhere in this report (i.e., issuing degrees), the vast majority of the schools are colleges. Exhibit C-1 shows the percentage of colleges providing records to NSC by year and by type of school. As shown, coverage of public two-year and four-year schools was more than 95 percent. Coverage was lower among private nonprofit four-year schools, considerably lower among private for-profit four-year schools, and very low for private two-year schools (both for-profit and nonprofit).

Exhibit C-1: NSC College-Level Cooperation Rates by College Control and Level from 2013 through 2016

Control and Level of College	2013 (%)	2014 (%)	2015 (%)	2016 (%)
Public, four-year	99.2	99.4	99.5	99.6
Private, nonprofit, four-year	93.6	95.2	95.8	96.1
Private, for-profit, four-year	74.4	79.9	81.7	81.0
Public, two-year	99.1	99.2	99.4	99.5
Private, nonprofit, two-year	39.5	40.8	40.4	42.1
Private, for-profit, two-year	19.7	28.1	26.7	26.6

Source: National Student Clearinghouse (https://nscresearchcenter.org/wp-content/uploads/NSC_COVERAGE.xlsx).

Analyses of NSC data in this report are limited to enrollment records obtained from 2000 forward. All study participants gave their informed consent to have NSC share their records with the PACE research team. The team negotiated a contract with NSC to match relevant NSC records to the study participants. The team sent both Social Security numbers and names to NSC to make the matching more accurate. The abstracted records were then sent by encrypted secure methods to the research team, who have used them under tight security conditions.

C.2 Data and Measures

Information on outcomes other than enrollment tends to be less reliable.[35] Notably, standards and practices governing credential reporting are inconsistent across schools. So our primary use of NSC data was to measure enrollment. Counting the quarter during which random assignment occurred as quarter 0, we obtained an abstract from NSC in October of 2018 covering enrollment through quarter 15 for all 943 WTA Connect study participants (470 in the treatment group and 473 in the control group).

Records from NSC are arranged in a spell format with starting and ending dates. We translated these first into a set of person-month-level records, reconciling multiple and conflicting spells as seemed most sensible. The team derived two variables for each person-month. The first was a simple binary indicator of “any enrollment” (yes/no). The second was a measure of full-time-equivalent (FTE) enrollment that took the values 1 (for full-time enrollment), 0.75 for three-quarter-time enrollment, 0.5 for half-time enrollment, 0.25 for some but less than half-time enrollment, and 0 for no enrollment.[36] To translate these to person-quarter-level outcomes, a student was counted as enrolled for the quarter if they were enrolled in any of the three months of that quarter, and FTE enrollment was calculated by summing the student’s total FTE months for the quarter.

C.3 Program Impacts on NSC-Measured Outcomes

Exhibit C-2 compares a selection of estimated impacts of WTA Connect using both NSC records and survey data. We included this table as a check on the impacts estimated in the main body of the report using survey data. The use of survey data allowed us to estimate impacts on variables not measurable with NSC data (such as receipt of particular types of credentials).

The pattern of effects of WTA Connect based on the two records systems is broadly consistent—except for receipt of a college credential. NSC is capturing very few of the credentials that these students are earning, and therefore fails to capture the effect of WTA Connect. The reason for this difference is not clear. NSC relies on college staff to decide which credentials to report. Respondents who reported earning a credential almost always identified Des Moines Area Community College as the issuer. Evidently, college staff there chose not to report to NSC, perhaps because these credentials are not degrees.

---

35 Dundar and Shapiro (2016) indicate that schools that choose to submit information on type of credential pursued or earned do so voluntarily and with minimal processing by NSC staff. About 90 percent of students attend schools that do submit information on credential types, but there is no systematic classification scheme for credentials that are not degrees. Schools merely submit names of certifications and diplomas awarded. The authors also specifically note that information on earned credits is weak. In addition, Dynarski, Hemelt, and Hyman (2015) report that only about 80 percent of degrees from Michigan colleges were reported to the NSC in the 2008-2010 period.

36 Because informed consent had been collected from all study participants, NSC shared full/part-time status for everyone in the sample, something that it does not otherwise share with researchers.

Exhibit C-2: Comparisons of Impacts of Workforce Training Academy Connect Based on Survey Data vs. Impacts Based on NSC Records

Outcome	NSC Records				Survey Data					
	Treatment Group	Control Group	Impact (Difference)	Standard Error	Treatment Group	Control Group	Impact (Difference)	Standard Error	Difference in Impacts	Standard Error
Any College Enrollment (%)										
In Q4	9.4	8.2	+1.1	1.9	16.1	13.5	+2.6	2.8	−1.5	2.9
In Q8	5.5	5.5	+0.0	1.5	9.4	12.2	−2.7	2.4	+2.8	2.5
In Q12	5.5	5.3	+0.2	1.5	9.5	9.4	+0.1	2.3	+0.1	2.2
Cumulative Number of FTE Months of College Enrollment										
Through Q12	1.2	1.2	+0.1	0.2	2.5	2.5	+0.1	0.5	+0.0	0.4
Any Completions from a College (%)										
Through Q12	1.8	1.1	+0.7	0.8	19.5	14.0	+5.5*	2.9	−4.8*	2.9
Sample size	470	473			348	350				

Source: PACE three-year follow-up survey; National Study Clearinghouse.

Note: Statistical significance is based on two-tailed tests. Statistical significance levels based on differences between research groups are summarized as follows: *** at the 1 percent level; ** at the 5 percent level; * at the 10 percent level.

Appendix D: NDNH's Unemployment Insurance Wage Data

Through the 1990s, many social program evaluations relied on administrative earnings data provided by state Unemployment Insurance (UI) agencies. State agencies maintained these data, and privacy concerns sometimes precluded sharing with outside researchers. UI records have become more accessible since 1996 with the advent of a centralized national database—the National Directory of New Hires (NDNH). Among NDNH's virtues is that, unlike state data, it captures earnings for study participants who move to another state during the follow-up period.

The federal Office of Child Support Enforcement (OCSE) in the U.S. Department of Health and Human Services' Administration for Children and Families (ACF) operates NDNH.[37] It contains new hire, quarterly wage, and UI information submitted by State Directories of New Hires, employers, and state workforce agencies. OCSE also supplements the state reports with records about earnings from federal civilian and military jobs (which are otherwise not covered by state UI data). Given this supplementation, the most important sources of uncaptured earnings are from self-employment, firms' employment of independent contractors, unreported tips, and informal employment.[38]

D.1 Data Collection Process

The primary purposes of NDNH are to assist state child support agencies to locate noncustodial parents, putative fathers, and custodial parents to establish paternity and child support obligations and to enforce and modify orders for child support, custody, and visitation. It is also used by state UI agencies and the federal Social Security Administration to identify overpayments of benefits. However, subject to federal law, regulation, guidance, and other requirements to protect data privacy and security,[39] OCSE may disclose certain information contained in NDNH to requesting local, state, or federal agencies for research likely to contribute to achieving the purposes of part A or part D of title IV of the Social Security Act. Part A governs the federal Temporary Assistance for Needy Families (TANF) program. Part D governs the state/federal child support program. Such disclosures may not include the names, Social Security numbers (SSNs), or other personally identifying information.

If the disclosure is approved, the agency and OCSE must work together on the operational issues surrounding the technical and procedural aspects of the disclosure, such as mitigating

---

37 More detail is available at: https://www.acf.hhs.gov/css/training-technical-assistance/guide-national-directory-new-hires.

38 According to the U.S. Bureau of Labor Statistics, about 10 percent of workers are self-employed: https://www.bls.gov/spotlight/2016/self-employment-in-the-united-states/home.htm.

39 The legal authority for this disclosure for research purposes is contained in subsection 453(j)(5) of the Social Security Act and Section 5507 of the Patient Protection and Affordable Care Act. For more information, see: https://www.govinfo.gov/app/details/USCODE-2010-title42/USCODE-2010-title42-chap7-subchapIV-partD-sec653.

the risks of identifiability and establishing appropriate data retention and disposition schedules of data files.

ACF's Office of Planning, Research, and Evaluation (OPRE) and OCSE negotiated a memorandum of understanding allowing access to NDNH data for the PACE project. Among other provisions, the memorandum dictates what self-reported data from study subjects may be merged with NDNH data, the computing environment where these merges are conducted, and procedures for review of tables prior to release.

The PACE research team transmits match request files to OCSE quarterly. These match request files contain the names and SSNs of PACE study participants. OCSE verifies with the Social Security Administration that the reported SSNs belong to the named persons. For those SSNs that pass this test, OCSE copies NDNH records for that quarter and the preceding seven quarters to a secure folder on the ACF server.[40] (Ordinarily, these records would be destroyed after two years.) These copied records contain a pseudo-SSN; the records are stripped of all personal identifiers.

States are required to submit earnings records to OCSE within four months, but there are stragglers and corrections. To be safe, PACE analyses limit NDNH-based measures to time periods that ended at least six months prior to the extract date.

Once we are ready to analyze the collected data, we submit a "passthrough" file to OCSE containing a variety of PACE-assigned variables (such as treatment status and program ID) and self-reported variables (such as the baseline information described in Appendix A). OCSE then strips the personal identifiers out of the passthrough file and replaces the actual SSNs with the same pseudo-SSNs previously assigned to the archived wage records. The study then uses these pseudo-SSNs to merge program and self-reported data with NDNH quarterly wage data on ACF's secure server to estimate program impacts on earnings and employment.

D.2 Data and Measures

Random assignment for Workforce Training Academy Connect (WTA Connect) started in April 2012 and ended in December 2014. Given the lag of up to six months in processing of employer reports by the states and transfer of state data to OCSE, wage records from NDNH were available through Q4 2018; this means that we had 26 post-randomization quarters of earnings data for the earliest randomized study participants and 16 post-randomization quarters of earnings data for the last randomized study participants. In addition, we had eight quarters of pre-randomization data for the entire sample (we included the four most recent pre-randomization quarters in our regression-adjustment models).

Of the 943 treatment and control group members randomized as part of the WTA Connect evaluation, 920 study participants reported a name and SSN that OCSE deemed to be of

40 Those study participants who are not matched in the Social Security Administration database are considered "missing" for these purposes, because their employment records are not available.

sufficient quality for its matching purposes.[41] Analyses in this three-year report thus are based on the 97 percent of the sample the agency deemed suitable. This sample's earnings in each quarter were based on earnings records found for each sample member in matching. As usual in use of such data, we defined sample members as "not working" when there was no match to wage records in a given quarter.

Each quarter, we submitted a match request file to OCSE that contained the names and SSNs for everyone randomized to that date. For those where the SSNs and names aligned, OCSE returned earnings data for the eight most recent quarters in NDNH, which is lagged by two quarters from the date of the match. This meant that we had up to eight wage reports for each quarter. We used the last version for each quarter within a window. For example, for earnings in the second quarter of 2014, we used reports from the match file for the third quarter of 2016 and discarded the seven earlier sets of earnings data for the second quarter of 2014.

When the earnings data for a quarter contained two or more reports for the same person from the state, we assumed that these reports reflected either different payments by the same employer or payments from different employers. Consistent with the logic discussed in Appendix F, we reviewed quarterly earnings for any values that were clearly impossible, but failing to find any such values, did not discard or top-code any large earnings amounts.[42]

We calculated two outcomes for each quarter: a binary indicator of "any earnings" (yes/no) and the total reported wages for the quarter ($). The result was two series of 23 measures for each person (employment and earnings for the four quarters before randomization, the quarter of randomization, and the 16 quarters after randomization). In addition, we calculated average quarterly earnings for Q12 and Q13 after random assignment (the confirmatory earnings outcome, established to align with the theory of change) and annual earnings for Q10-Q13.

[41] The acceptability of the combination of a name and an SSN can vary over time. OCSE reviews the SSN ownership every quarter for the entire sample.

[42] Top-coding means values above a threshold are set equal to the threshold.

Appendix E: Comparing NDNH- and Survey-Based Employment and Earnings Estimates

Barnow and Greenberg (2015) review findings from evaluations including both the National Directory of New Hires (NDNH) and surveys as data sources. Although average survey-reported earnings tend to be higher than average total Unemployment Insurance (UI) earnings, impact estimates still may be nearly unbiased (Kornfeld and Bloom 1999). In the evaluation of Workforce Training Academy Connect (WTA Connect), average quarterly earnings agree rather well between the two measurement systems, but correlational analysis shows that there must be considerable measurement noise in one or both. The correlation in person-level quarterly earnings between the two systems at Q12 is just 0.63 for the treatment sample and 0.65 for the control sample.[43]

This section compares estimates of employment and earnings impacts based on NDNH data and survey self-reports.[44] It also presents estimates of the impact of WTA Connect on self-employment earnings.

The top panel in Exhibit E-1 below shows the degree of agreement of impact estimates for WTA Connect derived from the two sources. The estimated impact based on UI records of +$180 for average earnings in Q12 is quite close to the estimated impact of +$167 for Q12 based on three-year follow-up survey data.[45] We explored whether earnings from self-employment could explain the difference between +$180 and +$167 if we were to treat the difference as real. Though the difference between these two impacts is small (just $13), earnings from self-employment are too small to explain it. It could be that the difference is just due to random memory errors by respondents.

Another plausible contributing cause to the discrepancy is differential undercoverage in NDNH. Barnow and Greenberg (2015) noted that state UI tax databases do not cover federal workers; out-of-state records; most workers at small farms, at railroads, at selected nonprofit organizations (particularly churches); and some casual or irregular jobs. Hiding of tip income

---

43 The survey figures convert the available survey measure—earnings in the prior week (calculated as hourly wage multiplied by number of hours worked)—to a calendar-quarter-level estimate by multiplying by 13 (the average number of weeks in a quarter).

44 From the follow-up survey, we had a complete history of jobs, with the starting wage and hours for each job as well as the last wage and hours for each job. We combined these to establish weekly earnings for the first and last weeks of a job. We then interpolated to get wages for each intervening month. We then summed weekly wages across jobs for multiple-job holders to get weekly earnings for every week between randomization and interview. Finally, we summarized these to the person-quarter level.

45 Assuming a correlation of 0.64 between the two person-level latent effects (the average of the correlations between NDNH- and survey-reported earnings for the two groups), the standard error between the two estimated impacts is $220, which is larger than the difference between the two impact estimates.

and income from household employment (such as childcare and cleaning) are additional important sources of undercoverage. In some states, independent contractors are not included. NDNH remedies the undercoverage of federal workers and of out-of-state workers, but not the other causes of undercoverage. If control group members are more likely to find employment of the types undercovered by NDNH, then that could lead to positive bias in the NDNH-based impact. However, because the NDNH-estimated impact is not statistically significant and the difference between the NDNH- and survey-estimated impacts is also not statistically significant, there seems to be no reason to be concerned about this issue.

The second panel of Exhibit E-1 shows that NDNH-based employment estimates are similar to the survey-based estimates for control group members (both roughly 65 percent). For treatment group members, however, the NDNH-based employment estimate is larger (73 percent, compared with 65 percent). Theoretically, the percentage of study participants with any earnings over three months is bound to be higher than the percentage employed on a particular day. This could explain why NDNH-measured employment is more common than survey-reported employment in the treatment group. However, this does not explain why the impact on employment is significantly larger when measured with NDNH data instead of survey data.[46]

We could identify no explanation for this discrepancy. It is striking though that survey data on employment appear to be more consistent with earnings than are the NDNH data. It is not clear how WTA Connect could have such a strong effect on employment without also increasing earnings.

Exhibit E-1: Impacts of Workforce Training Academy Connect on Earnings and Employment around Q12 Based on Wage Records and Self-Reports

Outcome	Treatment	Control	Impact	Standard Error
Quarterly Earnings				
Average NDNH earnings in Q12 ($)	3,844	3,663	+180	233
Self-reported earnings in Q12 ($)	4,156	3,989	+167	278
Self-reported earnings from self-employment in Q12 ($)	13	17	−3	12
Employment				
Average percentage with employer-reported wages in Q12	72.7	65.1	+7.6**	3.0
Percentage working in the week prior to survey interview	65.2	64.5	+0.7	3.7
Sample sizes				
NDNH	461	459		
Survey	348	350		

Source: National Directory of New Hires; PACE three-year follow-up survey.

Note: Self-reported earnings are calculated for the week prior to the survey interview, based on reported work hours and wages, and multiplied by 13 weeks for a quarterly estimate. A majority of survey interviews occurred in the 12th and 13th follow-up quarters.

Statistical significance levels, based on two-tailed tests, are summarized as follows: *** 1 percent level; ** 5 percent level; * 10 percent level.

46 We did not calculate correlations between NDNH- and survey-reported employment for the two groups; however, using the average correlation for the earnings as we did before, we obtain an approximate standard error for the difference between the two estimated impacts of 2.9 percentage points.

Appendix F: Treatment of Outliers

We took a conservative approach to outliers, retaining extreme values except where they were clearly impossible. This approach is based on the general difficulty of discriminating between errors and legitimate large values and on the fact that remedies require assumptions about true values that may not be correct.

Trimming observations could easily introduce non-ignorable nonresponse by making nonresponse a function of Y.[47]

Winsorizing observations (also known as “top-coding,” where values above a threshold are set equal to the threshold) could introduce bias if there is a treatment impact but the same threshold is used for treatment and control group members (and there is no reasonable basis for setting different thresholds for the two groups).

Furthermore, evidence suggests that results are generally robust to extreme values. In particular, research by Judkins and Porter (2016) and Lumley et al. (2002) indicates that for the sample sizes available in this evaluation, ordinary least squares inference on the reported data should be robust to outliers.

Outcomes assessed for extreme values included instructional hours (by type of instruction), credits, and National Directory of New Hires earnings. We found no values that were clearly impossible, and thus retained all reported values in the analysis.

47 Trimming by definition creates item nonresponse because the provided response is discarded. If trimming is a function of observed Y, as is standard, and if there is some relationship between observed Y and true Y, then item nonresponse becomes a function of true Y, which is known as “non-ignorable nonresponse.” Because there is no known way to remove bias due to non-ignorable nonresponse, trimming is likely to create uncorrectable biases in estimated treatment effects.

Appendix References

Barnow, B. S., and D. Greenberg. 2015. "Do Estimated Impacts on Earnings Depend on the Source of the Data Used to Measure Them? Evidence from Previous Social Experiments." *Evaluation Review* 39 (2): 179-228. Doi:10.1177/0193841X14564154.

Beicht, Ursula, and Michael Friedrich. 2008. "Anlage und Methode der BIBB-Übergangsstudie." In *Ausbildungschancen und Verbleib von Schulabsolventen*, edited by Ursula Beicht, Michael Friedrich, and Joachim Gerd Ulrich, 79-99. Bielefeld, Germany: W. Bertelsmann.

Betz, N. E., and K. M. Taylor. 2001. *Manual for the Career Decision Self-Efficacy Scale and CDMSE—Short Form.* Columbus, OH: The Ohio State University.

Bühlmann, P., and S. van de Geer. 2011. *Statistics for High-Dimensional Data.* Berlin, Heidelberg, Germany: Springer.

Ciolino, Jody D., Hannah L. Palac, Amy Yang, Mireya Vaca, and Hayley M. Belli. 2019. "Ideal vs. Real: A Systematic Review on Handling Covariates in Randomized Controlled Trials." *BMC Medical Research Methodology* 19: 136. https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-019-0787-8.

Cohen, S., R. Kamarck, and R. Mermelstein. 1983. "A Global Measure of Perceived Stress." *Journal of Health and Social Behavior* 24 (4): 385-396.

Cutrona, C., and D. Russell. 1987. "The Provisions of Social Relationships and Adaptation to Stress." *Advances in Personal Relationships*, 1.

Deke, J., and H. Chiang. 2017. "The WWC Attrition Standard: Sensitivity to Assumption and Opportunities for Refining and Adapting to New Contexts. *Evaluation Review* 41: 130-154. https://journals.sagepub.com/doi/10.1177/0193841X16670047.

Deville, J. C., and C. E. Särndal. 1992. "Calibration Estimation in Survey Sampling." *Journal of the American Statistical Association* 87: 376-382. https://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475217.

Duckworth, Angela L., C. Peterson, M. D. Matthews, and D. R. Kelly. 2007. "Grit: Perseverance and Passion for Long-Term Goals." *Journal of Personality and Social Psychology* 92 (6): 1087-1101. https://psycnet.apa.org/record/2007-07951-009.

Dundar, A., and D. Shapiro. 2016. *The National Student Clearinghouse as an Integral Part of the National Postsecondary Data Infrastructure.* Retrieved from the National Student Clearinghouse Research Center website: https://nscresearchcenter.org/wp-content/uploads/NSC-as-an-Integral-Part-of-the-National-Postsecondary-Data-Infrastructure.pdf.

Dynarski, S. M., S. W. Hemelt, and J. M. Hyman. 2015. "The Missing Manual: Using National Student Clearinghouse Data to Track Postsecondary Outcomes." *Educational Evaluation and Policy Analysis* 37(1s): 53S–79S. https://journals.sagepub.com/doi/pdf/10.3102/0162373715576078.

Farrell, Mary, and Karin Martinson. 2017. Pathways for Advancing Careers and Education (PACE). The San Diego County Bridge to Employment in the Healthcare Industry Program: Implementation and Early Impact Report. OPRE Report 2017-41. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. https://www.acf.hhs.gov/opre/report/san-diego-county-bridge-employment-healthcare-industry-program-implementation-and-early.

Fein, D. J. 2012. Career Pathways as a Framework for Program Design and Evaluation: A Working Paper from the Pathways for Advancing Careers and Education (PACE) Project. OPRE Report 2012-30. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. https://www.acf.hhs.gov/opre/resource/career-pathways-as-a-framework-for-program-design-and-evaluation-a-working.

Folsom, R. E. 1991. "Exponential and Logistics Weight Adjustments for Sampling and Nonresponse Error Reduction. In *Proceedings of the American Statistical Association, Social Statistics Section,* 197-202. Alexandria, VA: American Statistical Association.

Folsom, R. E., and A. C. Singh. 2000. "The Generalized Exponential Model for Sampling Weight Calibration for Extreme Values, Nonresponse, and Post-Stratification." In *Proceedings of the American Statistical Association, Section on Survey Research Methods,* 598-603. Alexandria, VA: American Statistical Association.

Folsom, R. E., and M. Witt. 1994. "Testing a New Attrition Nonresponse Adjustment Method for SIPP." In *Proceedings of the American Statistical Association, Section on Survey Research Methods,* 428-433. Alexandria, VA: American Statistical Association.

Goldrick-Rab, S., and K. Sorensen. 2010. "Unmarried Parents in College." *Future of Children* 20 (2): 179-203.

Hamadyk, J., and M. Zeidenberg. 2018. *Des Moines Area Community College Workforce Training Academy Connect Program: Implementation and Early Impact Report.* OPRE Report 2018-82. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. https://www.acf.hhs.gov/opre/resource/des-moines-area-community-college-workforce-training-academy-connect-program-implementation-early-impact-report.

Hendra, Richard, and Aaron Hill. 2018. "Rethinking Response Rates: New Evidence of Little Relationship between Survey Response Rates and Nonresponse Bias." *Evaluation Review.* doi:10.1177/0193841X18807719.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945–960. doi:10.1080/01621459.1986.10478354.

Hoven, M. R. 2012. "Investigating the Relationship between Perceived Social Support and Parent Self-Efficacy in Parents of Preschool-Aged Children." Master's Thesis. University of British Columbia. https://circle.ubc.ca/bitstream/handle/2429/43343/ubc_2012_fall_hoven_michaelyn.pdf?sequence=3. Last accessed 8/28/2015.

Izrael, David, David C. Hoaglin, and Michael P. Battaglia. 2000. "A SAS Macro for Balancing a Weighted Sample." In *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference,* Paper 275. Cary, NC: SAS Users Group International. https://pdfs.semanticscholar.org/f777/e121632ccc23bc2332efa8d1d2b4a5a311d3.pdf.

Judge, T. A. 2009. "Core Self-Evaluations and Work Success." *Current Directions in Psychological Science* 18 (1): 58-62.

Judge, Timothy, and Joyce E. Bono. 2001. "Relationship of Core Self-Evaluation Traits — Self-Esteem, Generalized Self-Efficacy, Locus of Control, and Emotional Stability — with Job Satisfaction and Job Performance: A Meta-Analysis." *Journal of Applied Psychology* 86 (1): 80-92.

Judge, Timothy, Edwin A. Locke, and Cathy C. Durham. 1997. "The Dispositional Causes of Job Satisfaction: A Core Evaluations Approach." *Research in Organizational Behavior* 19: 151-188.

Judge, Timothy, Edwin A. Locke, and Cathy C. Durham. 1998. "Dispositional Effects on Job and Life Satisfaction: The Role of Core Evaluations." *Journal of Applied Psychology* 83 (1): 17-34.

Judkins, David. 2019. "Covariate Selection in Small Randomized Studies." Presentation at the Joint Statistical Meetings, Denver, Colorado. https://ww2.amstat.org/meetings/jsm/2019/onlineprogram/AbstractDetails.cfm?abstractid=307372.

Judkins, David, David Fein, and Larry Buron. 2018. *Analysis Plan for the PACE Intermediate (Three-Year) Follow-up Study.* OPRE Report 2018-95. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. https://www.acf.hhs.gov/opre/resource/analysis-plan-for-the-pace-intermediate-three-year-follow-up-study.

Judkins, David R., and Kristin E. Porter. 2016. "Robustness of Ordinary Least Squares in Randomized Clinical Trials." *Statistics in Medicine* 35 (11): 1763-1773. https://www.statisticsviews.com/details/journalArticle/9169971/Robustness-of-ordinary-least-squares-in-randomized-clinical-trials.html.

Judkins, D., D. Morganstein, P. Zador, A. Piesse, B. Barrett, and P. Mukhopadhyay. 2007. "Variable Selection and Raking in Propensity Scoring." *Statistics in Medicine* 26: 1022-1033. https://onlinelibrary.wiley.com/doi/10.1002/sim.2591.

Judkins, David, Randall Juras, Samuel Dastrup, and Mary Ferrell. 2020. *The San Diego Workforce Partnership's Bridge to Employment in the Healthcare Industry Program: Appendices for Three-Year Impact Report.* OPRE Report 2020-105 Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. https://www.acf.hhs.gov/opre/resource/the-san-diego-workforce-partnerships-bridge-to-employment-in-the-healthcare-industry-program-three-year-impact-report-0.

Kessler, R. C., G. Andrews, D. Mrocek, B. Ustun, and H. U. Wittchen. 1998. "The World Health Organization Composite International Diagnostic Interview Short-form (CIDI-SF)." *International Journal of Methods in Psychiatric Research* 7 (4): 171-185. https://onlinelibrary.wiley.com/doi/abs/10.1002/mpr.47.

Koch, Gary G., Catherine M. Tangen, Jin-Whan Jung, and Ingrid A. Amara. 1998. "Issues for Covariance Analysis of Dichotomous and Ordered Categorical Data from Randomized Clinical Trials and Non-parametric Strategies for Addressing Them." *Statistics in Medicine* 17: 1863-1892. https://onlinelibrary.wiley.com/doi/abs/10.1002/(SICI)1097-0258(19980815/30)17:15/16%3C1863::AID-SIM989%3E3.0.CO%3B2-M.

Kornfeld, R., and H. Bloom. 1999. "Measuring Program Impact on Earnings and Employment: Do Unemployment Insurance Wage Reports from Employers Agree with Surveys of Individuals?" *Journal of Labor Economics* 17 (1): 168-197. https://www.journals.uchicago.edu/doi/pdfplus/10.1086/209917.

Le, H., A. Casillas, S. Robbins, and R. Langley. 2005. "Motivational and Skills, Social, and Self-Management Predictors of College Outcomes: Constructing the Student Readiness Inventory." *Educational and Psychological Measurement* 65 (3): 482-508. https://www.academia.edu/527739/Motivational_and_skills_social_and_self-management_predictors_of_college_outcomes_Constructing_the_Student_Readiness_Inventory.

Lin, W. 2013. "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique." *The Annals of Applied Statistics* 7: 295-318. https://projecteuclid.org/download/pdfview_1/euclid.aoas/1365527200.

Lumley, T., P. Diehr, S. Emerson, and L. Chen. 2002. "The Importance of the Normality Assumption in Large Public Health Data Sets." *Annual Review of Public Health* 23: 151-169. https://www.annualreviews.org/doi/pdf/10.1146/annurev.publhealth.23.100901.140546.

Peterson, C. H., A. Casillas, and S. B. Robbins. 2006. "The Student Readiness Inventory and the Big Five: Examining Social Desirability and College Academic Performance." *Personality and Individual Difference* 41 (4): 663-673. https://isiarticles.com/bundles/Article/pre/pdf/76798.pdf.

Research Triangle Institute. 2012. *SUDAAN Language Manual, Volumes 1 and 2, Release 11.* Research Triangle Park, NC: Author.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Stumpf, S. A., S. M. Colarelli, and K. Hartman. 1983. "Development of the Career Exploration Survey (CES)." *Journal of Vocational Behavior* 22 (2): 191-226. https://www.sciencedirect.com/science/article/abs/pii/0001879183900283.

Tukey, John W. 1991. "Use of Many Covariates in Clinical Trials." *International Statistical Review* 59(2):123-137. https://www.jstor.org/stable/1403439?seq=1.

Walker, Joan M. T., Andrew S. Wilkins, James R. Dallaire, Howard M. Sandler, and Kathleen V. Hoover-Dempsey. 2005. "Parental Involvement: Model Revision through Scale Development." *The Elementary School Journal* 106 (2): 85-104.

Williams, R. L., and R. E. Folsom. 1981. "Weighted Hotdeck Imputation of Medical Expenditures Based on a Record Check Subsample." In *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 406-411. Alexandria, VA: American Statistical Association.