Department of Labor Evaluation Design Pre-Specification Plans

Background

The Department of Labor (DOL)'s Chief Evaluation Office (CEO) is committed to upholding the department's Evaluation Policy principles of rigor, relevance, transparency, independence, and ethics in independent evaluations. For all rigorous experimental studies and studies using methods described as quasi-experimental, the CEO will publish Evaluation Design Pre-Specification Plans during the planning stages of evaluations to promote transparency and replicability. It is important to note that changes may occur during the course of conducting research after the publication of design plans, and final evaluation products will clearly note where and why research is altered in published plans.

This document provides a template that evaluators must use to meet the pre-specification practices articulated in OMB Memo M-20-12 *Phase 4 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Program Evaluation Standards and Practices*. OMB Memo M-20-12 calls for making an "evaluation's design and methods available before the evaluation is conducted and in sufficient detail to achieve rigor, transparency, and credibility by reducing risks associated with the adoption of inappropriate methods or selective reporting of findings, and instead promoting accountability for reporting methods and findings." The information reported must also provide sufficient information so that final reporting could be assessed per the DOL Clearinghouse for Labor Evaluation and Research (CLEAR) evidence guidelines. Evaluators may also find it helpful to refer to their Office of Management and Budget's Paperwork Reduction Act Information Collection Request requirements submissions.

Document Control

Table 1. Document Information

Title:	DOL Evaluation Design Pre-Specification Plan: Apprenticeship Evidence-Building Portfolio
Evaluator	Urban Institute (prime)/Mathematica/Capital Research Corporation
Security Level:	Public; no access restrictions
Contact Info:	chiefevaluationoffice@dol.gov

Table 2. Document History

Version	Date	Summary of Change
1	10/2020	Initial version published.
2	03/2023	Transferring the initial version into this new DOL template.

Evaluation Design Report for the Apprenticeship Evidence-Building Portfolio Project

Item 1 – Purpose, Research Questions, and Hypotheses. *Briefly describe objective of the evaluation (its relevance). Include primary and secondary questions and hypotheses to be tested, including ancillary or exploratory questions.*

Purpose

The U.S. DOL awarded two sets of grants in 2019 and 2020 to expand apprenticeships. First, in 2019, DOL invested $184 million in the **Scaling Apprenticeship through Sector-Based Strategies** grants (referred to throughout as Scaling Apprenticeship grants) to expand both apprenticeships registered with the U.S. DOL or a State Apprenticeship Agency (SAA) and unregistered apprenticeships. Apprenticeship programs train participants across traditional and new industry sectors and occupations. Twenty-three grantees representing community colleges and college consortia in 18 states received awards ranging from $2 to $12 million over a four-year grant period to expand apprenticeship programs in sectors with high demand for skilled workers, most notably health care, information technology (IT), and advanced manufacturing. In addition, in 2020, DOL awarded nearly $100 million through the **Apprenticeships: Closing the Skills Gap** grants (referred to throughout as Closing the Skills Gap grants) to 28 public–private partnerships to expand apprenticeship in the same key sectors with grantee leads located in 23 states, with a particular focus on cybersecurity and artificial intelligence occupations. The awards ranged from $500,000 to $6 million over a four-year grant period.

The DOL CEO contracted with the Urban Institute and its partners, Mathematica and Capital Research Corporation, to conduct the Analysis of Strategies for Expanding Apprenticeship Portfolio project with the primary goal of understanding the impact and implementation of recent investments in apprenticeship sponsored by the Department, including the Scaling Apprenticeship and Closing the Skills Gap grants. The project includes designing and conducting rigorous evaluations to expand the evidence base on apprenticeships as well as an implementation study of apprenticeship strategies used by DOL grantees.

Apprenticeship models involve an industry- and employer-driven, structured approach to occupational training that combines on-the-job training (OJT) and related technical instruction (referred to as educational or instructional components by both the Scaling Apprenticeship and Closing the Skills Gap grant programs). Apprentices are paid, productive employees of an employer who either sponsors the apprenticeship program or partners with a program sponsor. An apprenticeship "program" is a structured training program for a specific occupation that includes an employer who provides OJT and paid employment to the apprentice; a related instruction provider; and a nationally recognized credential. Programs can also include one employer or multiple employers. If multiple employers sign on to the same apprenticeship standards, they are considered a single program. Employers or other program partners may

operate multiple programs in different occupations, or they may only operate a single program.[1] Apprenticeship programs are referred to as "registered" if they have program standards registered either with U.S. DOL or an SAA. Unregistered apprenticeship programs have all the characteristics of a registered program but are not registered with or monitored by U.S. DOL or an SAA.

This design options report presents strategies for rigorously studying the impact of the Scaling Apprenticeship and Closing the Skills Gap grants. In each section of this report, we first discuss the design for the evaluation of the Scaling Apprenticeship grants, and then follow with a discussion of how the design does or does not differ for the Closing the Skills Gap grants.

Research Questions

This report discusses potential study designs that could answer the following primary research questions of interest:

1. What is the impact of registered apprenticeships on earnings and employment of participants in the 9th and 10th quarters following program enrollment?
2. What is the impact of unregistered apprenticeships on earnings and employment of participants in the 9th and 10th quarters following program enrollment?

In addition, the report discusses designs to answer secondary research questions:

1. What are the impacts of unregistered and registered apprenticeships for different types of apprentices and their pathways to apprenticeship programs, such as incumbent worker apprentices, those referred to apprenticeship from the workforce system, and those that participate in apprenticeships after enrolling in community colleges?
2. What are the impacts of unregistered and registered apprenticeships on earnings and employment for subgroups defined by
    a. key participant characteristics, such as race, gender, and age;
    b. program receipt status; that is, those who received grant-funded services versus those that were hired as apprentices;
    c. key program features, such as the program length, or whether a grantee is directly sponsoring its apprenticeship programs or acting as an intermediary.

Although there have been quasi-experimental impact studies of registered apprenticeships (Hollenbeck and Huang 2016; Reed et al. 2012), there are no rigorous studies of the effectiveness of unregistered apprenticeships or more recent registered apprenticeship initiatives, and this project is designed to contribute to building a greater understanding of apprenticeships to inform future practice, policy, and grantmaking.

[1] For more information, see DOL's web page on registered apprenticeship programs at https://www.apprenticeship.gov/employers/registered-apprenticeship-program.

Hypotheses

Our hypotheses are informed by the existing research on the impact of apprenticeship training on participants' earnings and employment. Apprenticeships combine classroom learning with OJT and provide a credential upon completion. Registered Apprenticeships are programs that are registered under either DOL's Office of Apprenticeship or through recognized SAAs. Unregistered apprenticeships are independent programs that use the same earn-and-learn model, but do not go through the same review process for occupational standards that is required for registered apprenticeships. Pre-apprenticeship programs are a set of strategies designed to expand access and prepare individuals for entry into an apprenticeship program.[2] Unregistered apprenticeships can include a wide variety of approaches for upskilling an employee with occupation-specific training. Unregistered apprenticeships are often shorter than registered apprenticeships because they do not have to meet the same requirements for the number of hours of classroom instruction or workplace training. They are also not eligible for public funding that is earmarked for registered apprenticeships. For the Scaling Apprenticeship and Closing the Skills Gap grants, any program considered an unregistered apprenticeship has to meet the five hallmarks or characteristics of apprenticeship program quality outlined in the funding announcements. These include (1) a paid, work-based component, (2) OJT and mentorship, (3) an educational and instructional component, (4) an industry-recognized credential, and (5) safety, supervision, and equal employment opportunity.[3]

Apprenticeship is one of the most intensive workplace-based training models, and research in the U.S. has found that apprenticeships generate substantial benefits for individual apprentices and employers. Reed and colleagues (2012) found that nine years from the start of their program, 21,426 registered apprenticeship participants in ten states earned nearly $1,500 more in quarterly earnings than similar nonparticipants who enrolled in the program but did not participate in it would have earned. Nonparticipant earnings were estimated by comparing apprentices to one another in a dosage model (a statistical model estimating earnings outcomes as a nonlinear function of time in training and other covariates) and predicting earnings outcomes for nonparticipants with a dosage of zero. They also found positive impacts when comparing earnings gains of apprentices to similar people in the same states but using unemployment insurance (UI) wage records for apprentices and Current Population Survey data for nonparticipants. A study in Washington State used propensity score matching and found an impact of nearly $3,500 in quarterly earnings compared with several hundred thousand nonapprentices served by the Wagner-Peyser Employment Services program (Hollenbeck and Huang 2016). Helper et al. (2016) indicates that employers of apprentices also benefit from the reliable talent pipeline that apprenticeship provides, increased worker productivity, and reduced

[2] See the *Scaling Apprenticeship* grants funding announcement for details on these definitions, which are used for the purposes of the grant program: https://www.grants.gov/web/grants/view-opportunity.html?oppId=307212.
[3] See page 14 of the *Scaling Apprenticeship* grants funding announcement for details on the five hallmarks of an apprenticeship program: https://www.grants.gov/web/grants/view-opportunity.html?oppId=307212.

turnover. Studies estimated a rate of return on apprenticeship investment for one health care system of at least 40 percent, and for the company Siemens USA, of at least 50 percent (Helper et al. 2016). In other words, for every dollar invested in apprenticeship, the health care system resulted in $1.40 in returns. For Siemens USA, every dollar invested in apprenticeship resulted in at least $1.50 in returns to the company.

Apprenticeship has long been dominated by the construction trades (Boren et al. 2022). More programs are being developed in the health care, services, and IT industries (Gardiner et al. 2021; Walton, Gardiner, and Barnow 2022). The populations studied have also been largely male because males represent a large percentage of apprentices (Hollenbeck and Huang, 2016; Reed et al. 2012), although sectors such as health care and childcare have greater female representation (Walton, Gardiner, and Barnow 2022).

Based on this research literature, our hypothesis is that apprenticeship training will have a positive impact on participants' earnings and employment. We expect that these impacts will vary by subgroup. For example, in the health care sector where direct care positions can receive low pay and direct care pay is constrained by Medicare and Medicaid reimbursement (Lerman, Eyster, and Kuehn 2014), apprenticeship training may not have as large of an impact on earnings. Similarly, the impact for registered and unregistered apprenticeships may differ depending on the length and intensity of training. As registered and unregistered apprenticeship programs vary by length and intensity, we hypothesize that the impacts on earning gains from these two types of programs might differ.

Background on Grant Programs

To provide context on the design options for testing these hypotheses presented in this report, the rest of this section provides summary information on the apprenticeship programs funded by the Scaling Apprenticeship and Closing the Skills Gap grants. We describe the purpose of the grant programs, and then discuss the variation across the grants in each program in terms of the models used and other key features. Tables 1 and 2 present data on the key characteristics of the grantees' approaches for each grant program respectively.

Table 1

Planned Characteristics of Scaling Apprenticeship Grants, 2019–2024

#	Grantee name	Industry or Occupation: Advanced manufacturing	Industry or Occupation: Information technology	Industry or Occupation: Health care	Planned Apprenticeship Models: Registered	Planned Apprenticeship Models: Unregistered	Planned Apprenticeship Models: Pre-apprenticeship	Planned Recruitment Sources for Apprentices: Incumbent workers	Planned Recruitment Sources for Apprentices: High schools or colleges	Planned Recruitment Sources for Apprentices: Workforce system, community, or industry partners	Grant Targets: Programs (new and expanded)	Grant Targets: Apprentices employed*	Grant Targets: Employers engaged
1	**Colorado Department of Higher Education**			X	X		X	X	X	X	44	5,000	70
2	**County College of Morris**	X			X	X	X	X	X	X	52	1,360	104
3	**Connecticut State Colleges and Universities**	X			X		X	X		X	16	2,710	64
4	**Community College of Baltimore County**			X	X		X	X	X	X	4	436	11
5	**Columbus State Community College**		X			X	X	X	X		51	1,152	650
6	**Alabama Community College System**	X				X	X	X		X	59	2,500	75
7	**Bergen Community College**			X	X	X	X		X	X	15	3,500	170
8	**Dallas County Community College District**			X	X	X	X	X	X	X	53	5,910	159
9	**The Florida International University Board of Trustees**		X		X			X			22	800	18
10	**St. Louis Community College**	X			X	X	X	X	X	X	79	2,280	136
11	**State University of New York**	X			X		X	X	X	X	1,000	3,200	1,000
12	**Purdue University**		X		X			X		X	60	5,000	52
13	**Illinois Community College Board**		X		X	X	X	X	X	X	113	842	92

14	**Lorain County Community College**	X			X	X	X	X	X	X	70	5,000	500
15	**Miami Dade County**		X		X	X	X			X	43	518	80
16	**Pennsylvania College of Technology**	X			X		X	X		X	91	1,660	596
17	**Pima County Community College District**	X			X	X	X	X	X	X	19	386	16
18	**San Jacinto Community College District**		X		X	X	X		X	X	54	3,700	20
19	**Trustees of Clark University**		X		X		X			X	15	900	30
20	**University of Cincinnati**		X		X	X	X	X	X	X	225	3,778	110
21	**Weber State University**		X			X	X		X	X	11	650	35
22	**West Los Angeles College**	X			X	X	X	X	X	X	55	4,400	21
23	**West Virginia Council for Community and Technical College Education**		X		X			X	X	X	68	960	200
	Totals	**9**	**10**	**4**	**20**	**14**	**20**	**18**	**16**	**21**	**2,219**	**56,642**	**4,209**

Source: Grant applications and initial and follow-up phone calls with grantees.

Notes: Registered apprenticeship programs have program standards approved by U.S. DOL or an SAA. Unregistered programs are not registered with these agencies. Pre-apprenticeship programs are a set of strategies designed to expand access and prepare individuals for entry into an apprenticeship program.

*Apprentices employed include incumbent workers and nonincumbent workers expected to be employed out of the total population served.

Table 2

Planned Characteristics of Closing the Skills Gap Grants, 2020–2024

#	Grantee name	Industry or Occupation: Advanced manufacturing	Industry or Occupation: Information technology	Industry or Occupation: Health care	Planned Apprenticeship Models: Registered	Planned Apprenticeship Models: Unregistered	Planned Apprenticeship Models: Pre-apprenticeship	Planned Recruitment Sources for Apprentices: Incumbent workers	Planned Recruitment Sources for Apprentices: High schools or colleges	Planned Recruitment Sources for Apprentices: Workforce system, community, or industry partners	Grant Targets: Programs (new and expanded)	Grant Targets: Apprentices employed*	Grant Targets: Employers engaged
1	**Aerospace Machinist Joint Training Committee**	X		X				X	X	9	305	500	
2	**AFL-CIO Working for America Institute**	X		X					X	550	2,320	600	
3	**Alamo Colleges**		X	X	X	X		X	X	6	330	12	
4	**American Association of Port Authorities**	X	X	X	X	X			X	94	5,122	55	
5	**Argentum**		X	X	X			X	X	11	6,255	45	
6	**Arizona State University**	X	X	X	X			X		8	330	30	
7	**Arkansas Division of Higher Education**		X		X		X		X	X	125	2,000	110
8	**Colorado Community College System (CCCS)**		X		X	X			X	X	34	224	15
9	**Electrical Training ALLIANCE (EA)**	X			X		X	X	X	X	1	5,000	1500
10	**Florida Alcohol and Drug Abuse Association**			X	X	X				X	4	2,570	30
11	**Goodwin College, Inc.**	X			X			X		X	7	1,152	125
12	**Health care Career Advancement Program (H-CAP), Inc.**			X	X			X		X	35	2,426	50

13	**Idaho State Board of Education**	X	X	X	X			X	X	X	35	619	15
14	**Ivy Tech Community College of Indiana**	X	X		X	X		X	X	X	120	2,720	150
15	**Missouri Chamber Foundation**		X		X			X		X	24	3,285	15
16	**North Carolina State University**		X		X				X		52	8,200	53
17	**Oakland Community College**	X			X					X	179	720	179
18	**Regents of University of Colorado Springs**		X		X					X	5	5,196	950
19	**Rhode Island Office of the Postsecondary Commissioner**	X			X				X	X	6	500	10
20	**Society for Human Resource Management Foundation, Inc.**	X	X	X	X			X	X	X	1	940	100
21	**Southern Utah University**	X	X		X			X	X		11	1,705	145
22	**Southwest Tennessee Community College**		X		X	X		X		X	66	192	136
23	**The Regents of the University of California (Davis)**		X		X			X	X	X	20	856	300
24	**The Regents of the University of California (Riverside)**	X	X		X	X			X	X	35	264	125
25	**University of Louisville Research Foundation, Inc.**	X	X	X	X	X		X	X	X	49	1,804	5
26	**University of Wisconsin-Whitewater**	X	X	X	X				X	X	1	2,512	700
27	**Wireless Infrastructure Association**		X		X	X	X			X	1	7,400	27
28	**Wisconsin Regional Training Partnership, Inc. (WRTP)**	X	X		X					X	83	1,717	300
	Totals	**16**	**19**	**9**	**28**	**9**	**5**	**11**	**17**	**25**	**1,572**	**66,664**	**6,282**

Source: Grant applications and phone calls with grantees.

Notes: Registered apprenticeship programs have program standards approved by U.S. DOL or an SAA. Unregistered programs are not registered with these agencies. Pre-apprenticeship programs are a set of strategies designed to expand access and prepare individuals for entry into an apprenticeship program. The Closing the Skills Gap grants cannot be used to fund pre-apprenticeship, although many programs use pre-apprenticeship as a strategy in their services.

*Apprentices employed include incumbent workers and nonincumbent workers expected to be employed out of the total population served.

Together, DOL, states, and industry have invested billions of dollars over the past decade to encourage, develop and expand industry-driven apprenticeship training nationwide. The breadth of apprenticeship investments has resulted in a diverse sectoral, geographic, and institutional mix of apprenticeship programs and projects. Apprenticeship has traditionally been used in the building trades but is now also used in food preparation and serving, personal care and services, and sales occupations (Kuehn 2019), health care (Lerman, Eyster, and Kuehn 2014), advanced manufacturing, science and engineering (Kuehn, Hecker, and Simon 2019; Kuehn and Jones 2018), and finance (Elejalde-Ruiz 2016). Expansion into more occupations, particularly those with high demand for skilled workers, has led to new models of apprenticeship (Copson et al. 2021).

To assist with development and adaptation of apprenticeship models, DOL provides program and capacity development grants to strengthen intermediaries (organizations that coordinate apprenticeship partners), industry, and other partners. The Scaling Apprenticeship and Closing the Skills Gap grant programs are two of the largest recent federal apprenticeship investments and the two grant programs that are the primary focus of the Analysis of Strategies for Expanding Apprenticeship Portfolio project. The Scaling Apprenticeship grant awards (table 1), announced in June 2019 and totaling $184 million, focus on accelerating the expansion of apprenticeships to sectors with high demand for skilled workers and many H-1B visas to hire temporary foreign workers. The Closing the Skills Gap grants (table 2), announced in February 2020 and totaling nearly $100 million, focus on expanding apprenticeships in sectors where apprenticeships are not traditionally used as a training strategy.

DOL awarded Scaling Apprenticeship grants to 23 community colleges and college consortia based in 18 states and Closing the Skills Gap grants to 28 college systems, industry associations, nonprofits, and state education agencies based in 23 states. The grantees across both programs, though alike in their mission to implement effective, high-quality apprenticeship programs, vary in the number of individuals they plan to serve, industries and occupations targeted, apprenticeship models, and recruitment strategies, among other factors. Below we discuss the variation across the grants, as described by grantees in their applications and in phone calls with selected grantees.

Eleven of the 23 Scaling Apprenticeship grantees plan to offer a combination of both unregistered and registered apprenticeships. Among the remaining grantees, nine grantees offer only registered apprenticeships, while three offer only unregistered apprenticeships. All 28 Closing the Skills Gap grantees plan to offer registered apprenticeships. Nine of the grantees plan to offer a combination of registered and unregistered apprenticeships.

Twenty of the 23 Scaling Apprenticeship grantees plan to offer pre-apprenticeships in addition to their chosen apprenticeship models. Some grantees indicated they would be establishing new pre-apprenticeship programs; others indicated they were working to expand existing pre-apprenticeship programs. Many of the grantees offering pre-apprenticeships described their

purpose as providing remedial academic training. For example, if an employer identified an incumbent worker to begin apprenticing, but that worker lacked the prerequisite reading and math skills to begin college coursework for the apprenticeship program, the worker could gain those skills in the pre-apprenticeship program and then transition into the apprenticeship upon completion.

Pre-apprenticeship is not an allowable funded program activity under the Closing the Skills Gap grant. However, five grantees reported that they plan to provide pre-apprenticeship services in the overall program, though likely to be funded by other resources. These grantees hoped to include pre-apprenticeship offerings to expand their apprenticeship programs, meet their recruitment target numbers, and provide remedial support to potential apprentices. Nine additional grantees stated that pre-apprenticeship would be offered through partners to increase the pipeline for apprentices and to assess potential participants for basic skills.

Twenty-one of the 23 Scaling Apprenticeship grantees noted they would rely on referral assistance from public agency partners like workforce development boards, American Job Centers, state offices of adult education, and veteran resource centers, as well as community organizations. Sixteen planned to recruit high school and community college students to apprenticeship and pre-apprenticeship programs. Eighteen grantees planned for employers to recruit apprentices from among their own employees (incumbent workers). Among grantees whose target population was mostly or entirely made up of incumbent workers, grantees indicated in initial phone calls that identification and recruitment of apprentices would be largely left to employers.

Most Closing the Skills Gap grantees (25 of 28) stated that they would rely on external recruitment partners for apprentices. Recruitment partners varied, including workforce development boards, specialized recruitment agencies, industry groups, and community-based organizations that serve their target population. Seventeen grantees mentioned recruiting through high schools or colleges, either through their own college networks or through educational partnerships. Most (27 of the 28) grantees reported plans to train incumbent workers as apprentices.

The target number of employed or hired apprentices across the grant period varied widely by grantee in the Scaling Apprenticeship grant programs, from under 400 to nearly 6,000 (grantees did not specify separate target numbers for registered versus unregistered programs). Nine grantees had targets of 3,000 or more. Eight grantees set targets lower than 1,000 for total apprentices employed across the entire four-year grant period. In 2019 and 2020, conversations the study team held with grantees to clarify questions about their programs, many noted that they were behind schedule in their enrollment for the first year. Although one volunteered that they hoped to catch up by the end of the year, the current restrictions on work as a result of the COVID-19 pandemic, including work closures, vaccine and mask mandates, and college closures

or enforcement of remote learning, made it unlikely that many grantees would reach their first-year targets.

The target number of employed or hired apprentices for the Closing the Skills Gap grantees ranged from 192 to 8,200 (including incumbent workers). Eleven of the 28 grantees set employment targets lower than 1,000 apprentices hired. Seven grantees expected to employ more than 3,000 apprentices across the four-year grant period. Unlike the Scaling Apprenticeship grantees, Closing the Skills Gap grantees reported targets for incumbent and nonincumbent apprentices separately. Eight grantees planned to serve more incumbent workers than nonincumbents. Based on clarification calls with grantees, we found that COVID-19 had unprecedented effects on the ability of grantees to stay on schedule, however, and they were uncertain at the time if they would meet their first-year target numbers.

Like the number of apprentices, the targets set by Scaling Apprenticeship grantees for new and expanded apprenticeship programs as well as employers engaged varied widely. The average target number of new and expanded apprenticeship programs was 96, with a low of 4 and a high of 1,000. The average target number of employers engaged among all the grantees was 183, with a low of 11 and a high of 1,000. Most grantees (15 of 23) indicated a larger target number of employers engaged than their target number of apprenticeship programs, suggesting that some programs would serve multiple employers. Eight grantees had the opposite, suggesting that they would have fewer multiple employer programs and would in some cases have multiple programs for each employer to participate in for different occupations or positions at the employer's business.

Wide variation also existed in the target number of employers engaged and the target number of new and expanded apprenticeship programs among the Closing the Skills Gap grantees. The average target number of new and expanded programs was 56 with a low of 1 program and a high of 550. The average target number of employers engaged was 224, with a low of 5 and a high of 1,500 employers. Twenty-one grantees expected to have more employers engaged than new and expanded programs, while seven expected to have more programs than employers.

Item 2 – Evaluation Design. *Briefly describe the overall evaluation methodological approach, based on a logic model of the program or policy being evaluated. Briefly discuss the program of interest and the feasibility of the planned approach, including the process for developing credible control or comparison groups. Include any anticipated challenges that could result in changes in the methodological approach and plans for how to address those challenges.*

The conceptual framework for evaluating the impact of unregistered and registered apprenticeship programs under the two grant programs, developed by the study team, is presented in figure 1 below. The framework identifies the challenges and objectives the grants seek to address—the needs of business and industry, workers, and state and local apprenticeship systems. It includes resources that may be used to support grant activities—the Scaling Apprenticeship and Closing the Skills Gap grants themselves, as well as other relevant national initiatives or existing state and local apprenticeship systems and partnerships. The framework also specifies program models and components of apprenticeship programs the grantees and their partners design and implement, and strategies and partnerships the grantees and partners use to expand apprenticeship. Finally, it includes the expected short-term outcomes and long-term outcomes of unregistered and registered apprenticeship programs.

To estimate the impact of the apprenticeship program on earnings and employment, we propose a quasi-experimental design (QED) that relies on the "selection on observables" assumption (see Imbens 2004 for a review). This assumption is based on the idea that observational characteristics can account for key factors that relate to both enrolling in an apprenticeship program and earning and employment outcomes. Although this assumption can never be fully tested, our design adheres to the following principles, found in the literature, for generating credible comparison groups when studying workforce development programs using QEDs (Heckman et al. 1998; Heckman, Ichimura, and Todd 1997; Glazerman, Levy, and Meyers 2003):

1. selecting treatment and comparison groups from the same local areas so that they face the same local labor markets and service environments;
2. using a rich set of socio-demographic variables from a common data source for both samples; and
3. using preprogram earnings histories, in temporal periods no longer than a quarter, to capture pre-enrollment differences in earnings and employment that can influence later employment outcomes.

Apprentices and Comparison Groups

In our primary analysis, we define the apprentices used in the treatment group for this study as anyone ever hired as an apprentice in a registered or unregistered apprenticeship program funded by the Scaling Apprenticeship or Closing the Skills Gap grants (referred to in this report as the "program"). This includes those who do not complete the program or earn a credential but excludes pre-apprentices and other participants who may receive educational or instructional training and other services but are not hired as apprentices. As a secondary analysis, we will

estimate impacts using everyone who received grant-funded services, whether or not they were hired as apprentices (intention-to-treat analysis).

Rather than comparing apprenticeship to any one specific alternative training model, the research questions focus on the difference between the average outcomes of the apprentices and the average outcomes apprentices would achieve if the program did not exist (also referred to as the average treatment effect on the treated). Because there are a number of services or activities apprentices might have engaged with in the absence of the program, we considered a broad range of comparison groups. To guide the selection of comparison groups, table 3 shows the three types of apprentices that we consider: (1) those who enroll in the apprenticeship program after being referred through the public workforce system, or who come from unemployment or underemployment, (2) incumbent workers, and (3) those who are recruited to the apprenticeship program from the population of community college students. The strength of the QED depends on the extent to which we are able to identify a defensible comparison group for each of these recruitment sources, the counterfactual condition experienced by the comparison group, and the common data available for both the apprentices and the comparison groups (Heckman et al. 1998). In the next sections, we will introduce the planned comparison groups for each of the three apprentice types, as well as the data sources available to investigate them.

TABLE 3

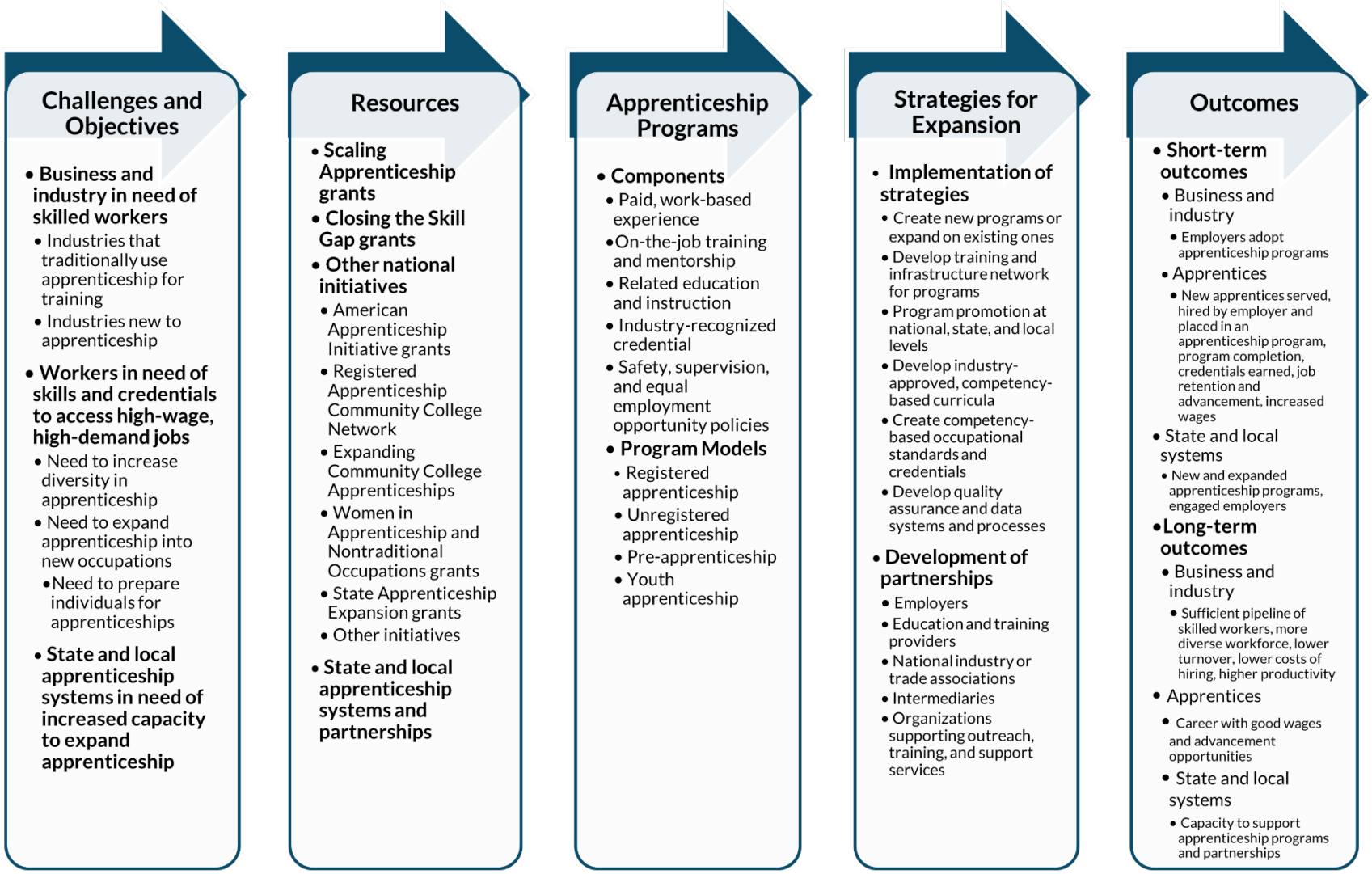
Planned Treatment and Comparison Group Members

Group	Treatment Group	Comparison Group
1	Apprentices referred to the program from the public workforce system	Wagner-Peyser participants
2	Apprentices who are incumbent workers	Wagner-Peyser participants, community college students
3	Apprentices who are community colleges students prior to program enrollment	Community college students

Source: Authors' analysis.

FIGURE 1

Conceptual Framework for Documenting and Assessing Impact



Group 1. Apprentices Referred to the Program from the Public Workforce System

Some apprentices are referred to grant programs from an American Job Center (AJC) or other public workforce agencies or providers. In the absence of the apprenticeship program, these apprentices would presumably be referred to another workforce development program or service. Therefore, Wagner-Peyser participants receiving some service from the public workforce system and whose data are in the Workforce Integrated Performance System (WIPS), with similar levels of education and prior earnings and living in similar geographic areas, constitute an appropriate comparison group. This comparison group includes people that receive more intensive training, as well as those who receive light-touch case management or job search assistance. This is appropriate because apprentices referred from an AJC might have received similar public workforce services in the absence of the apprenticeship program.

Group 2. Apprentices Who Are Incumbent Workers

Some grants in the program enroll incumbent workers, defined by the funding opportunity announcements as apprentices working for the sponsoring employer before beginning their program. For example, administrators of the Federation of Advanced Manufacturing Education (FAME) manufacturing technician apprenticeship program in Alabama reported that auto manufacturers might offer the program to production workers (workers on the assembly line who use machines to make cars), who could become higher-skilled maintenance workers (those who maintain the machines on the assembly line) upon completing the apprenticeship. Some apprenticeship programs are designed exclusively for incumbent workers.

In the absence of the apprenticeship program, we might assume that these incumbent workers would still be employed by the same employer. This makes the selection of the comparison group challenging because the apprentices are already in jobs. However, incumbent workers are an important population of apprentices, and some grantees may end up serving incumbent workers almost exclusively. Thus, it is important to include them in the QED. The following two groups are potential comparison groups for incumbent workers:

> (1) Individuals served by Wagner-Peyser and the WIOA Adult, Dislocated Worker, and Youth programs. Most people served by Wagner-Peyser Employment Services and WIOA programs are unemployed or underemployed, though some work full-time and are using these services to find better jobs. Incumbent workers, on the other hand, are all employed even before entering an apprenticeship program. This presents a considerable challenge and could prevent the identification of a similar enough comparison group from this data source in practice. However, there might be incumbent worker apprentices who are underemployed at baseline, and there may be nonapprentices in the WIPS data who are currently employed, at wages similar to those earned by the incumbent worker apprentices. Thus, while the overall population in these data would not be an appropriate comparison group for incumbent worker apprentices, a subset may be sufficiently similar regarding their employment history as well as socio-demographic characteristics. It is an empirical question as to whether credible comparison groups can be selected using this

approach. Once we have WIPS data and preprogram earnings, we will be able to evaluate whether we are able to form a comparison group that is sufficiently similar to incumbent apprentices.

(2) Community college students. Since many community college students are employed prior to and during their enrollment, we can use students who are enrolled in the same community colleges as the incumbent worker apprentices but do not pursue an apprenticeship. That said, because the National Directory of New Hires (NDNH) does not include data on industry and occupation, community college students in this comparison group could be employed in very different types of jobs than the incumbent worker apprentices.

Group 3. Apprentices Who Are Community College Students Prior to Program Enrollment

Some apprentices enroll in the participating community college prior to becoming apprentices, and likely would have enrolled in the community college even in the absence of the program. Indeed, some grantees reported that they have started or plan to recruit apprentices from their community college student bodies. In addition, some report that they are conducting outreach to high school seniors who can enroll upon graduating. Such apprentices may have attended a community college in the absence of the apprenticeship program. Thus, a subset of community college students from the partnering community college systems likely represents an appropriate comparison group for many apprentices. But in contrast with the comparison group for incumbent worker apprentices, these community college students do not need to have been employed prior to enrollment. This comparison group—community college students taking similar courses and with similar background characteristics as apprentices—would only be found in data provided by community colleges. Hence, we will need to collect data from community colleges. We are attempting to collect data for both credit and noncredit students, although our understanding is that reporting requirements mostly apply to degree-seeking students. If feasible, we will include noncredit students as comparison group members in our study.

Combining Comparison Groups

Ultimately, we will generate three distinct, independent analysis samples: one for each group described above. We will estimate and report impacts separately for each sample and also generate an average of the three impact estimates, weighting each impact equally or according to the number of apprentices. This is the same, in essence, as a stratified design. In this case, however, the actual data in each stratum (subset of the entire sample) is different, so the estimation will be done separately.

Challenges and Solutions

First, the project's success depends on obtaining data-sharing agreements with state workforce agencies, grantees, community colleges, and NDNH. As of February 2023, we have in place data agreements with NDNH, 22 grantees, community college systems from eight states (Alabama,

California, Illinois, Indiana, Missouri, New Jersey, Ohio, Texas),[4] and state workforce agencies from ten states (Alabama, Arkansas, Connecticut, Florida, Indiana, Michigan, Missouri, New Jersey, Pennsylvania, Utah). We are in the process of reaching out to four additional grantees in Indiana and Ohio. Since grant agreements require grantees to participate in an evaluation, we are optimistic to obtain these grantee data as well.

Second, because all apprentices appear in the WIPS data, one consequence of our strategy is that a single apprentice may appear in multiple groups. For example, they might appear in the community college student group as well as the WIPS data. (This will not always be the case if we cannot obtain data including personally identifiable information [PII] from all community colleges that apprentices attend.) For these apprentices, we will attempt to determine which sample, or stratum, is likely to provide a better comparison group and use them in that stratum only. Similarly, we will attempt to include each comparison group member in only one stratum. This ensures that the estimated impacts in the two strata are independent, making aggregation straightforward. If, on the other hand, we are unable to cleanly assign each sample member to a single group, we will account for the correlation in the impact estimates when averaging them. Apprentices may not fall into any of these categories if they are new hires who are not associated with either the public workforce system or the college system. In these cases, we will attempt to determine which sample—or stratum—is likely to provide a better comparison group and use them in that stratum only. The analysis section in item 6 contains further details on impact estimation.

Third, we have multiple options to construct comparison groups for incumbent workers, but we can only assess the credibility and therefore the feasibility of this subgroup analysis at a later stage of this project.

[4] For some states, we receive all available community college data in that state. For other states, we receive data from a few individual colleges.

Item 3 – Evaluation Data. *Describe data sources, the key outcomes and primary constructs of interest (including the level of measurement, such as individual, industry, firm or geographic area), and how they will be measured, including any variables that will be examined in existing administrative datasets. Describe any demographic data points, such as age, gender, race and ethnicity, etc., that will be available, and whether they may be meaningfully analyzed based on anticipated observations (including anticipated sample size or number of observations available after linking observation units across datasets, if merging administrative or other data sources). Include information about how the collected data will be verified or verifiable, and how it will accurately capture the intended information to address the questions of interest.*

Apprenticeship is a workforce development strategy and thus the confirmatory (primary) outcomes of interest are individual labor market outcomes: employment and earnings. An apprentice is, by definition, employed and earning wages while enrolled as an apprentice, whereas some members of the comparison group may not be employed, so the study would likely find impacts if examining earnings and employment outcomes soon after program enrollment. One hypothesis is that apprentices could ultimately attain positions that place them on a higher earnings trajectory than the comparison group, and this higher trajectory may not be fully realized for several years after completing the program. Thus, it is important to examine long-term earnings and employment to the extent possible.

Our primary employment and earnings outcomes will use the 9th and 10th quarters of available postenrollment data to balance the objectives of maximizing sample size and having the latest follow-up period possible, given the current project timeline.[5] We will examine two quarters to smooth anomalous findings that could occur if only one quarter was used.[6] The impact evaluation will rely on four main data sources: (1) DOL's WIPS, (2) community college data, (3) county-level data from the 2016 American Community Survey (ACS), and (4) earnings and employment records from the NDNH. This section gives a brief overview of these data sources before demonstrating the data collection process. We will discuss anticipated sample sizes and power calculations in item 5.

WIPS

The WIPS represents a national database that collects data on participants in workforce programs funded by DOL (as well as some programs funded by the Department of Education), including Wagner-Peyser Employment Services and the apprenticeship grants. Table 4 displays the key WIPS data items required for our evaluation. This list includes those variables that may be associated with both enrollment in an apprenticeship program and labor market outcomes, such as gender, race, age, veteran status, education level, ex-offender status, and low-income status

[5] For the treatment group, enrollment is defined as the start of the apprenticeship program. For the comparison group, enrollment is defined as the start quarter of Wagner-Peyser services, or the first term enrolled at the community colleges in our sample.

[6] It is important to pre-specify primary outcomes to avoid the multiple comparison problem of finding spurious, statistically significant impacts when testing many contrasts (Schochet 2008). Other outcomes will be considered secondary but will be important for contributing to the interpretation of findings on the overall pattern of effects.

(all these fields are self-reported by the participant). To ensure data accuracy, DOL has several data integrity mechanisms in place. A detailed overview of all data elements and their definitions provided by DOL can be found here.

The literature suggests that credible comparison groups for QEDs of workforce development programs require that they be obtained from the same local areas as the treatment group to balance local economic conditions and service environments (Heckman, Ichimura, and Todd 1997; Heckman et al. 1998; Glazerman, Levy, and Myers 2003). The WIPS data contain residential zip codes and county codes for the comparison group. However, the Scaling Apprenticeship and the Closing the Skills Gap grant programs do not require grantees to submit location data to the WIPS, so the county and zip code data would not be available in the WIPS for apprentices. To gather location data on apprentices, we will first request program data from grantees on apprentices' residential zip codes. If we cannot obtain zip codes from grantees, we will instead request information on the zip code for the community college campus where the apprentices take their educational or instructional training or the zip code of their sponsoring employer. In either case, we will ask that grantees link these location data to apprentices' PII or WIPS IDs so that it can be used in the analysis.

Table 4

WIPS Data Items Required for the Evaluation

Data Category	Study Uses
Identifiers for apprentices vs. comparison group	To define the apprentice and comparison samples
Grant-funded apprentice status	
Program in which enrolled (i.e., WIOA Adult, WIOA Dislocated Worker, WIOA Youth, Wagner-Peyser)	
Nongrant apprentice status	
Geographic identifiers	Key information needed to select the apprentice and comparison samples from the same local areas
3-digit county FIPS code	
State code of residence	
Zip code	
Demographics and other characteristics	To construct balanced apprentice and comparison samples
Entry and exit quarters	
Age	
Race and ethnicity	
Disability status	
Education level	
Low-income status	
Industry of employment (incumbent workers only)	

Occupation (incumbent workers only)	
Ex-offender status	
AJC program enrollment and service receipt	To screen the comparison sample and examine differences in AJC service receipt
Dates of AJC services and activities (self-services, staff-assisted services, career services)	
Types and dates of basic career services (individualized, training, and other support services)	

Source: Authors.

Notes: AJC = American Job Centers; FIPS = Federal Information Processing Standard; WIOA = Workforce Innovation and Opportunity Act; WIPS = Workforce Integrated Performance System; Zip = Zone Improvement Plan.

The WIPS data also include people receiving incumbent worker training supported by DOL. However, we do not recommend including these workers in the comparison group because they do not provide a sufficiently strong contrast with the apprentices. An impact evaluation using other DOL incumbent worker services as the counterfactual would measure the impact of different forms of incumbent worker training programs against each other, which is not a treatment contrast that helps to answer a research question of interest for this evaluation. Therefore, we will consider using community college data instead.

Community College Data

Based on other studies using community college data (for example, Anderson et al. 2017), we expect the community college data to have many of the same demographic characteristics as the WIPS data, except for public benefit receipt. Community college data also typically contain reports of Pell grant receipt, an indicator of family income. We also plan to obtain data from assessments of academic skills, either from standardized entrance exams or standardized, state-wide high school accountability tests if assessment scores are not widely available in the community college data.

All community colleges that we target participate in Title IV Federal Financial Assistance programs. Hence, they are obliged to report their institutional data to the Integrated Postsecondary Education Data System (IPEDS) of the National Center for Education Statistics in a timely and accurate manner, including key demographics like students' gender and race/ethnicity. Through their participation with IPEDS, community colleges in our sample have significant experience in preparing institutional data and reports for external use; therefore, we expect they will be able to provide reliable student enrollment data for our study.

ACS Data

The American Community Survey (ACS) is collected by the U.S. Census Bureau and provides annual estimates of income, education, employment, health and living condition for residents of the United States. We will collect county-level data for apprentices and comparison group members from the appropriate ACS three-year sample data to control for differences in county-

level characteristics in the analysis. These characteristics include the unemployment rate, poverty rate, population, median household income, and the share of the population in urban areas.

NDNH Data

The NDNH serves as a legally mandated nationwide database housing employment, unemployment insurance, and quarterly wage data provided by state directories of new hires (SDNH), state workforce agencies (SWA), and federal employers. We will use the NDNH data on earnings and employment to obtain preprogram earnings of apprentices and comparison group members, as well as to measure their earnings and employment outcomes. The NDNH data contain quarterly earnings information collected by all state UI agencies and submitted to the Office of Child Support Enforcement of the U.S. Department of Health and Human Services (Solomon-Fears 2011). The NDNH data contain outcomes only for people with reportable earnings in covered jobs. Thus, anyone in the study sample not found in the NDNH data during a relevant quarter will be counted as not employed and having no earnings in that quarter.

The NDNH data cover most wage and salary employment but do not cover all types of jobs and industries. NDNH data do not cover self-employed workers, railroad employees, workers in service for relatives, most agricultural labor, some domestic service workers, and part-time employees of nonprofit organizations (Czajka, Patnaik, and Negoita 2018). In prior studies, these sectors have made up about 10 percent of U.S. employment (Hotz and Scholz 2001; Kornfeld and Bloom 1999). NDNH data also omit workers whose employers do not report their earnings to their UI agency, even in the formal sector, because of the prevalence of flexible staffing arrangements or illegal neglect of reporting (Abraham et al. 2018; Blakemore et al. 1996; Hotz and Scholz 2001; Houseman 2001; Katz and Krueger 2016, 2019). Additionally, NDNH data do not cover workers who are casually employed, such as day laborers or part-time helpers, and exclude most work that is part of the gig economy (Abraham et al. 2018; Katz and Kruger 2016, 2019). Finally, there could be inconsistencies in reports of social security numbers (SSNs) that lead to inaccuracies in the NDNH.

Data Collection Process

Figure 2 illustrates the data collection process which requires five key steps:

1. Collecting WIPS data from DOL to identify apprentices and comparison group members and their background characteristics (for groups 1 and 2), as well as student records from community colleges (for group 3 and perhaps for group 2, but not for group 1)

2. Collecting geographical data for apprentices from grantees

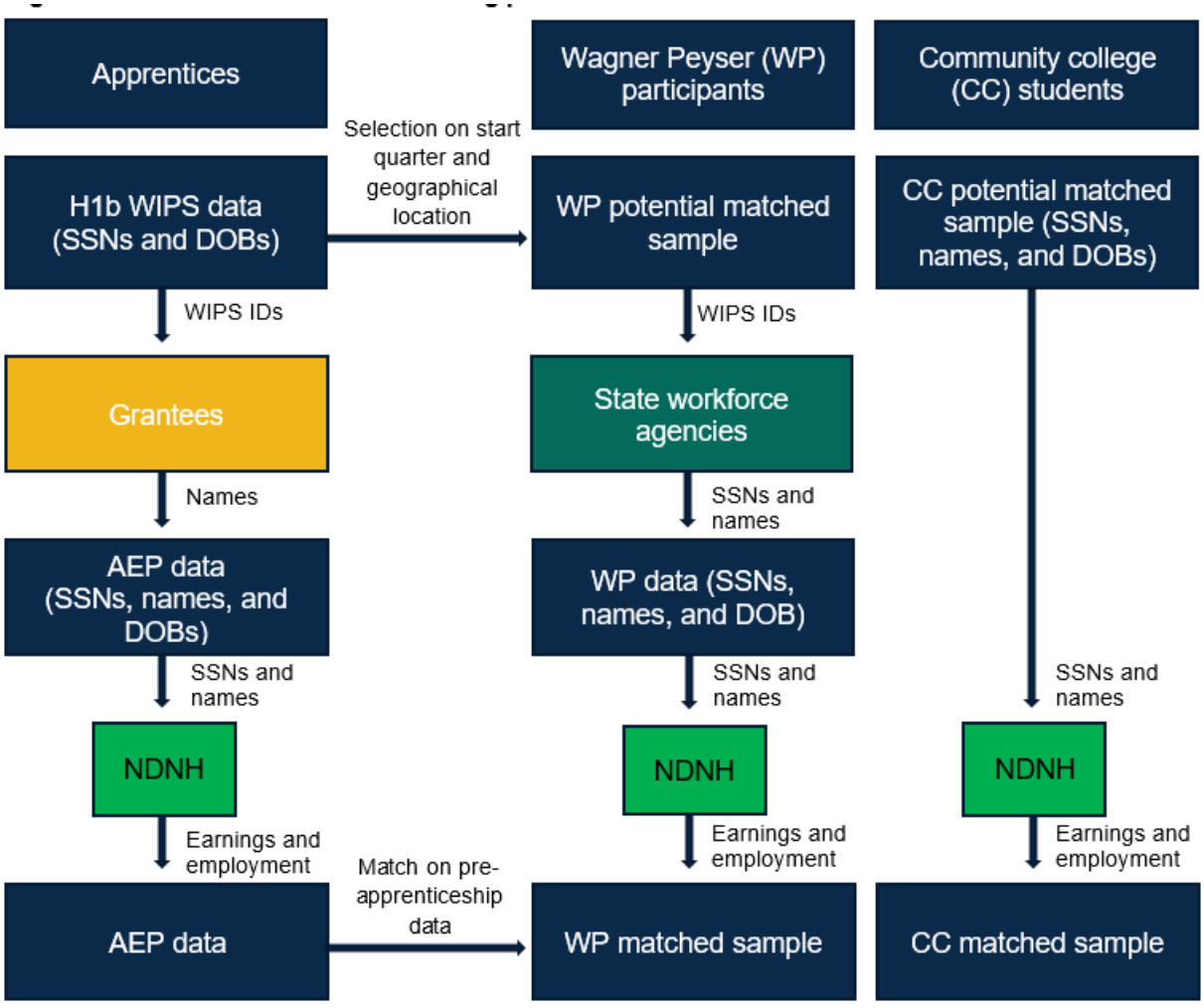
3. Collecting PII on the apprentices and comparison group members in the WIPS data from states who participate in the study (for groups 1 and 2)

4. Using the PII to link records with preprogram and postenrollment data from NDNH

5. Adding county-level characteristics using ACS data

FIGURE 2

Data Collection and Mapping Process by Personal Identifiable Information



Source: Authors' analysis.
Notes: AEP = Apprenticeship Evidence-Building Portfolio; AEP data = Refers to the combined data set the project team will be using; DOB = Date of birth; NDNH = National Directory of New Hires; SSNs = Social security numbers; WIPS = Workforce Integrated Performance System

For ease of exposition, we focus first on the data necessary for groups 1 and 2, and then discuss the additional steps needed for group 3.

Data collection process for apprentices referred from the public workforce system and incumbent workers and their comparison groups (Groups 1 and 2)

(1) Collecting WIPS Data

Using a range of criteria described below, we will submit a request to DOL for program year (PY) 2020, PY 2021, and PY 2022 WIPS data for (1) apprentices served by the grant program (WIPS data contain a field that identifies Scaling Apprenticeship participants) and (2)

nonapprentices that received services under Wagner-Peyser Employment Services or WIOA Adult, Youth, and Dislocated Worker programs. We will request these data as soon as the public (or restricted) data use files become available in mid-2021, mid-2022, and mid-2023, although we will discuss with DOL the possibility of obtaining these data on a quarterly basis if feasible. We will also request the PY 2019 WIPS file as soon as possible so that we can set up our procedures and computer programs to clean the data, set sample restrictions, and identify appropriate comparison groups.

After collecting the WIPS data on apprentices and participants served by WIOA and Employment Services in the study period, we will restrict the treatment and potential comparison group samples based on the following factors:

States with Data Use Agreements (DUAs). First, we will restrict the sample to those states that agree to participate in the study by signing a data use agreement (DUA) for the sharing of participants' PII. As we discuss in the next section, the signed DUA is a necessary condition for obtaining PII and thus being able to obtain outcome data from NDNH. The states included in the study will most likely represent a convenience sample of states those agree to participate in the study. Therefore, study results will not necessarily be representative of the program as a whole.

Apprentices from other programs. Although apprentices served under the grant are necessarily in the treatment group, we may exclude from the sample any potential comparison group members who participated in any nongrant apprenticeship program, using flags in the WIPS data. This restriction will avoid the undesirable situation where apprentices in the grant programs are compared with people who attended other apprenticeship programs. As such, while we are studying the impact of these specific grant programs, the counterfactual condition is defined as the services apprentices would receive if they were not apprentices, rather than the stricter definition of not being grant-funded apprentices.

Enrollment period. We will restrict the sample based on when apprentices and comparison group members enrolled or started receiving services to ensure a sufficient follow-up period for NDNH data collection. The study sample will consist of apprentices that started between July 2020 and June 2023. The comparison groups will be Wagner-Peyser participants (group 1), incumbent workers served by the Wagner-Peyser and the WIOA Adult, Dislocated Worker, and Youth program (group 2), and community college students (group 3) with similar characteristics to the treatment group.

(2) Collecting Geographical Data from Grantees

To avoid comparing apprentices concentrated in certain counties and zip codes to group members in different areas, which could lead to differences in the availability of other services and labor market conditions, we will restrict the comparison group to people living in areas in which at least one member of the apprentice group lives. Since WIPS data does not provide geographical data on apprentices, this will be based on either zip code or county of residence for

an apprentice or the community college campus or employer for an apprentice, depending on the data provided to us by grantees.

(3) Collecting PII from States

To obtain NDNH data for the individuals in our treatment and comparison groups, we need to submit their SSNs and names (PII) to Office of Child Support Enforcement (OCSE). However, the WIPS data only contain PII for apprentices—not for Wagner-Peyser participants. Therefore, we will need to obtain information on the first name, last name, and SSN of sample members from the study states.

We will negotiate DUAs with states to participate in the study by sending us the PII for the sample we select from the WIPS. We will place particular emphasis on states containing grantees with large numbers of apprentices, which we will identify from grantees' quarterly reports. We will prioritize states with which the study team has existing DUAs from other projects and will seek to amend those DUAs to incorporate the samples for this study. We will also request and analyze PY 2019 WIPS data to provide information on the grant programs for planning and recruiting purposes. Once executed, the DUA agreements will allow us to send the set of state-provided identifiers from our WIPS sample to each state and receive a file with the PII for each sample member.

(4) Collecting NDNH Data

Once our WIPS sample is linked to the PII we receive from states, we can send several data requests to NDNH to retrieve preprogram and follow-up earnings and employment data. The sample will include those who enroll in WIPS programs from July 2020 through June 2022. The timeline of the study will allow all sample members to have at least 8 quarters of follow-up data, and more than 12 quarters for some, depending on when they enrolled in the program. For apprentices, the first quarter of enrollment would be the quarter they were hired as apprentices or the quarter they began educational or instructional training, whichever came first (both dates are recorded in the WIPS). However, we will exclude from the sample apprentices that started educational or instructional training more than a quarter before they were hired. For comparison group members from the WIPS, the first quarter of enrollment would be the quarter they started receiving public workforce services. For community college students, we will use the start of their enrollment in the first of these courses as their program enrollment date.

Our primary employment and earnings outcomes will use the ninth and tenth quarters of available postenrollment data to balance the objectives of maximizing sample size and setting the latest follow-up period possible. We will examine two quarters to smooth anomalous findings that could occur if only one quarter was used. It is important to prespecify primary outcomes to avoid the multiple comparison problem of finding spurious, statistically significant impacts when testing many contrasts (Schochet 2008). Other outcomes will be considered secondary but will be important for contributing to the interpretation of findings on the overall pattern of effects.

NDNH deletes data older than eight quarters from its system. Thus, we cannot obtain preprogram data at the same time that we request outcome data. Instead, we will need to make early NDNH data requests for preprogram earnings soon after we receive the WIPS data to ensure the preprogram period is fully covered. We will submit requests for baseline data to NDNH in Q3 2021 for the PY 2020 sample and in Q3 2022 for the PY 2021 sample. This will provide at least four quarters of preprogram earnings for all sample members.

If we are unable to obtain four quarters of preprogram earnings data from the NDNH because we do not receive PII data from the states quickly enough, we will instead ask the states to provide these UI wage records directly. This is not as ideal as receiving data from NDNH, because it will be more costly to manage this data access process with multiple states rather than the single database of NDNH. In addition, state UI wage records would not contain preprogram earnings for people who worked in other states before enrolling. However, if we need to get preprogram earnings data directly from states, we may be able to obtain more than four preprogram quarters.

(5) ACS Data

In addition, we will also incorporate the county-level characteristics obtained from the appropriate ACS three-year sample data.

Data Collection Process for Community College Data

(1) Collecting community college data

For comparison group 3, we will need community college data. To obtain these data, we will negotiate DUAs with community college systems that serve as Scaling Apprenticeship and Closing the Skills Gap grantees. We will only request the data if the community colleges can provide PII, including SSNs. Community colleges generally have these data because they need them for purposes of student financial aid. The community college data itself will not contain information that will allow us to identify apprentices. However, we will request SSNs from the grantees and link them to the community college data to identify apprentices served by the grants.[7] If we are able to obtain community college data in a state in which we also have a DUA for the analysis of group 1 and 2 apprentices, we will attempt to identify whether any of the comparison group members in the community college data actually participated in nongrant apprenticeships in the WIPS data, and remove them from the sample.

After collecting community college data, we will restrict the treatment and potential comparison group samples based on the following factors:

[7] Alternatively, we can also link the community college data with SSNs to the WIPS data, which has SSNs for grant-supported apprentices. However, this data is available with a lag, and it would be more efficient to collect this data from the grantees.

Enrollment period. We will restrict the sample based on when apprentices enrolled and when comparison group members started taking the courses related to the apprenticeship to ensure a sufficient follow-up period for NDNH data collection.

Geographic region. If we can gather data on geographic location from all community colleges, we may restrict the data to students that have nonmissing information on location.

Preprogram earnings and employment. We will attempt to acquire data on preprogram earnings from the NDNH data. For the reasons discussed above, we cannot wait to obtain preprogram data while we request outcome data. We will submit requests for baseline data to NDNH in Q3 2021 and Q3 2022. This will provide four to eight quarters of preprogram earnings for all sample members, unless we acquire the earnings records directly from states.

Standardized test scores. We will negotiate a DUA with the associated state department of education to acquire data on standardized tests taken during high school, such as the Scholastic Assessment Test (SAT), American College Testing (ACT), or other state-administered tests. We will use the PII collected from the community college system and the states to link to the state department of education data system. As available, we will use standardized test score data for all three groups to generate a comparison group that is similar to apprentices in order to estimate credible impacts. These data will be especially important for younger members of the sample, such as those that matriculate in community college directly after high school and are likely to have a shorter duration of reported earnings in the preprogram period.

(2) Collecting NDNH Data

We will first identify appropriate comparison group samples using the available community college demographic data (absent the preprogram earnings). We will then submit the PII we collect from the community colleges to NDNH to obtain preprogram earnings and employment data and use the NDNH data to generate the comparison group weights.

Item 4 – Response rates and attrition*. Describe methods to maximize response rates and to deal with issues of nonresponse. The accuracy and reliability of information collected must be shown to be adequate for intended uses. Describe potential selection or response rate issues and other potential sources of bias, and resulting limitations for analyses, including limitations related to the ability to examine specific subpopulations of interest (e.g., disaggregation by gender, ethnicity, race, etc.). For collections based on sampling, a specific justification must be provided for any collection that will not yield "reliable" data that can be generalized to the universe or population of interest.*

The population of interest for this study consists of participants in the apprenticeship programs offered by the Scaling Apprenticeship and Closing the Skills Gap grantees and members of three comparison groups. The study uses administrative data and does not depend on individual responses to primary data collection efforts or sampling. However, the factors that follow could lead to a smaller sample size than the full population of interest.

Missing Data on Key Background Variables

The WIPS data are generally of high quality, and most key variables have no missing values. This is likely due to reporting requirements and WIPS-integrated data checks on required variables. Even still, our experience with the WIPS data suggests that some individuals may have missing data on key variables. For example, our experience suggests that about one percent are likely to be missing geographic identifiers. Given the importance of these variables for estimating the propensity score, we will exclude from the analysis anyone missing geographic data. We will also exclude individuals with missing data for variables that identify whether individuals were in an apprenticeship program or received the services that qualify them for the comparison group. Similarly, we will exclude community college students with missing data on key variables.

PII Data from State Workforce Agencies

We rely on state workforce agencies' willingness to provide us with PII for the Wagner-Peyser comparison group members. We anticipate recruiting ten states and will place particular emphasis on states in which grantees with many apprentices are located. Our power calculations in item 5 (Sampling and Power Analyses) show how different sample sizes could look like and which statistical power they could provide. Although our sample is purposeful and will yield meaningful impact estimates, we cannot formally generalize our findings to the full grant programs because they only include a subset of grantees and participants.

Another hurdle is that participant identifiers in state workforce data systems differ from those in the WIPS. This mismatch occurs because participant identifiers in state systems are scrambled when they are submitted by states to DOL. Thus, DOL must provide a crosswalk between the scrambled identifiers in WIPS and the true identifiers in state systems. With this crosswalk, we can send states the applicable set of state identifiers for the study sample and then obtain PII

back from the states. This will allow us to submit an analysis file with PII to get access to the NDNH data.

Geographic Characteristics

For apprentices, we need to collect apprentices' names and residential zip and/or county information from the Scaling Apprenticeship and the Closing the Skills Gap grantees. If we are unable to retrieve the geographical data for some individuals, we may restrict the sample to apprentices with nonmissing location data to account for differences in labor market conditions between treatment and comparison group members. In addition, we will use the county-level characteristics obtained from the appropriate ACS three-year sample data.

Timing of NDNH Data Availability

NDNH deletes data older than eight quarters. The later we submit PII to NDNH to request earnings and employment data, the fewer apprentices we can include in our study. In addition, we will have only four quarters of preprogram earnings for some individuals, a relatively short period. The literature clearly identifies the value of longer earnings windows for reducing selection biases when performing matching exercises for workforce programs (Heckman and Smith 1999; Mueser, Troske, and Gorislavsky 2007), and DOL's own systematic reviews of research reflect this finding. Specifically, DOL's CLEAR review protocol for studies in the Employment and Training Topic Area requires that studies demonstrate similar preprogram earnings from "greater than one year before program participation." Therefore, we will create measures of employment dynamics that more completely account for preprogram employment.

First, we will use the full dynamics of four quarters of preprogram employment as participant characteristics in the propensity score model. Specifically, we will create indicators for every possible combination of employment statuses over four preprogram quarters (such as (0,0,0,0), (1,0,0,0), (1,1,0,0) and so on)[8], for a total of 16 employment history types; we will use these indicators as characteristics. This is important because it will help us to capture the steep drop in earnings right before program entry (for nonincumbent workers) that other research has demonstrated often exists for individuals enrolling in workforce programs (Ashenfelter 1978; Heckman and Smith 1999). The approach of using a full history of employment dynamics was first implemented by Card and Sullivan (1988) using annual earnings. More recently, Mueser, Troske, and Gorislavsky (2007) included dynamics with a limited set of four binaries to capture employment dynamics over eight quarters, while Dolton and Smith (2011) implemented a matching approach after stratifying their population on a full year of program benefit dynamics broken into six-week intervals. Each of these studies found this to be an important strategy, despite it not being widely used. Second, we will include job switch indicators capturing individual job turnover, which we will generate from the masked employer IDs in the NDNH

[8] Zero stands for a quarter in that an individual was unemployed, while one stands for a quarter in that an individual was employed.

data. These indicators help represent job stability and employment disruptions prior to program entry.

Item 5 – Sampling and Power Analyses. *Describe (including a numerical estimate) the sampling frame and any sampling or other respondent selection method to be used. Describe the procedures for the collection of information including statistical methodology for stratification and sample selection; estimation procedure; degree of accuracy needed for the purpose described in the justification; unusual problems requiring specialized sampling procedures. Data on the number of entities (e.g., establishments, State and local government units, households, or persons) in the universe covered by the collection and in the corresponding sample are to be provided in tabular form for the universe as a whole and for each of the strata in the proposed sample. Indicate expected response rates for the collection as a whole. If the collection had been conducted previously, include the actual response rate achieved during the last collection. Include clear description of groups to be studied or compared and anticipated sample sizes. Also outline power calculations that align with each hypothesis to be tested to clearly demonstrate sufficient sample to examine the primary research questions with the selected methodology.*

The study relies on obtaining PII data from state workforce agencies and community colleges for comparison group members to request their earnings and employment data from NDNH. Since these data are administrative, selecting which sample members are included in the study depends on states' and community colleges' willingness to share data. The population of interest in the selected states will include apprentices who enrolled in one of *the* Scaling Apprenticeship or Closing the Skills Gap grantee programs between July 2020 and June 2022, along with comparison group members who enrolled either in Wagner-Peyser and the WIOA Adult, Dislocated Worker, and Youth programs or at a community college during the same period.

To guarantee that the study's sample size is sufficient to estimate statistically significant effects, we estimated minimum detectable impacts (MDIs) for the primary outcomes, earnings and employment. First, we present the MDIs for the evaluation of unregistered apprenticeships, pooling together the Scaling Apprenticeship and Closing the Skills Gap grantees. The MDIs reflect a number of key assumptions that the study team made:

1. Although it is reasonable to expect that we will have more comparison group members than apprentices because WIOA and Wagner-Peyser serve more people than apprenticeship programs, to be conservative, we assume that the comparison sample will be the same size as the apprenticeship sample. This assumption, even if conservative, does not have important implications because increasing the size of just one of the two groups has a limited effect on statistical power.

2. For each of the grantees shown in tables 1 and 2, we assume that if they plan to have both registered and unregistered apprentices, then apprentices will be equally divided between the two.

3. We assume that we will not be able to obtain data on all apprentices, due to sample restrictions and an inability to obtain data from all states and community colleges

targeted for outreach. The other rows project smaller and larger sample sizes, to account for the possibility that grantees do not achieve or exceed their targets, more or fewer states execute DUAs, or there is greater or less sample loss due to restrictions and missing data.

Table 5 displays the MDIs as a function of the sample size for the evaluation of unregistered apprenticeships. The table displays MDIs for all groups combined, for smaller sample sizes to provide more conservative estimates if we collect fewer data, and for one of the three apprentice groups (those referred to the program from the public workforce system, incumbent worker apprentices, and apprentices who are community college students prior to enrollment) that represent one-third of the overall sample. We do not show MDIs for individual programs, because sample sizes will typically be too small to yield precise estimates at the site level. However, the MDIs by apprentice group pertain also to MDIs for subgroup analyses where individuals or programs are grouped for analysis.

Table 5

Minimum Detectable Impacts (MDIs) for Impact of Unregistered Apprenticeship

Sample size (apprentice and comparison group)	Average quarterly earnings (MDI)	Employment (MDI; percentage points)
12,147 (target, average across three groups)	$141	2.1 pp.
16,000	$123	1.8 pp.
6,000	$201	3.0 pp.
4,049 (target, single group)	$244	3.6 pp.

Source: Authors' calculations.

Notes: Assumptions include equal sample sizes between apprentice group and comparison group; $3,102 standard deviation of earnings (based on WIA core services group in the 9th and 10th quarters after random assignment, personal communication from Dana Rotz); 70 percent employment rate for comparison group (based on WIA core services group in the 9th and 10th quarters after random assignment in Rotz et al. 2017); covariates explain 20 percent of the variation in outcomes; alpha level 0.05, two-sided test, 80 percent power.

Even with substantial drops in the sample that could be caused by grantees having fewer apprentices than they targeted—perhaps as a result of the COVID-19 pandemic and its effects on the economy—and some states and community colleges not agreeing to provide the necessary PII, the MDIs at the group and aggregate levels would still be much lower than $1,460, the estimated impact of registered apprenticeships on quarterly earnings at the ninth quarter after enrollment found in Reed and colleagues (2012).

Table 6

Minimum Detectable Impacts for Impact of Registered Apprenticeship

Sample size (apprentice and comparison group each)	Average quarterly earnings (MDI)	Employment (MDI; percentage points)
37,175 (target, average across three groups)	$81	1.2 pp.

45,000	$73	1.1 pp.
18,000	$116	1.7 pp.
12,392 (target, single comparison group)	$140	2.1 pp.

Source: Authors' calculations.

Notes: Assumptions include equal sample sizes between apprentice group and comparison group; $3,102 standard deviation of earnings (based on Workforce Investment Act [WIA] core services group in the 9th and 10th quarters after random assignment, personal communication from Dana Rotz); 70 percent employment rate for comparison group (based on employment rate of WIA core services group in the 9th and 10th months after random assignment in Rotz et al. 2017); covariates explain 20 percent of the variation in outcomes; alpha level 0.05, two-sided test, 80 percent power.

Next, we present the MDIs for the evaluation of registered apprenticeships in table 6. We make the same assumptions as we do for unregistered apprenticeships. However, based on our assumptions and grantee reports of target apprentices served, more apprentices will be served through registered apprenticeships and unregistered apprenticeships, leading to the larger sample size. The MDIs for registered apprentices are even smaller than those calculated for unregistered apprenticeships (table 5). As such, there is ample power to detect impacts that are much smaller than what have been found in the literature (for example, the impacts estimated in the Reed et al.'s 2012 study are $1,460 higher quarterly earnings and 8.6 percentage points employment increase for registered apprentices), even if there are substantial reductions in the sample size from the most optimistic scenario.

Item 6 – Analyses. *Outline key models, plans for tabulation, coefficients, tables and descriptive statistics. Outline methodological approaches for regressions and other analytical methods selected by research question and hypothesis. Cite relevant literature for models used or otherwise outline the basis for the specific analytic approach. Address any complex analytical techniques that will be used. Describe how the data will be prepared and analyzed. Specify what data will be removed from final reporting due to disclosure risks. Outline dummy variables, coefficients or table cells that will be included in final public reporting (as well as those that may be removed due to disclosure risk).*

The parameter of interest for the impact estimation is the average treatment effect on the treated (ATT). That is, we are estimating the average difference between the apprentices' outcomes and what the apprentices' outcomes would have been in the absence of the apprenticeship programs.[9] We will estimate impacts separately for registered and unregistered apprenticeship programs. Our primary estimation approach for both impact estimations will be inverse probability weighting (Chesnaye et al. 2022; Horvitz and Thompson 1952), but we will examine whether the impact estimates are robust to using methods that match on the propensity score instead. We discuss two key features of the impact analysis: (1) estimating the propensity score and (2) using the propensity score to estimate impacts.

Estimating the Propensity Score

The propensity score is the predicted probability of being in the treatment group, conditional on preprogram characteristics (Rosenbaum and Rubin 1983; Dehejia and Wahba 1999, 2002). To estimate the propensity score, researchers often estimate a regression, where the outcome is the treatment indicator, and the predictors are preprogram characteristics that influence outcomes and program enrollment decisions. The predicted probability of treatment from this regression, for each individual, is the propensity score.

One key concern regarding propensity score estimation is that the specification of the propensity score model can influence the impact estimates (Drake 1993). A number of newer methods have been suggested to select propensity score models based on the data, without overfitting. Our main approach to estimating the propensity score will be the generalized boosted regression, which is a procedure based on logistic regression that searches over a core set of provided covariates to create new partitions and interactions that most predict participation (McCaffrey et al. 2004), with out-of-sample predictions used to prevent overfitting. We will also use least absolute shrinkage and selection operator (LASSO) linear probability regression models that search over a large set of covariates to identify those that best predict participation, subject to a regression penalty for overfitting (Tibshirani 1996). Finally, we will also use simple logistic regression as a baseline strategy for comparison. We will estimate separate propensity score

[9] We focus on estimating the partial equilibrium effect of registered apprenticeship, holding other effects of apprenticeship on labor market dynamics constant. General—as opposed to partial—equilibrium effects would account for the (assumed) downward effect on the price of the apprentices' skills generated by the increased supply of apprentices with these skills, among other factors. We do not plan to estimate general equilibrium effects.

models (and impacts) for groups 1, 2, and 3 because their populations and program services could differ, leading to different program effects.

Although generalized boosted regression generates and includes interactions and higher-order terms based on an algorithm, the LASSO selects among variables identified by the study team, and the logistic model uses the full set of variables. For the logistic model, we will use the main covariates associated with assignment to treatment discussed in table 3, with a cubic specification for preprogram earnings. For the LASSO, we will specify this same set of variables, *K*. Once the LASSO selects covariates from this set, *J*, a second LASSO specification will include both *J* and the interactions between *J* and *K*.

To ensure that the selection of a propensity score modeling approach is not influenced by the impact estimates they lead to, we will select an estimation approach through empirical exercises on the data before they are linked to the NDNH outcome data (although we will use NDNH preprogram data for the exercise).

The literature suggests that comparison and treatment groups should be selected within the same local areas because the availability of services and labor market conditions vary geographically (Glazerman, Levy, and Meyers 2003; Heckman, Ichimura, and Todd 1997; Heckman et al. 1998). At the same time, we are concerned that estimating separate models by local area will lead to model overfitting and instability because some programs may not have large apprentice samples (and this approach could be expensive to implement). To balance these objectives, our plan is to estimate a unified model across grantees within states, where the models include both demographic and county-level variables. We will then use the estimated propensity scores to identify comparison groups within the same local areas and to also compare their quality to those obtained across areas. Because of larger potential comparison groups, it might be the case that the weighting across area yields more balanced comparison groups on demographic and employment history variables at the expense of balance on the county variables. On the other hand, available comparison group samples might be sufficient to generate high quality within-area weighting (the preferred approach).

We will base the selection of the propensity score model on how well it generates treatment and comparison samples that are balanced on preprogram characteristics. We will focus on the standardized difference in each characteristic (that is, the effect size), but also present two-tailed *p*-values resulting from *t*-tests.[10] The four summative measures across characteristics will be:

1. Average absolute value of the standardized differences: provides an overall sense of balance.

[10] We acknowledge that the *t*-test may not be the best measure of similarity due to statistical significance being directly tied to sample size (Imbens and Rubin 2015), but we include it for its familiarity.

2. Share with an absolute value of the effect size greater than 0.25 standard deviations: a relevant threshold for the CLEAR standards.

3. Share with an absolute value of the effect size greater than 0.10 standard deviations: generates a higher standard for identifying credible matches than the 0.25 threshold.

4. Share of characteristics with *p*-value less than 0.05: provides a traditional sense of statistically-significance differences.

Using the Propensity Score to Estimate Impacts

After selecting the propensity score estimator that generates the greatest balance between the apprentice and comparison group, we will use the estimated propensity scores to estimate impacts. We propose the inverse probability weighted (IPW) estimator as our main approach because it can make use of a larger sample and thus is generally more precise than a matching estimator. Indeed, Hirano, Imbens, and Ridder (2003) show that IPW can reach the theoretical efficiency bounds identified by Hahn (1998), and this is confirmed in practice by the improved precision that is demonstrated by the empirical Monte Carlo studies—particularly when overlap of the propensity score is considerable. Under our main approach, we will first trim the sample to protect against the influence of scores at extreme values. Specifically, we will implement the approach suggested by Crump et al. (2009) whose study suggests dropping all units with propensity scores below 0.1 and above 0.9 as a selection rule.

Next, we will generate weights, equal to one for all apprentices and equal to a scaled version of $\hat{p}_i/(1-\hat{p}_i)$ for the comparison group, where $\hat{p}_i$ is the estimated propensity score. For within-area analyses, the propensity score model is still estimated across areas, but the comparison group weights are adjusted to sum to a constant within each local area.

Finally, we will estimate impacts using weighted least squares, and use generalized method of moments (GMM) to account for estimation error in the propensity score (Abadie and Imbens 2016). That is, once the generalized boosted regression or LASSO procedure selects a propensity score model that generates the greatest balance, we will estimate that propensity score model jointly with the impact model. This allows the estimated standard errors of the impact estimates to account for the estimation error in the propensity scores, but not the estimation error associated with selecting the propensity score model itself. We will generate doubly robust impact estimates by including the background characteristics described above in the impact model to account for the remaining differences between the apprentice group and comparison group, and to account for variation in the outcome (Bang and Robins 2005).

We might expect that program effects could vary across groups 1, 2, and 3 due to differences in background characteristics, prior work experience, and service needs. Further, the received service contrasts between the treatment and comparison groups could also differ across the groups—such as the type of occupational training, the duration of program participation, and the balance of educational or instructional training versus OJT. Therefore, we will present the

impacts for each group separately. We will also construct an average impact using the number of apprentices in each stratum as a weight. We will use the number of apprentices in the stratum sample as the weight because doing so approximates the impact we would have estimated if all apprentices were in one group, rather than three separate groups.

We will attempt to generate the strata to be independent (that is, no apprentices or comparison group members will be in multiple strata), and thus will not need to account for correlation in the impact estimates when calculating the standard error of the average estimate. If we are unable to do so, however, we will use generalized method of moments in estimating the average impact estimate, which allows to account for the correlation induced by having overlapping samples when estimating the standard error of the average impact. Finally, we will also compare the impacts for these three groups to one another as part of the subgroup analysis, although power may be limited for these analyses. For this reason, subgroup analyses may need to be interpreted with caution.

Sensitivity Analysis

As discussed, IPW will be our main approach, largely due to statistical power gains relative to other approaches. However, as a robustness check, we will examine the sensitivity of the impact estimates to two approaches to matching propensity scores: (1) nearest-neighbor one-to-one match with replacement and (2) caliper matching.

Nearest-neighbor matching identifies a comparison group by selecting a single comparison group member with the closest estimated propensity score for each apprentice. Each apprentice is included with a weight of one and each selected comparison member is weighted by the number of times they are selected. Weights for the comparison group are then normalized to sum to one (Imbens 2015). It is considered the baseline approach because it is simple and has been shown to have the smallest bias across a range of approaches. When implementing this strategy, we will correct the standard errors for the estimation error that is introduced from the matching procedure (Abadie and Imbens 2008), and from the estimation of the propensity score.

Caliper matching works by selecting all comparison group members within a given distance as representing the comparison group for each apprentice. Comparisons of outcomes are then made across the groups while modeling the influence of covariates (Imbens and Wooldridge 2009). This is the primary strategy selected by Heinrich et al. (2013) when estimating the causal impact of Workforce Investment Act (WIA) training and other workforce programs. The caliper is measured by the log-odds ratio, which has the benefits of linearizing distance across the distribution (Smith and Todd 2005), and weights are additively assigned to comparison group members as their representative fraction when matched to each apprentice. Although there are sophisticated options for selecting the caliper (Galdo, Smith, and Black 2008), we rely on the suggestion of Lechner et al. (2011) of using the greatest distance in log-odds ratios of any of the nearest-neighbor matches. Thus, for each apprentice, we will select all comparison group members with propensity scores (in log odds) within the pre-specified bandwidth.

Finally, to examine the degree of selection on observables that would be necessary to change the conclusions of the impact estimates, we will formally assess sensitivity using the method introduced by Rosenbaum (2002). This method allows one to assess selection in QEDs without specifying a particular structure for the unknown selection parameter. We will use available software packages to implement the analysis.

Subgroup Analysis

A key part of the evaluation will be to conduct analyses to examine what works and for whom. This information is valuable for purposes of ongoing program improvement and targeting services appropriately. To address these research questions, we will estimate impacts for subgroups defined by key individual baseline characteristics and key program services received by the apprentices.

Individual Baseline Characteristics

We will examine impacts on the following subgroups defined by individual and local area characteristics:

- **Underserved groups.** The Scaling Apprenticeship and Closing the Skills Gap grants place an emphasis on serving populations that have been traditionally underrepresented in apprenticeship programs, such as veterans, individuals with disabilities, women, people of color, and ex-offenders. All these characteristics are available in the WIPS data and many will generally exist in community college data as well.

- **Age.** The effects of apprenticeships may differ by age for a number of reasons. Younger workers typically have less work experience than older ones and may have fewer skills. At the same time, younger individuals may be more eager to seek additional education and training services and have a longer time horizon to benefit from program services. We will estimate impacts for those younger than 23, those that are 23 to 30 years old, and those that are older than 30.

- **Education level.** Apprentices with different preprogram education levels may benefit differently from the programs because of varying skills and job readiness. We will estimate impacts for those with a high school credential or less, as well as those with any postsecondary education.

- **Recent employment experience.** Job readiness, marketability, and motivation to work may be greater for those who worked near the time of entry than for those with less labor market involvement, suggesting that program impacts could differ based on recent employment experiences. We will estimate impacts for those never employed in the prior three quarters, versus those who worked in any of the previous three quarters.

- **The local unemployment rate.** Job opportunities for apprentices may be greater in local areas with lower unemployment rates than higher rates. However, the same may be true for the comparison group members. Thus, an important empirical question is whether a stronger economy is associated with larger or smaller effects of program participation.

We will estimate impacts for these subgroups by modifying the IPW estimation approach to include terms formed by interacting subgroup indicators with the treatment status indicator in the outcome model, and using F-tests to assess whether differences in impacts across subgroup levels are statistically significant. For example, to assess whether impacts are larger for apprentices older than age 30, we will interact this indicator with the treatment status indicator and include it as a covariate in the regression models.

Services Received by the Apprentices

We will explore collecting management information system data from the grantees on program services received by their participants. If these data are available, of sufficiently high quality, and collected somewhat consistently across grantees, we can create indicators of key dimensions of service receipt for apprentices in the study sample. We could then estimate subgroup impacts by comparing the outcomes of apprentices who received a particular array of services to those of their comparison group.

Service indicators could measure occupation, time spent in the program, service mix environment (for example, classroom or OJT), and the receipt of supportive or case management services in addition to occupational training. We will explore using cluster analysis to group program features that are correlated with each other to reduce the dimensionality of the subgroup analysis that, if successful, can help improve the interpretation of the findings. In standard subgroup analyses, researchers examine each subgroup in isolation without regard to the other subgroups. But these analyses can be difficult to interpret if the subgroups are related to each other, and they can also suffer from the multiple comparisons problem, where the chances of finding spurious significant subgroup effects increase substantially when many hypothesis tests are conducted across many subgroups. For example, apprentices might stay longer in occupations that require more extensive training. Cluster analysis can help address these issues by using statistical algorithms to form "clusters" or "typologies" that group together similar variables. In essence, the cluster analysis forms groups so that the "distance" between the variables within the clusters is much smaller than between the clusters. If the formed clusters have policy relevance, this approach can lead to more focused and rigorous subgroup analyses. We will discuss with DOL and the technical working group (TWG) the choice of program features to enter the cluster routines (which could also include demographic variables), and the promise of this approach using the resulting typologies formed by the cluster routines. For the WIPS sample, we will link apprentices to specific programs using the grantee identifier, variable "Grantee ID," in WIPS. If these data are missing for many grantees, we can link an apprentice to the nearest grantee in their county of residence. Linkages will be straightforward for the samples obtained using community college data.

To obtain weights for IPW estimation for these subgroup analyses, we will estimate separate propensity score models for each service group, where the dependent variable will be "1" for apprentices who received a particular service and "0" for all comparisons. This approach will generate a new set of weights for each service group analysis. A simpler approach will be to use nearest-neighbor matching, where treatments in a service category are compared with their matched comparison group members. We will assess the merits of both approaches using the data.

We will interpret the impact estimates for subgroups defined by service receipt with greater caution and caveats than the main impacts estimates and the estimates for other subgroups because there is an additional, unobserved selection mechanism for certain aspects of service receipt. For example, even conditional on being hired as an apprentice, those apprentices that spend a longer time in the program may be more motivated or have higher noncognitive skills than those who exit earlier (Heckman and Rubinstein 2001), and we do not have additional baseline characteristics to account for these differences.

Interpretation of Findings

We will report the regression-adjusted mean differences between the treatment and comparison groups and impact estimates with standard errors. We will also conduct a statistical significance test at the 0.05 level to evaluate the impact estimates. For transparency, we will report p-values rather than simply indicating if an estimate was statistically significant at the 0.05 level. In the presentation of our findings, we will prioritize the primary research questions and describe exploratory or secondary analyses as suggestive. We will provide a comprehensive interpretation of our findings, including a detailed discussion about the similarities and differences in services received by apprentices and comparison groups.

In our public reports, we will consolidate the summary statistics and avoid including small cell sizes that could potentially reveal identifiable participant information. However, apart from this, we do not foresee any exclusion or removal of data based on the risk of disclosure.

Item 7 – Expert and stakeholder inputs. *Include a description of a process for soliciting input and feedback through peer review, technical working groups, and/or other consultation from independent, unbiased experts.*

The impact study has been and will be informed by input and feedback from independent, unbiased experts in the subject matter of the study (apprenticeship training) and in experimental and nonexperimental impact study methods. Methods for gathering expert and stakeholder feedback include the project's TWG; review of study publications by stakeholders in the CEO and the Employment and Training Administration; and review by Urban Institute experts outside of the project team.

The Apprenticeship Evidence-Building Portfolio project organized a TWG in May 2020 to provide feedback on all aspects of the evaluation, including the final version of this Evaluation Design Plan. The TWG includes the following five individuals who have a wide variety of skills and backgrounds useful for providing input to the study in the areas discussed above.

1. Susan Helper, Professor of Economics, Case Western Reserve University,[11] with relevant expertise in apprenticeship and workforce development. Dr. Helper is an expert on the globalization of supply chains and on how U.S. manufacturing might be revitalized. She was formerly Chief Economist at the U.S. Department of Commerce and a member of the White House Staff. Having written on the subject, Dr. Helper will bring to the study the business perspective of the benefits and costs of apprenticeships.

2. Chris Magyar, Chief Apprenticeship Officer, Techtonic,[12] with relevant expertise in apprenticeship/workforce development. Mr. Magyar is responsible for the overall operations and strategy of the Techtonic Apprenticeship Program. In 2016, Techtonic launched its apprenticeship and became the first company in the State of Colorado (and one of the first in the country) to establish a DOL-registered apprenticeship for software development. Mr. Magyar's employer perspective will be valuable to the study.

3. Mary Alice McCarthy, Director of the Center on Education and Skills, New America,[13] with relevant expertise in apprenticeship/workforce development and implementation evaluation. Dr. McCarthy is an expert in higher education, workforce development, and job training policies. She also has expertise in providing technical assistance having led technical assistance initiatives in career pathways, credentialing, and competency-based education. She formerly worked at DOL and the Department of Education. She also wrote policy guidance on credentialing and career pathways and supported the Trade Adjustment Assistance Community College and Career Training (TAACCCT) and Workforce Innovation Fund grant programs. Her experience could play an important role in the implementation evaluation and understanding of the findings.

[11] For more information on Susan Helper, see https://faculty.weatherhead.case.edu/helper/.
[12] For more information on Chris Magyar, see https://www.air.org/experts/person/chris-magyar.
[13] For more information on Mary Alice McCarthy, see https://www.newamerica.org/our-people/mary-alice-mccarthy/.

4. Ron Painter, CEO, National Association of Workforce Boards,[14] with relevant experience in apprenticeship/workforce development. Mr. Painter is an expert in workforce development programs and would bring his practitioner perspective to the study. Prior to the NAWB, he was the founding CEO of the Three Rivers Workforce Investment Board in Pittsburgh. His background and contacts with community colleges, Workforce Investment Boards, and employers could play an important role in the implementation evaluation and understanding of the findings.

5. Jeffrey Smith, Professor of Economics and Applied Econometrics, University of Wisconsin-Madison,[15] with relevant experience in impact evaluation. Dr. Smith is a recognized expert in experimental and nonexperimental methods for the evaluation of interventions. Over his career, he has consulted with governments in the U.S., Canada, the U.K., and Australia on evaluation issues. Dr. Smith is particularly well positioned to advise on creative design strategies to overcome common evaluation challenges.

The responsibilities of the TWG include reading study design reports for the Apprenticeship Evidence-Building Portfolio and providing written and oral commentary and criticism of the team's design. The mix of subject matter and methodological experts on the TWG helps to ensure that all dimensions of study design are considered. In the initial meeting of the TWG on the impact study design, members provided feedback related to the importance of a rigorous counterfactual for the matching design and key variables for baseline balance. TWG members also emphasized the importance of learning about the impacts of unregistered apprenticeships and apprenticeships in nontraditional occupations. All study design reports will be revised in response to TWG member comments.

Although the TWG's purpose is to review design plans and decisions, other reviewers external to the study team will provide comments on study publications. These external reviewers will ensure that all study publications are rigorous and present results in a way that is consistent with the evidence. Similar to the TWG, external reviewers will bring substantive and methodological expertise. Two types of external reviewers will review and comment on all study publications. First, staff from the CEO and the Employment and Training Administration will provide several rounds of review, until a research product is determined to be of publishable quality by the COR. Second, the Urban Institute has a policy of engaging colleagues external to the project to provide peer review for all publications.

[14] For more information on Ron Painter, see https://www.linkedin.com/in/ron-painter-aa70055/.
[15] For more information on Jeffrey Smith, see https://econ.wisc.edu/staff/smith-jeffrey/.

Item 8 – Timelines, Challenges, and Changes*. Indicate where, when, and how data will be collected. Include, clear timelines and plans for releasing findings to relevant stakeholders and specify how departures from the plan, including changes related to timelines and methodological decisions, will be documented. Outline potential vulnerabilities to the timeline related to data collection or access and plans to mitigate risks. Provide the time schedule for the entire project, including beginning and ending dates of the collection of information, completion of the report, publication dates, and other actions.*

Table 6 displays a potential timeline for this impact evaluation of registered and unregistered apprenticeships. The dates to accomplish key data collection and analysis tasks were built to generate a final report with impact estimates by the end of the project timeline in Fall 2025.

Table 6

Illustrative Project Schedule for Scaling Apprenticeship and Closing the Skills Gap Impact Study

Task/Activity	Period
Study Design	
Technical working group meeting	May 2020
Final design report	November 2020
OMB/IRB approval	Fall 2020
Data Collection Agreements	
WIPS data use agreement	October 2020–December 2020
NDNH agreement	October 2020–January 2021
Data use agreements with grantees, states, and community colleges	October 2020–March 2021
Data Requests/Collection	
WIPS PY 2019, 2020, 2021, and 2022 data files	January 2021, July 2021, July 2022, July 2023, or quarterly extracts if feasible
Data from grantees	March 2021, September 2021, September 2022, September 2023, or more frequently if quarterly extracts from WIPS are available
Data from community colleges	June 2021, June 2022, June 2023
SSNs from states	July 2021, July 2022, July 2023
Preprogram earnings	August 2021, August 2022, August 2023
Follow-up earnings	August 2022, August 2023, January 2024, January 2025
Data Analysis	
Impact analysis with NDNH analysis file	January 2025–May 2025
Reporting/Dissemination	
Final Report Draft	June 2025
Final Report Final	September 2025
Briefs and briefings	July 2025–September 2025

Source: Authors.

Item 9 – Other relevant information. *Include any other information relevant to supporting the transparency and reproducibility of the study.*

The Urban Institute and Mathematica study team keep a detailed accounting of all data collected and analyzed for the study to ensure transparency. This record-keeping includes data dictionaries provided by community college systems, public workforce agencies, and grantees, as well as statistical code used to clean and analyze the data. Much of this information will be included in a methodological appendix or section of the final impact study report, and all can be made available on request.

The study team cannot provide data files for reproduction because the data-sharing agreements established with community colleges and grantees do not allow the publication or distribution of the data. However, the study team can facilitate any effort to obtain this data and reproduce the study through a detailed description of the study data and data cleaning procedure.

Item 10 – References. *Provide references and cite any relevant literature.*

Abadie, Alberto, and Guido W. Imbens. 2008. "On the Failure of the Bootstrap for Matching Estimators." *Econometrica* 76 (6):1537–557.

Abadie, Alberto, and Guido W. Imbens. 2016. "Matching on the Estimated Propensity Score." *Econometrica* 84 (2): 781–807. https://doi.org/10.3982/ECTA11293.

Abraham, Katharine G., John C. Haltiwanger, Kristin Sandusky, and James R. Spletzer. 2018. *Measuring the Gig Economy: Current Knowledge and Open Issues*. Working Paper no. w24950. Cambridge, MA: National Bureau of Economic Research.

Anderson, Theresa, Daniel Kuehn, Lauren Eyster, Burt S. Barnow, and Robert I. Lerman. 2017. *New Evidence on Integrated Career Pathways*. Washington, DC: Urban Institute.

Ashenfelter, Orley C. 1978. "Estimating the Effect of Training Programs on Earnings." *The Review of Economics and Statistics* 60 (1): 47–57. https://doi.org/10.2307/1924332.

Bang, Heejung, and James M. Robins. 2005. "Doubly Robust Estimation in Missing Data and Causal Inference Models." *Biometrics* 61 (4): 962–73. https://doi.org/10.1111/j.1541-0420.2005.00377.x.

Blakemore, Arthur E., Paul L. Burgess, Stuart A. Low, and Robert S. St. Louis. 1996. "Employer Tax Evasion in the Unemployment Insurance Program." *Journal of Labor Economics* 14 (2): 210–30.

Boren, Zach, Andrew Campbell, Bhavani Arabandi, John Marotta, Daniel Kuehn, Jacqueline Rayfield. 2022. *Union-Based Apprenticeships for Young People: Creating Good Jobs and Meeting Employers' Needs for Skills*. Washington DC: Urban Institute.

Card, David, and Daniel Sullivan. 1988. "Measuring the Effect of Subsidized Training Programs on Movements in and out of Employment." *Econometrica* 56 (3): 497–530. https://doi.org/10.2307/1911698.

Chesnaye, Nicholas C., Vianda S. Stel, Giovanni Tripepi, Friedo W. Dekker, Edouard L. Fu, Carmine Zoccali, and Kitty J. Jager. 2022. "An Introduction to Inverse Probability of Treatment Weighting in Observational Research." *Clinical Kidney Journal* 15 (1): 14–20. https://doi.org/10.1093/ckj/sfab158.

Copson, Elizabeth, Tresa Kappil, Karen Gardiner, Andrew Clarkwest, Hannah Engle, Alexander Trutko, John W. Trutko, Asaph Glosser, Riley Webster, Daniel Kuehn, Robert Lerman, and Jessica Shakesprere. 2021. "Implementing Registered Apprenticeship Programs: Experiences of 10 American Apprenticeship Initiative Grantees." Washington, DC: Department of Labor, Employment and Training Administration.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2009. "Dealing with Limited Overlap in Estimation of Average Treatment Effects." *Biometrika* 96 (1): 187–99. https://doi.org/10.1093/biomet/asn055.

Czajka, John L., Ankita Patnaik, and Marian Negoita. 2018. *Data on Earnings: A Review of Resources for Research*. Report for the U.S. Department of Labor. Oakland, CA: Mathematica Policy Research.

Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94 (448): 1053–62. https://doi.org/10.2307/2669919.

Dehejia, Rajeev H., and Sadek Wahba. 2002. "Propensity Score-Matching Methods for Nonexperimental Causal Studies." *The Review of Economics and Statistics* 84 (1): 151–61.

Dolton, Peter, and Jeffrey Andrew Smith. 2011. "*The Impact of the UK New Deal for Lone Parents on Benefit Receipt*." IZA Discussion Papers, No. 5491. Bonn, DEU: Institute for the Study of Labor.

Drake, Christiana. 1993. "Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect." *Biometrics* 49 (4): 1231–36. https://doi.org/10.2307/2532266.

Elejalde-Ruiz, Alexia,."Apprenticeships Come to Insurance and Financial Services Industries." March 7, 2016, *Chicago Tribune*, available at http://www.chicagotribune.com/business/ct-insurance-apprenticeships-aon-zurich-0308-biz-20160307-story.html.

Galdo, Jose C., Jeffrey Smith, and Dan A. Black. 2008. "IZA DP No. 3095: Bandwidth Selection and the Estimation of Treatment Effects with Unbalanced Data." *Annales d'Economie et de Statistique*: 189–216.

Gardiner, Karen, Daniel Kuehn, Elizabeth Copson, and Andrew Clarkwest. 2021. "Expanding Registered Apprenticeship in the United States: Description of American Apprenticeship Initiative Grantees and Their Programs." Washington, DC: Department of Labor, Employment and Training Administration; Rockville, MD: Abt Associates; Washington, DC: Urban Institute.

Glazerman, Steven, Dan M. Levy, and David Myers. 2003. "Nonexperimental Versus Experimental Estimates of Earnings Impacts." *The Annals of the American Academy of Political and Social Science* 589: 63–93. https://doi.org/10.1177/0002716203254879.

Hahn, Jinyong. 1998. "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica* 66 (2): 315–31.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *The Review of Economic Studies* 64 (4): 605–54. https://doi.org/10.2307/2971733.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66 (5): 1017–98. https://doi.org/10.2307/2999630.

Heckman, James J., and Yona Rubinstein. 2001. "The Importance of Noncognitive Skills: Lessons from the GED Testing Program." *American Economic Review* 91 (2): 145–49.

Heckman, James J., and Jeffrey A. Smith. 1999. "The Pre-Programme Earnings Dip and the Determinants of Participation in a Social Programme. Implications for Simple Programme Evaluation Strategies." *The Economic Journal* 109: 313–48.

Heinrich, Carolyn, J., Peter R. Mueser, Kenneth R. Troske, Kyung-Seong Jeon, and Daver C. Kahvecioglu. 2013. "Do Public Employment Training Programs Work?" *IZA Journal of Labor Economics* 2 (1): 1–23.

Helper, S., R. Noonan, J. Nicholson, and D. Langdon. 2016. *The Benefits and Costs of Apprenticeship: A Business Perspective*. Washington, DC: U.S. Department of Commerce, Economics and Statistics Administration, Office of the Chief Economist; Cleveland: Case Western Reserve University.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71 (4): 1161–89. https://doi.org/10.1111/1468-0262.00442.

Hollenbeck, Kevin, and Wei-Jang Huang. 2016. *Net Impact and Benefit-Cost Estimates of the Workforce Development System in Washington State*. Upjohn Institute Technical Report No. 16-033. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research. https://doi.org/10.17848/tr16-033.

Horvitz, Daniel G., and Donovan J. Thompson. 1952. "A Generalization of Sampling without Replacement from a Finite Universe." *Journal of the American Statistical Association* 47 (260): 663–85. https://doi.org/10.1080/01621459.1952.10483446.

Hotz, Joseph V., and John K. Scholz. 2001. *Measuring Employment and Income for Low-Income Populations with Administrative and Survey Data*. Madison, WI: University of Wisconsin–Madison Institute for Research on Poverty.

Houseman, Susan N. 2001. "Why Employers Use Flexible Staffing Arrangements: Evidence from an Establishment Survey." *Industrial and Labor Relations Review* 55 (1): 149–70. https://doi.org/10.2307/2696191.

Imbens, Guido W. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *The Review of Economics and Statistics* 86 (1): 4–29. https://doi.org/10.1162/003465304323023651.

Imbens, Guido W. 2015. "Matching Methods in Practice: Three Examples." *The Journal of Human Resources* 50 (2): 373–419. https://doi.org/10.3368/jhr.50.2.373.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge, GBR: Cambridge University Press. https://doi.org/10.1017/CBO9781139025751.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47 (1): 5–86. https://doi.org/10.1257/jel.47.1.5.

Katz, Lawrence F., and Alan B. Krueger. 2016. *The Rise and Nature of Alternative Work Arrangements in the United States, 1995–2015*. Working Paper no. w22667. Cambridge, MA: National Bureau of Economic Research.

Katz, Lawrence F., and Alan B. Krueger. 2019. *Understanding Trends in Alternative Work Arrangements in the United States*. Working Paper no. w25425. Cambridge, MA: National Bureau of Economic Research.

Kornfeld, Robert, and Howard S. Bloom. 1999. "Measuring Program Impacts on Earnings and Employment: Do Unemployment Insurance Wage Reports from Employers Agree with Surveys of Individuals?" *Journal of Labor Economics* 17 (1): 168–97. https://doi.org/10.1086/209917.

Kuehn, Daniel. 2019. "Registered Apprenticeship and Career Advancement for Low-Wage Service Workers." *Economic Development Quarterly* 33 (2). https://doi.org/10.1177/0891242419838605.

Kuehn, Daniel, and Diane Jones. 2018. *Sub-baccalaureate STEM Education and Apprenticeship*. Washington, DC: Urban Institute.

Kuehn, Daniel, Ian Hecker, and Alphonse Simon. 2019. *Registered Apprenticeship in Science and Engineering*. Washington, DC: Urban Institute.

Lechner, Michael, Ruth Miquel, and Conny Wunsch. 2011. "Long-Run Effects of Public Sector Sponsored Training in West Germany." *Journal of the European Economic Association* 9: 742–84. https://doi.org/10.1111/j.1542-4774.2011.01029.x.

Lerman, Robert, Lauren Eyster, and Daniel Kuehn. 2014. "Can We Upgrade Low-Skill, Low Wage Occupations? The Case of Apprenticeships in the Long-Term Care Occupations." *Journal of Women, Politics and Policy* 35 (2): 110–32. https://doi.org/10.1080/1554477X.2014.890835.

McCaffrey, Daniel F., Greg Ridgeway, and Andrew R. Morral. 2004. "Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies." *Psychological Methods* 9 (4): 403–25. https://doi.org/10.1037/1082-989x.9.4.403.

Mueser, Peter R., Kenneth R. Troske, and Alexey Gorislavsky. 2007. "Using State Administrative Data to Measure Program Performance." *The Review of Economics and Statistics* 89 (4): 761–83.

Reed, Debbie, Albert Yung-Hsu Liu, Rebecca Kleinman, Annalisa Mastri, Davin Reed, Samina Sattar, and Jessica Ziegler. 2012. *An Effectiveness Assessment and Cost-Benefit Analysis of Registered Apprenticeship in 10 States.* Report for the U.S. Department of Labor Employment and Training Administration. Oakland, CA: Mathematica Policy Research.

Rosenbaum, Paul R. 2002. *Observational Studies, 2nd ed*. New York: Springer.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55. https://doi.org/10.2307/2335942.

Rotz, Dana, Paul Burkander, Kenneth Fortson, Sheena McConnell, Peter Schochet, Mary Grider, Linda Molinari, and Elias Sanchez-Eppler. 2017. *Providing Public Workforce Services to Job Seekers: 15-Month Impact Findings on the WIA Adult and Dislocated Worker Programs (Technical Supplement)*. Washington, DC: Mathematica Policy Research.

Schochet, Peter Z. 2008. "Statistical Power for Random Assignment Evaluations of Education Programs." *Journal of Educational and Behavioral Statistics* 33 (1): 62–87. https://doi.org/10.3102/1076998607302714.

Smith, Jeffrey A., and Petra E. Todd. 2005. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125: 305–53.

Solomon-Fears, Carmen. 2011. *The National Directory of New Hires*. Washington, DC: Congressional Research Service.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the LASSO." *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–88. https://www.jstor.org/stable/2346178.

Walton, Doug, Karen Gardiner, and Burt Barnow. 2022. *Expanding Apprenticeship to New Sectors and Populations: The Experiences and Outcomes of Apprentices in the American Apprenticeship Initiative*. Rockville, MD: Abt Associates.