Department of Labor Evaluation Design Pre-Specification Plans

A. Background

The Department of Labor's Chief Evaluation Office 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, 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 altered from 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 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 (PRA) Information Collection Request requirements submissions.

B. Document Control

Table 1. Document Information

Title:	DOL Evaluation Design Pre-Specification Plan: Reentry Programs
Evaluator	Mathematica (prime)/Social Policy Research Associates
Security Level:	Public; no access restrictions.
Contact Info:	chiefevaluationoffice@dol.gov

Table 2. Document History

Version	Date	Summary of Change
1	September 2021	Initial version
2	June 2022	Revised based on feedback from Technical Working Group and updates to data collection
3	August 2022	Revised based on DOL external review

Evaluation Design Report for the Reentry Projects grants

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.*

For two decades, the U.S. Department of Labor (DOL) has invested in reentry services by committing substantial funding toward programs serving justice-involved young adults and adults, under a funding umbrella currently known as the Reentry Projects (RP) grants. Between 2016 and 2019, DOL awarded almost $300 million through these grants to improve participants' labor market and criminal justice outcomes. Grantees include both intermediary organizations that serve large numbers of participants across multiple subgrantees and states, and smaller, community-based organizations (CBOs). The services offered vary depending on the grant stream and target group, but all offer an array of services, including career preparedness, employment-focused services, and case management. In addition, all RP grantees were required to use at least one of the following employment strategies: registered apprenticeship, work-based learning, and career pathways.

DOL's Chief Evaluation Office has contracted with Mathematica and Social Policy Research Associates to build evidence about effective strategies to serve people with prior justice involvement and facilitate their successful reentry into the community. This comprehensive evaluation aims to determine the impacts of the program on labor market and criminal justice outcomes (*impact study*), understand how the grant programs were implemented across a broad range of intermediaries and CBOs (*implementation study*), and measure the outcomes of a broader set of RP participants than those included in the impact study (*outcomes study*). The remainder of Item 1 and Items 2 through 9 describe the quasi-experimental design (QED) of the impact study. Other design documents discuss the implementation and outcomes studies.

Impact study

The impact study is designed to answer key questions about the effects of reentry programs on participants' labor market and criminal justice outcomes. As discussed in Lacoe and Betesh (2019), most prior studies of adult reentry programs do not find consistent, positive effects, but this may be due to variation in program models, implementation quality, and study designs. Other evidence indicates that strategies that emphasize longer-term attachment to work through training and work experience have been effective with disadvantaged populations (Anderson et al. 2017; Copson et al. 2016; Hendra et al. 2016). Thus, it is important to determine if reentry programs that use these and other evidence-based strategies can improve outcomes for justice-involved youth and adults.

With this in mind, the impact study will compare RP participants' outcomes to outcomes of individuals with recent criminal justice involvement who receive employment services available

through the Wagner-Peyser Act. The study is designed to answer three primary research questions:

1. What is the impact of RP on the likelihood of being employed in the 9th and 10th quarters after enrollment compared with Wagner-Peyser employment services?
2. What is the impact of RP on participants' earnings in the 9th and 10th quarters after enrollment compared with Wagner-Peyser employment services?
3. What is the impact of RP on the likelihood of being convicted of a crime over the 10 quarters after enrollment compared with similar Wagner-Peyser employment services?

In addition, our exploratory research questions are as follows:

1. What is the impact of RP on participants' arrest rates and incarceration rates over the 10 quarters after enrollment compared with Wagner-Peyser employment services?
2. What is the impact of RP on participants employment and earnings outcomes in the 4th and 5th quarters after enrollment compared with Wagner-Peyser employment services?
3. What is the impact of RP on participants' conviction, arrest, and incarceration rates over the 5 quarters after enrollment compared with Wagner-Peyser employment services?
4. What is the impact of RP on the frequency and severity of criminal justice outcomes in the 9th and 10th – as well as the 4th and 5th – quarters after enrollment compared with Wagner-Peyser employment services?

Finally, the study will conduct exploratory analyses to determine whether impacts differed for the following key subgroups:

a. Adult versus young adult participants

b. Participants of different race or ethnicities

c. Participants of different gender

d. Participants with lower versus higher frequency of prior criminal justice involvement

e. Participants served by the different intermediaries (Opportunities Industrialization Center of America [OICA], The Dannon Project, PathStone Corporation, and AMIkids)

f. Participants served by grantees with different strategies or other characteristics uncovered by the implementation study (for example, participants served by grantees where the majority of people are referred directly by the court system versus those served by grantees with more community referrals)[1]

[1] We will refine our approach for defining these characteristics based on the information collected through the implementation study.

g. Participants who received different types of services (case management only, case management and work-based learning, and so forth)[2]

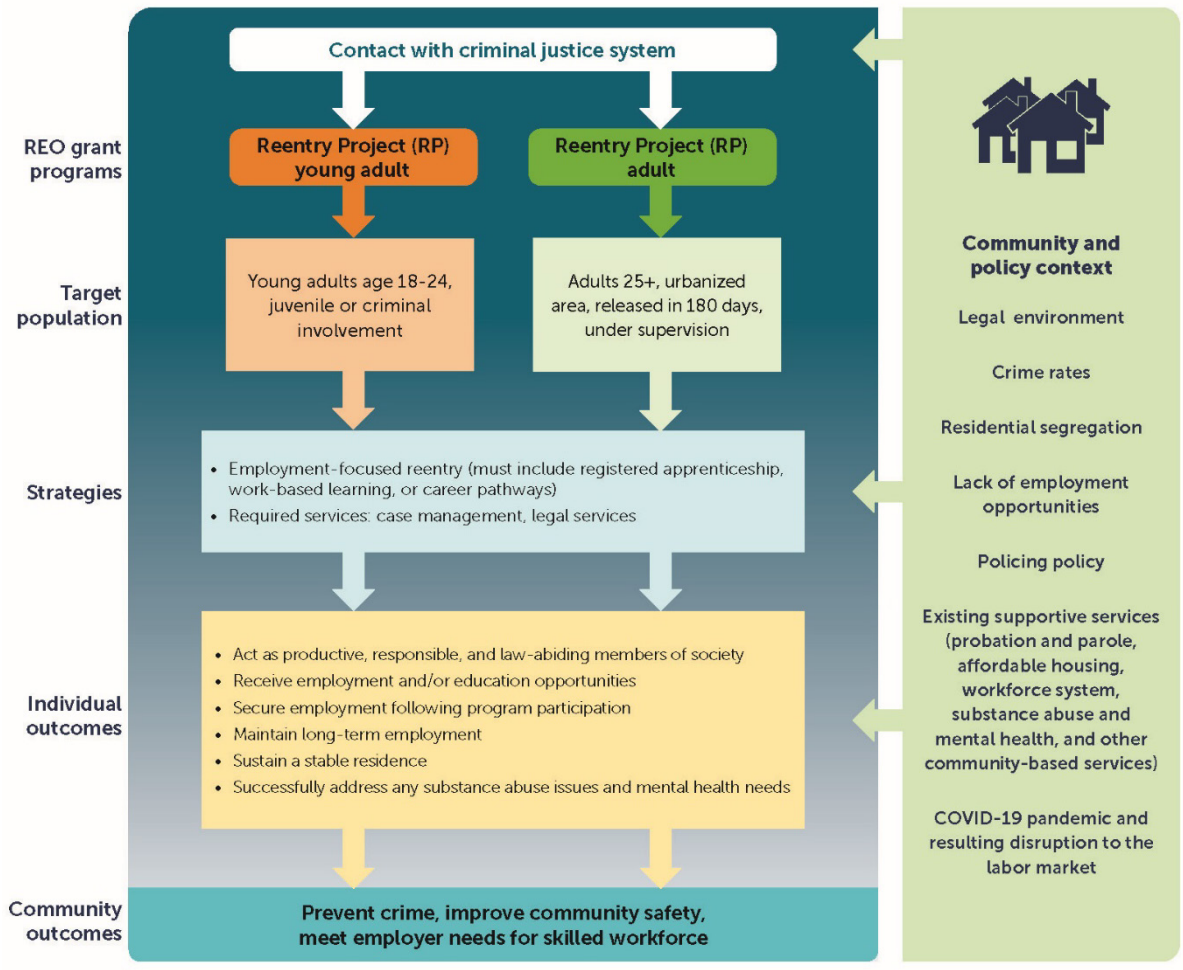
h. Participants that were mandated to participate by a court vs. those that were not, if this distinction is observable in the criminal justice data we receive from states.

[2] Because the types of services received are based on post-enrollment behavior, and are not pre-program characteristics, we will not give impacts for these subgroups a causal interpretation.

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.*

RP grantees combine structured employment experiences—through models such as apprenticeship, work-based learning, and career pathways—with case management to help participants transition to unsubsidized employment (see the logic model in Figure 2.1). The RP program's key objective is to improve participants' labor market and criminal justice outcomes, which motivates the selection of these outcomes for the primary research questions.

Figure 2.1. Logic model for RP grants



Note: REO = Reentry Employment Opportunities

RP participants will be compared to Wagner-Peyser participants who have similar prior criminal justice involvement and demographic characteristics, and who live in the same geographic areas.

This comparison group is credible because, in the absence of RP programs, many individuals with criminal justice backgrounds seeking to reenter the labor market may go to American Job Centers or otherwise enroll in Wagner-Peyser services.[3] In addition, just like RP participants, Wagner-Peyser participants with criminal justice backgrounds are clearly interested in seeking employment and employment-related services. In this sense, the study focuses on the impact of RP programs over and above normal workforce development services that are not specifically designed for reentering individuals. This is the effect of interest for DOL and the one that fills the most relevant gaps in the evidence base.

Challenges and solutions

As described in detail in the next section, the study depends on procuring data-sharing agreements with state workforce and criminal justice agencies. The study team has already been in contact with many of the relevant agencies, most of which have indicated they are willing and able to provide the necessary data, and a sufficient number of workforce agencies in these same states have indicated they can do the same. As of June 2022, the team is in the process of obtaining signed agreements with these agencies; the agreements will specify not only the participation of these agencies but also the timing with which data will be provided.[4]

[3] More than 3.5 million individuals received Wagner-Peyser services between July 2019 and June 2020.
[4] The study also requires data-sharing agreements with grantees; however, grant agreements require grantees to participate in an evaluation, so we do not view this as a challenge for the study.

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.*

The study will use four distinct types of data: (1) Workforce Integrated Performance System (WIPS) data, (2) National Directory of New Hires (NDNH) data, (3) state criminal justice data, and (4) personally identifiable information (PII) to link the other sources together. We will form our study sample using a treatment group of RP participants and a comparison group of Wagner-Peyser participants served in the same geographic areas with similar demographic characteristics and prior criminal justice involvement. To develop the sample, the design takes advantage of DOL's WIPS database, which has data on both RP and Wagner-Peyser participants, their geographic locations, and other important background characteristics. NDNH data will provide information on employment and earnings outcomes. Criminal justice data will provide information on criminal justice involvement both pre-enrollment (to use as matching variables) and post-enrollment (to use as outcomes).

Workforce Integrated Performance System. The WIPS is a national database that contains 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 RP grants. The WIPS contains data on individual-level demographic characteristics, including age, gender, race, ethnicity, disability status, education, employment status at program enrollment, and English learner status. The WIPS also includes data on employment and training services received through DOL workforce programs. In addition to using these data to form a matched comparison group, we will also use them to examine the impacts for key demographic subgroups and subgroups defined by service receipt, as described in Item 1. Finally, the WIPS contains data on the county of residence, a key matching variable that will allow the study to compare individuals facing the same local labor markets. Both grantees (for RP participants) and state workforce agencies (for Wagner-Peyser participants) collect these date uniformly and submit them to the WIPS. (DOL provides details on the full list of data elements in the WIPS and how each one is coded; DOL also has a validation procedure to confirm the accuracy of data elements). We will obtain program year 2018 (PY 2018) through the second quarter of program year 2021 (PY 2021 Q2) WIPS data for all RP and Wagner-Peyser participants. Because

program years start in the third quarter of each calendar year, we will obtain data on people who received RP and Wagner-Peyser services between July 2018 and December 2021.[5]

One data element in the WIPS data that is available for both RP and Wagner-Peyser participants is an indicator for whether an individual has prior criminal justice involvement. However, preliminary analysis on aggregate data reveals that this indicator, which is based on self-reports, is likely to represent a substantial undercount of justice-involved individuals. As such, in selecting Wagner-Peyser participants, we will not limit to those who have this indicator, but rather will use the state criminal justice data to identify prior criminal justice involvement of both RP and Wagner-Peyser participants, as described below.

National Directory of New Hires data. We will obtain quarterly employment and earnings data from NDNH, a database maintained by the Office of Child Support Enforcement (OCSE) at the U.S. Department of Health and Human Services. We will use these data to generate two constructs that serve as outcomes for primary research questions: (1) employment in the 9th and 10th quarters after program enrollment and (2) average quarterly earnings in the 9th and 10th quarters after program enrollment. We will also use the data to examine employment and earnings for the full post-enrollment period.

We can obtain NDNH data by submitting a list of Social Security numbers (SSNs) and names (called a "match-file") to OCSE. OCSE will then hold data for the individuals included in the request. We can then later request outcomes data from NDNH for these individuals (through what is called a "pass-through file"). The employment and earnings data from NDNH will be used only as outcome data. NDNH data are only available for eight prior quarters from when PII is submitted. By the time we will make our first submission to NDNH, the eight prior quarters will not cover the pre-enrollment period for many RP participants. Thus, we cannot use employment and earnings data as pre-program variables in the matching process. Although pre-program earnings are generally critical matching variables for impact evaluations of employment programs, in this case the sample members will have been incarcerated for all—or a large portion—of the relevant pre-program time period and thus do not have applicable earnings. We therefore expect that pre-program criminal justice involvement will be a more important set of variables to match on than pre-program reportable earnings.

Criminal justice data. The criminal justice data will serve two functions: as pre-program variables in the matching process and as outcomes. We will use at least two years of pre-program criminal justice data to provide sufficient information on background criminal justice involvement but will use a longer time span where possible.[6] We will also use the data to

[5] Because we will only be able to obtain outcome data for the 9th and 10th quarters after enrollment for individuals that received services by the end of PY 2020 (June 2021) – the follow-up period for our primary outcomes – we consider our main impact sample to be those that received services in PY 2018 through PY 2020. However, we will use the sample that enrolled in the first two quarters of PY 2021 (July through December of 2021) for other outcomes.

[6] We expect to have a full history for most states but may only have two years for some.

construct outcome measures for a primary research question—conviction rates in the 10 quarters following program enrollment—and for exploratory analyses, including arrest and incarcerations rates, and measures of the frequency and severity of criminal justice involvement. Once we have collected the criminal justice data, we will specify measures of frequency and severity that can be constructed consistently across the states.

Finally, we will use grantee information provided by DOL to identify participants associated with particular intermediaries to conduct the exploratory analysis on intermediaries. We will use qualitative data on grantee program models and strategies, collected for our implementation study, for exploratory analyses on these subgroups.

Personally identifiable information. To link the WIPS data to NDNH data and criminal justice data, we need three types of PII data: SSNs, names, and dates of birth. Table 3.1 indicates the three types of PII, what they will be used for, and the sources of those data for RP and Wagner-Peyser participants.

To collect quarterly employment and earnings data from NDNH, we will need names and SSNs.[7] To link study sample members to criminal justice data, we will need names and dates of birth. SSNs for RP participants are available in the WIPS data. We will request names and dates of birth for RP participants from grantees, linking them to the unique identifiers that grantees submit to the WIPS (WIPS IDs). For Wagner-Peyser participants, we will request SSNs, names, and dates of birth from state workforce agencies, because SSNs are not recorded in the WIPS for Wagner-Peyser participants.[8]

Table 3.1. Sources of key data elements for NDNH and criminal justice match

PII data element	Needed to obtain . . .	Data source	
		RP participants	Wagner-Peyser participants
SSN	NDNH data	WIPS database	State workforce agencies
Name	NDNH and criminal justice data	Grantees	State workforce agencies
Date of birth[a]	Criminal justice data	Grantees	State workforce agencies

Note: Names are not required for the NDNH match but are requested to improve the accuracy of the match. Names are required for linking to criminal justice data.

PII = Personally Identifiable Information, RP = Reentry Project, NDNH = National Directory of New Hires, WIPS = Workforce Integrated Performance System, SSN = Social Security Number.

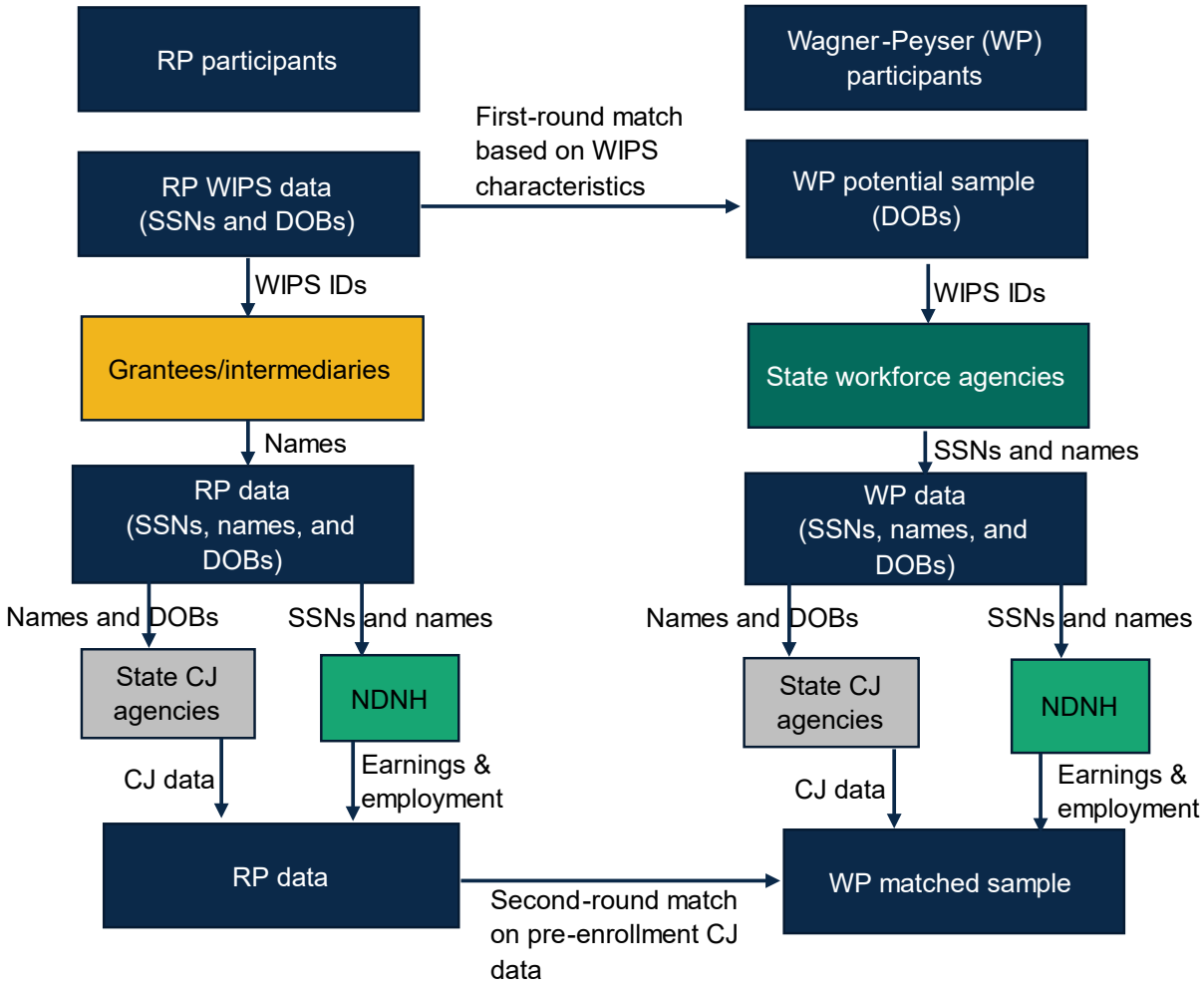
[7] Names are not formally required for matching with NDNH data; however, we will still collect them and send them to NDNH because doing so will provide a more accurate match to NDNH records, and we will need names to link to criminal justice data.

[8] Dates of birth are recorded in the WIPS for both RP participants and Wagner-Peyser participants, and we will use these data to calculate participant age, a key matching variable. However, it is not clear that we will be allowed to redisclose dates of birth to criminal justice agencies to link to criminal justice records. For the purpose of this report, we assume we will not have such permission.

[a] If we are permitted to redisclose dates of birth from the WIPS, then we will not need to collect this data element from intermediaries/grantees (for RP participants) or state workforce agencies (for Wagner-Peyser participants).

Figure 3.1 illustrates the process by which we will link various data sources for the treatment and comparison groups and the process for creating a comparison sample. We will collect data and create a comparison sample through a two-stage process. In the first stage, we will obtain PII for RP participants from the WIPS and RP grantees. We will then perform a first-round match, described in greater detail under Item 6, to identify a broad pool of potential comparison group members using WIPS data. Using PII collected for both RP participants and potential comparison group members, we will then obtain criminal justice data from state agencies in order to use pre-program criminal justice variables to construct a final comparison group.

Figure 3.1. Data collection and matching process



CJ = criminal justice; DOB = date of birth; RP = Reentry Project, NDNH = National Directory of New Hires, WIPS = Workforce Integrated Performance System, SSN = Social Security Number, WP = Wagner-Peyser.

There are several important caveats regarding the study's ability to capture key constructs with the available data. NDNH data contain outcomes only for people with reportable earnings in covered jobs. Although these data cover most wage and salary employment, they do not cover all types of jobs and industries. In particular, NDNH data only cover earnings submitted to Unemployment Insurance (UI) agencies. NDNH does not contain data on self-employed workers, most agricultural laborers, some domestic service workers, and part-time employees of nonprofit organizations (Czajka et al. 2018). In the past, these sectors have made up about 10 percent of U.S. employment (Kornfeld and Bloom 1999; Hotz and Scholz 2001). NDNH data also omit workers whose employers do not report their earnings to their UI agency, even in the formal sector (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, and exclude most work that is part of the gig economy (Abraham et al. 2018; Katz and Kruger 2016, 2019). Because we cannot distinguish between people who are truly unemployed and those who are employed but do not have reportable earnings, 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.

State criminal justice data come with a different caveat. We will only search for criminal justice records for RP and Wagner-Peyser participants in the states in which they lived when they enrolled in their program. We will drop sample members, in both the treatment and comparison groups, with no record of pre-program criminal justice involvement. This will mean excluding those who may have been in the criminal justice system in a different state from where they enrolled in their program (whether RP or Wagner-Peyser), those who were in the federal justice system, or those who are not found in state criminal justice systems for other reasons. Because we apply this restriction to both the treatment and comparison groups, we assume that it does not introduce bias.

In addition, we will not be able to distinguish between sample members who have no post-enrollment criminal justice involvement and those who were arrested, convicted, or incarcerated but do not appear in the post-enrollment criminal justice data. For participants who moved to a different state after enrollment and then became involved in the criminal justice system, or otherwise were not found in the criminal justice data in the post-enrollment period, their involvement would not be observed and the study will assume they do not have any post-program criminal justice involvement.[9] The study design assumes that this type of behavior is just as likely to occur in both the treatment and comparison groups and thus does not introduce bias into the impact estimates.

[9] The RExO study (Wiegand and Sussell 2015) of the justice-involved population was able to match about 87 percent of its study sample in state-provided administrative criminal justice records.

Item 4 – Response rates and attrition*. Describe methods to maximize response rates and to deal with issues of non-response. 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 the study consists of RP participants who received services from an RP program that was funded through grants awarded in 2017, 2018, or 2019. Because each RP program enrolls participants over an approximately three-year period, the study sample will therefore consist of RP participants who enrolled in an RP program from PY 2017 through PY 2020. The comparison group will consist of Wagner-Peyser participants with similar characteristics to the treatment group and who enrolled in the Wagner-Peyser program in the same geographic areas during the same time frame. We will drop sample members in both the treatment and comparison groups who have no record of pre-program criminal justice involvement in the criminal justice data.

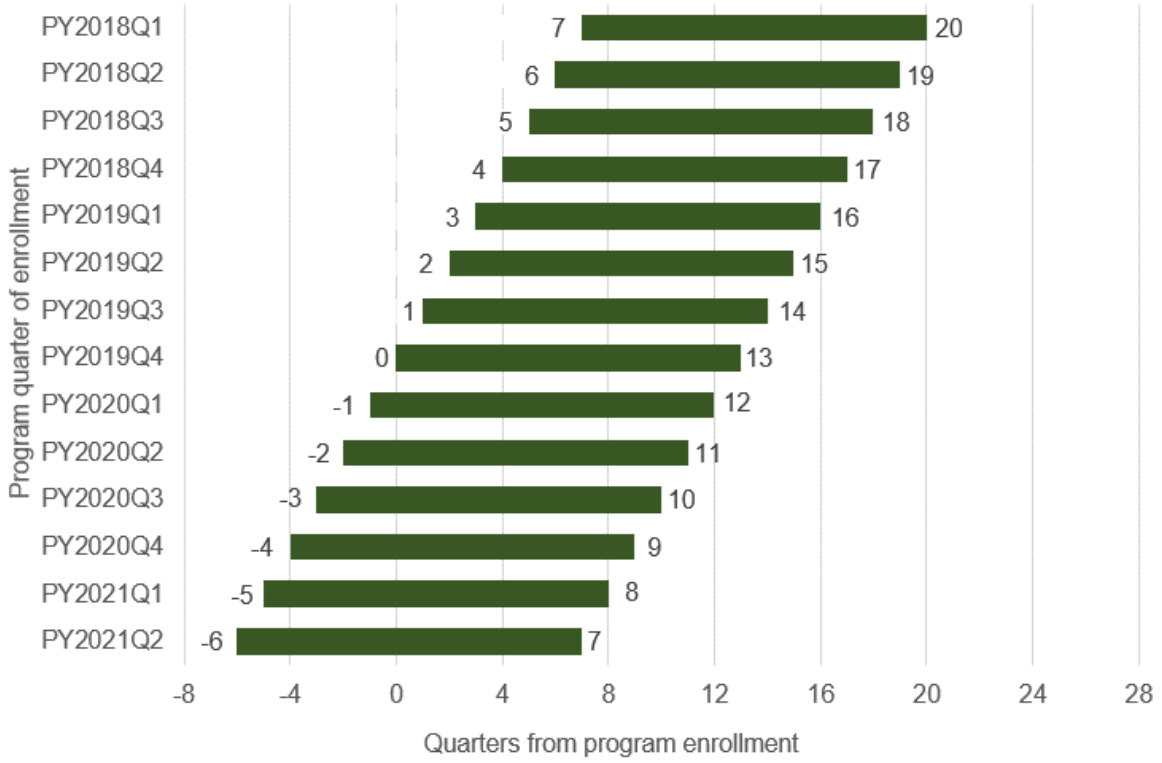
The design does not rely on individual study participants responding to any data collection instrument, other subjective reports, or sampling. However, the sample of individuals included in the study will be smaller than the full population of interest due to several factors:

1. **States included in study.** To use study resources most efficiently, we selected the 11 states where the four largest intermediaries operate, though the study also includes other grantees operating in those same states. Grantees in states outside of these 11 will not be included.
2. **State agencies' willingness to provide data.** For those states selected for the study, we cannot include RP participants or comparison group members in the study unless we have data from their state's workforce and criminal justice agencies (as indicated in Table 3.1). We discuss the implications of this in the power calculation section in Item 5.
3. **Timing of NDNH data availability.** The timing of our planned submissions to NDNH affects the number of RP participants for whom we can collect outcome data. At any given time, NDNH contains wage data going back only two years, limiting the employment and earnings outcome data we will have on RP participants. We plan to submit SSNs to OCSE in the second quarter (Q2) of 2022 to hold data on the sample of individuals for whom we intend to request outcome data. We then plan to submit a second request for NDNH data for this sample in Q4 2023. This schedule would give us employment and earnings outcomes covering Q2 2020 through Q3 2023. This, in turn, would allow us to measure employment and earnings in the 9th and 10th quarters following enrollment for anyone who started receiving RP services—and their counterparts who started receiving Wagner-Peyser services—between July 2018 and June 2021 (PY 2018 Q1 through PY 2020 Q4) (Figure 4.1). This timing affords the study the

maximum available sample given the spread of enrollment across grantees and grant years but excludes RP participants who enrolled in a program during PY 2017.

4. **Grantee data.** To use study resources most efficiently, we will only gather PII from grantees serving relatively few participants in the selected states if resources allow. Similarly, there could be grantees from the earliest grant year, 2017, whose grants have ended and no longer maintain the PII data we need. However, we think this latter concern is unlikely to be critical because (1) grantees received extensions due to the COVID-19 pandemic, and (2) even where grants have ended, the grantees themselves may still be in operation and may still have the necessary data.
5. **Missing or incorrect SSNs.** SSNs are necessary to obtain NDNH records for both treatment and comparison group members. It is likely that a small proportion of RP participants will have missing or incorrect SSNs in the WIPS data and a small proportion of Wagner-Peyser participants selected for the comparison group will have missing or incorrect SSNs in the state workforce data.

Figure 4.1. Expected timing of NDNH data availability, by quarter of program enrollment



NDNH = National Directory of New Hires.

The study will use demographic characteristics from the WIPS to (1) identify key subgroups to estimate impacts for specific sets of program participants, and (2) examine receipt of services for both RP and Wagner-Peyser participants. These data are verified through DOL's own data collection procedures. Grantees and state workforce agencies that submit data receive instructions and training on DOL's data collection procedures. Other key subgroups of interest

will be formed based on pre-program criminal justice involvement, which will be based on the criminal justice administrative data that state agencies manage.

Some pre-program data in these administrative data sources may be missing. We will impute missing variables so that we can use the information available in pre-program data without excluding individuals from the sample. We will conduct imputation separately for the RP and comparison groups. We will use a chained stochastic regression approach in which we impute variables using information from other variables (Rubin 1987; Raghunathan et al. 2001). The chained equation method runs a series of regression models that temporarily fill in missing values of variables when predicting other ones. This updating process continues until the change to the newly predicted values are below a pre-specified stopping criterion. We will then use predictive mean matching to impute missing observations. This method works, for example, by filling in a person's missing education level by (1) identifying a group of individuals with similar predicted education values to those of the person with the missing value and (2) using the actual education level of a randomly selected person in that group as the imputed value. This method is valid under the assumption that data are missing at random, conditional on the variables included in the imputation model.

Once we have identified the sample of RP participants and grantees included in the impact study, we will examine the extent to which they are representative of the full set of RP participants and grantees, using the data that will be available for all RP participants (WIPS and NDNH) and grantees. We will discuss this element of external validity in the final report.

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 RP impact study's approach relies on obtaining key data elements for RP and Wagner-Peyser participants from states and the NDNH. Because all data are administrative, the key selection mechanism determining which sample members to include in the study is based on states' willingness to share data.

Study sample. The population of interest for the impact study consists of RP participants who enrolled in PY 2018 through PY 2020, along with comparison group members who enrolled in Wagner-Peyser services in the same period. Table 5.1 presents the sizes of the RP participant sample in each state based on available enrollment data. These sample sizes incorporate our assumptions about the timing of the coverage of the NDNH data that we will receive, described in Item 4.

Table 5.1. RP participant sample sizes, by state and grantee type

State	Number of RP participants	Share of RP participants in state served by intermediary grantees (versus CBOs)	Intermediary in this state (% of intermediary sample in state)
Alabama[a]	631	99%	Dannon (88%)
Florida	1,536	44%	AMIkids (100%); OICA (15%)
Minnesota	262	59%	OICA (7%)
New Jersey	148	100%	PathStone (17%)
New York	1,790	11%	PathStone (22%)
North Carolina	781	71%	OICA (25%)
Ohio	800	23%	OICA (8%)
Oregon	462	67%	OICA (14%)
Pennsylvania	1,402	69%	OICA (32%); PathStone (30%)
Puerto Rico	278	100%	PathStone (32%)
South Carolina	87	100%	Dannon (12%)
Total	**8,177**	-	-

Note: Participants not served by intermediary grantees received RP program services from CBOs that received RP grants during PYs 2017 through 2019. Sample sizes assume a 90 percent match rate for RP participants in the criminal justice data.

RP = Reentry Project, CBO = community-based organization, OICA = Opportunities Industrialization Centers of America

[a] Conviction data in Alabama are only available for individuals older than 21. We anticipate that this will result in about half of the RP young adult participants in Alabama being included in the analysis. This assumption is incorporated in the sample sizes presented here and in the minimum detectable impacts below.

Power calculations. To ensure that the evaluation will have sufficient sample to estimate suitably precise impacts, we estimated minimum detectable impacts (MDIs) for each primary outcome. The MDI is the size of smallest impact of the program that we would be likely to detect as statistically significant based on the sample size and other parameters. Overall, we anticipate being able to detect impacts of 3.4 percentage points or more for employment, $311 or more for earnings, and 2.5 percentage points or more for conviction rates over the 9th and 10th quarters

after enrollment.[10] These MDIs are sufficiently small to be able to detect impacts that are substantively important and in the same range as the impacts that have been found in the literature.

Previous studies of employment-focused interventions for individuals with justice involvement have found a range of impacts on employment and recidivism outcomes. In the evaluation of the Center for Employment Opportunities program for former prisoners, Redcross et al. (2012) found an impact of 6.3 percentage points on average quarterly employment in unsubsidized jobs, no impact on unsubsidized earnings, and an impact of 5.6 percentage points on the probability of being convicted of a crime in the first three years after enrollment. The Re-integration of Ex-offenders (RExO) study (Wiegand and Sussell 2015) estimated, in the three years after enrollment, a decrease in employment of 2.6 percentage points, a decrease in quarterly earnings of $185 dollars, and an increase in conviction rates of 2.9 percentage points; however, none of these impacts was statistically significant. For the Enhanced Transitional Jobs Demonstration, the impact on employment was 4 percentage points, the impact on annual earnings was $701 dollars ($175 in quarterly earnings), and the impact on convictions in the 30th month after enrollment was 2.5 percentage points (Barden et al. 2018). The Second Chance Act Adult Demonstration Program (D'Amico and Kim 2018) found impacts of 4.6 percentage points on employment, $900 dollars in quarterly earnings, and 6.4 percentage points on convictions after 18 months. As discussed in Lacoe and Betesh (2019), however, the overall evidence on many employment-focused reentry programs is mixed, with several studies finding no significant impacts.

*MDIs under different scenarios***.** The MDIs for the impact study rely on assumptions about sample sizes and data availability that vary based on which states' workforce and criminal justice data we can obtain. We calculated MDIs to assess statistical power across two scenarios of data availability:

- **Scenario 1.** We obtain both workforce and criminal justice data from all 11 states.
- **Scenario 2.** We obtain both types of data from the five states that, based on ongoing data collection activities, are very likely to provide the data, have already provided data, or with whom we have executed a data sharing agreement.[11]

For each scenario, we restricted the power analysis to counties in which RP participants reside, based on the WIPS data, and assume we would be able to form a comparison group out of

[10] These MDIs are calculated under the assumption that we will only receive data from states for which we are in the process of obtaining both state workforce and criminal justice data. See Table 5.2 further detail on the assumptions underlying these calculations.
[11] These states are Alabama, Florida, New Jersey, Pennsylvania, and Oregon.

Wagner-Peyser participants that will have the same sample size as the group of RP participants.[12]

For estimates pooled across RP grantees, we will have adequate statistical power under all scenarios of data availability (Table 5.2). However, due to the spread of intermediary grantees across states, not every intermediary has sufficient sample to estimate suitably precise intermediary-specific impacts under each scenario using frequentist estimation methods. To estimate meaningful intermediary-specific impacts for each of the four intermediaries, we will use Bayesian methods, which require special considerations that we describe in further detail in Item 6.

Table 5.2. MDIs for each planned analysis under different scenarios

Outcome	Scenario 1	Scenario 2
Employment (MDIs)	2.3 pp.	3.4 pp.
Quarterly earnings (MDIs)	$216	$311
Convictions (MDIs)	1.9 pp.	2.5 pp.
Number of RP participants	6,537	3,761

Note: The last row provides the number of RP participants included in the sample under each scenario. The MDIs are calculated based on several assumptions: (1) the comparison sample will be equal in size to the number of RP participants, (2) the employment rate for the comparison group is 65 percent (based on the RExO study of a similar population; Weigand and Sussell 2015), (3) the standard deviation of quarterly earnings in the comparison group is $4,349 (based on individuals in the Workforce Investment Act Gold Standard Evaluation's core services group in the 9th and 10th quarters after random assignment), (4) the conviction rate for the comparison group is 25 percent (Wiegand and Sussell 2015), (5) 10 percent of the variation in the employment and earnings outcomes is explained by pre-program covariates, (6) 20 percent of the variation in conviction rates is explained by pre-program covariates, and (7) 90 percent of RP participants will show up in criminal justice records in their state of residence.

pp. = percentage points, RP = Reentry Project, MDI = minimum detectable impact

Power considerations associated with linking administrative justice records. The MDIs presented above assume that 90 percent of RP participants will appear in criminal justice records in their state of residence. This match rate assumption is based on match rates from prior studies of the impacts of reentry programs for justice-involved individuals (Wiegand and Sussell 2015). However, because participant records will be linked to state-provided justice records using names and dates of birth (rather than a common identifier such as a social security number or other numeric ID), this introduces the additional possibility of record linking error beyond this match rate. Recent evidence suggests that linking errors may have considerable impacts on statistical power by introducing attenuation bias in the impact estimates. These linking errors

[12] As shown in Table 5.1, the state with the largest number of RP participants, Pennsylvania, had fewer than 1,500 participants over a three-year period. Pennsylvania had more than 270,000 Wagner-Peyser participants in these same three years. As such, it is very likely that the number of Wagner-Peyser participants with prior criminal justice involvement would exceed 1,500. Because we will employ a weighting approach that allows us to include many more comparison group members than treatment group members, the assumption that the two groups will be equal in size is conservative.

may be false positives (matching an individual to a justice record that does not belong to them) or false negatives (failing to match an individual to a justice record that does belong to them). The larger the total linking error (the sum of both false positives and false negatives), the lower the statistical power to detect effects will be.

Using simulations presented in Tahamont et al (2021) as a benchmark, we anticipate that a total linking error rate of 10 percent would increase our MDI for conviction rate outcomes to 2.9 percentage points from 2.5 percentage points (holding all other assumptions constant). If our total linking error rate were 15 percent, our MDI for the conviction rate outcome would be 3.8 percentage points rather than 2.5.[13]

For the exploratory outcomes that represent subgroup analyses described under Item 1, we present the relevant anticipated sample sizes below. These sample sizes pertain to RP participants only. We are not able to anticipate sample sizes for subgroups defined by characteristics related to criminal justice – because we have not collected criminal justice data yet – or related to grant program strategies – because we have not completed qualitative analysis.

Table 5.3. Sample sizes for select subgroups

Subgroup defined by . . .	Anticipated number of RP participants
Gender	
Male	2727
Female	992
Race/ethnicity	
Hispanic	515
White, non-Hispanic	617
Black, non-Hispanic	2388
Other race, non-Hispanic	130
RP grant type	
Adult	1680
Young adult	2041
Intermediary grantee	
OICA	1332
The Dannon Project	436

[13] These are based on derivations in Tahamont et al. (2021) that present tradeoffs between statistical power (the probability of type II error) and sample size. This study's sample size is roughly twice as large as that used in their derivations, meaning that we expect a lesser effect of linking error on our MDIs than what is presented in their findings. Additionally, we expect our false negative rate on criminal justice outcomes to be lower than in the typical scenario presented in their paper, as our sample is restricted to both RP and Wagner Peyser participants that match to criminal justice records in the pre-program period.

Subgroup defined by . . .	Anticipated number of RP participants
PathStone	368
AMIKids	308

Note: Sample sizes are restricted to RP participants in the 5 states specified under Scenario 2 above.

RP = Reentry Project, OICA = Opportunities Industrialization Centers of America

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).*

To estimate the impact of RP program services, we will compare outcomes for individuals who received program services to a comparison group with similar characteristics. Identifying this comparison group and estimating impacts will involve a three-stage procedure.

1. First-stage match

The pool of potential comparison group members consists of participants enrolled in Wagner-Peyser employment services in the selected states. Given that it is infeasible to collect criminal justice data from state agencies for the full group of Wagner-Peyser participants, we will perform a first-stage match using key variables available in the WIPS data to narrow down the potential comparison group sample.[14] In this first stage, we aim to identify about 50 Wagner-Peyser participants for each RP participant. We plan to select this many comparison group members because we assume that most of them will not have criminal justice backgrounds and thus will not be used in the impact analysis. We will use the following procedure to select the Wagner-Peyser participants from the WIPS for each RP participant:

1. For each RP participant, we will identify individuals who resided in the same county and who enrolled in the Wagner-Peyser program in the same year and quarter that the RP participant enrolled in the RP program.
2. Within this group of Wagner-Peyser participants we will use coarsened exact matching (CEM) to match RP participants to the individuals with whom they share the closest overlap in demographic characteristics. We will match individuals without replacement. For this approach, we will match on key demographic characteristics available in the WIPS data for both RP and Wagner-Peyser participants, including age at enrollment,

[14] Given the magnitude of individuals who enroll in Wagner-Peyser programs each year (more than 3 million), we believe it would be infeasible to collect criminal justice data for a sample of that size.

gender, race/ethnicity, education level, employment status at program enrollment, receipt of dislocated worker services, English learner status, veteran status, and disability status.

3. In addition to matching RP participants to Wagner-Peyser participants residing in the same county, we will additionally use CEM as described in Step 2 to select matches based on demographic characteristics, using Wagner Peyser participants with the same state and quarter of enrollment as each RP participant. We will add any Wagner Peyser participant identified as a suitable match on state, program quarter of entry, and demographic characteristics who was not identified as a potential match in Step 2.[15]
4. We will additionally select all individuals identified in Step 1 who are flagged as having prior criminal justice involvement in the WIPS data (the "ex-offender" variable) and were not selected in Step 2.

Although we do not know the number of Wagner-Peyser participants with criminal justice involvement, the sample of RP participants is considerably smaller, so we expect to identify at least as many Wagner-Peyser participants with a criminal justice background in each county with RP participants.

At the end of this first round of matching, we will have a set of WIPS identifiers for the RP participants and their pool of potential comparison group members. We will then use these identifiers to obtain PII through state workforce agencies (for Wagner-Peyser participants) or RP grantees (for RP participants) that we will submit to state criminal justice agencies to receive data on pre-program criminal justice involvement to use as part of the second stage of matching.

2. Propensity score estimation

After receiving pre-program data from state criminal justice agencies, we will first exclude any participants, in both the treatment and comparison groups, who have no record of pre-program criminal justice involvement. Next, we will estimate propensity scores that capture the probabilities of receiving RP program services, conditional on pre-program characteristics.

We will pool the sample identified during the first-round match within state. We will then estimate the probability that an individual received RP program services (as opposed to Wagner-Peyser services) by state. The covariates will include the data on observed demographic characteristics used in the first-round match along with information on pre-program criminal justice involvement and county-level characteristics (including county-level unemployment and county-level poverty from the American Community Survey). Overall, we plan to use the following variables in the propensity score model:

- Age at enrollment
- Gender

[15] We do not expect that every RP participant will necessarily receive 50 unique matches. We intend to iterate through this process until the overall number of Wagner-Peyser participants we have identified is suitable in size.

- Race/ethnicity
- Education level
- Receipt of dislocated worker services
- English learner status
- Veteran status
- Disability status
- County-level unemployment rate
- County-level poverty rate
- Prior criminal justice involvement, potentially including[16]
  - Number of arrests
  - Number of convictions
  - Time since release
  - Release reason
  - Arrest charge class/category
  - Conviction charge class/category
  - Sentencing category/length
  - Time incarcerated

We will estimate propensity scores using the following methods:

- Generalized boosted regression model (GBM), a machine-learning approach that uses an algorithm to search over the set of provided covariates and select the interactions and data partitions that most predict participation (McCaffrey et al. 2004). In this approach, the algorithm generates and includes interactions and higher-order terms of the covariates.
- Bayesian additive regression trees (BART), a machine-learning method that uses a Bayesian statistical model to iterate over a series of "regression trees" to identify covariates and interactions that best fit the data (Chipman et al. 2010). One benefit of the BART approach is that its flexibility allows it to account for differences in the relationship between covariates and the propensity score across subgroups (for example, different relationships between the covariates and the propensity score across grantees).
- Least absolute shrinkage and selection operator (LASSO) regression (Tibshirani 1996), an alternative machine learning technique that selects covariates among a set of specified variables. The LASSO regression limits the number of covariates by penalizing each additional covariate added to the model.

[16] We will define the specific variables to be used for matching from the justice records once we have determined the availability of specific data elements across state justice agencies.

- Logistic regression to serve as a baseline comparison for the machine learning approaches.

We would ideally like to restrict comparisons to RP and Wagner-Peyser participants who reside within the same county, because labor market conditions, justice system characteristics, and available services vary geographically. However, the smaller sample sizes within a county raise the concern that restricting comparisons to local areas would reduce the quality of the predicted propensity scores. To balance these concerns, we plan to estimate the propensity score on the pooled sample across counties within each state, and estimate a separate propensity score within each county. We will then use these estimated propensity scores to assess the covariate balance of comparison groups constructed both across and within counties. If the within-county propensity scores achieve a similar covariate balance as the cross-county propensity scores, then we will use the within-county propensity scores, as they have the advantage of also achieving balance in the distribution of counties. If the within-county propensity scores perform markedly worse in terms of covariate balance, we will pool across counties.

Covariate balance. The goal of the propensity score estimation process is to construct treatment and comparison groups that are similar based on pre-program characteristics. We will therefore select the propensity score estimation approach based on how well the treatment and comparison samples are balanced on covariates. We will use the prognostic score as a summary measure to assess covariate balance. This score is calculated by estimating a regression model to predict an outcome (in our case, each of the three primary outcomes), separately for both the treatment and comparison groups. We will then compare the mean predicted values for the two study groups (Zhang et al. 2019). A smaller difference in mean outcomes for a given propensity score model versus a different model indicates that the model leads to better overall covariate balance, incorporating information on differences in means of the covariates between the study groups and how those covariates are associated with the outcome. We will select our primary method for estimating the propensity score as the one that produces the lowest standardized mean difference in prognostic scores. This method has been shown in simulations to outperform selection based on comparisons of means across predictors of the propensity score (Stuart et al. 2013).

Although GBM is our preferred approach, we will ultimately use the propensity score model, as well as the within-county versus across-county approach, that achieves the greatest covariance balance. Note that we will select the propensity score estimation approach and assess covariate balance before the employment and earnings outcome measures are available, ensuring that our choice will not be influenced by the potential impact estimates that would result.

3. Impact estimation

Outcomes. We will estimate impacts for a range of employment and criminal justice outcomes, both for the cross-grantee analysis and the intermediary-specific analysis.

Estimation approach. The parameter of interest for the impact study will be the average treatment effect on the treated (ATT). The ATT represents impacts for the population of individuals who received RP program services. After selecting the method to estimate propensity scores that results in the best balance on pre-program characteristics across the RP and comparison groups, we will use the propensity scores to estimate impacts on a range of employment and criminal justice outcomes. Our primary approach to estimation will be inverse probability weighting (IPW; Horvitz and Thompson 1952). IPW is an appealing approach as it allows for the inclusion of a larger sample: instead of dropping potential comparison group members who may be close, if not perfect, matches, the IPW approach gives those observations less weight in the regression. Having a larger sample can improve precision and efficiency (Hirano et al. 2003).

Before estimating impacts, we will trim the sample to remove any individual with an estimated propensity score outside of the range of [0.1,0.9], the rule of thumb suggested by Crump et al. (2009). This strategy improves the overlap of the research samples, avoids assigning very large weights, and prevents comparison group members from being included in the analysis if they are not similar to the treatment group based on pre-program characteristics.

To estimate the impact of RP program participation on employment, earnings, and criminal justice outcomes, we will generate weights based on our estimated propensity score. These weights will be equal to 1 for all RP participants and equal to a function of the estimated propensity score for each member of the comparison group. We will adjust comparison group weights to sum to a value proportional to the number of RP participants within each county.

To estimate impacts, we will then use a weighted least squares regression, controlling for key covariates, using the following regression model:

$$Y_{ig} = \alpha + \beta T_i + \gamma X_{ig} + \delta_g + \varepsilon_{ig}$$

Y_{ig} is the outcome Y for individual i in county g. T_i is an indicator for whether the individual received RP services. X_{ig} is a set of individual covariates, and δ_g is a county fixed effect (that is, an indicator for living in a specific county). We will control for the same set of covariates in the impact estimation as we include in the propensity score estimation, described above. This approach produces what is called "doubly robust" impact estimates (Funk et al. 2011). For binary outcomes, we will use weighted least squares estimation of the linear probability model.

To account for estimation error in the propensity score, we will use generalized method of moments to estimate the propensity score model jointly with the impact model (Abadie and Imbens 2016). With this approach, the standard errors of the impact estimates will also account for estimation error in the propensity scores (but not estimation error associated with selecting the propensity score model itself).

Intermediary-specific estimates. To estimate the impacts of specific intermediary grantees on outcomes, we will first estimate intermediary-specific impacts using the approach described above. However, we will likely not have a sufficient sample to estimate suitably precise effects for two of the four intermediaries. Therefore, we plan to estimate intermediary-specific impacts using a Bayesian approach. We will fit a prior distribution of treatment effects using the pooled impact estimate. Intuitively, this can be thought of as the best estimate of what a specific intermediary's impact will be, before looking at its data. Next, we will use a Bayesian model to incorporate the data from that intermediary to estimate the posterior distribution of the intermediary-specific impact estimate. Intuitively, this can be thought of as our best estimate of what a specific intermediary's impact will be, after looking at its data.

As with a frequentist regression, impact estimates for smaller intermediaries will be less precise than those for larger intermediaries. However, the Bayesian approach allows for greater precision of the intermediary-specific impacts than a frequentist approach would yield, because it accounts for information learned from other grantees.

Sensitivity analyses

For each confirmatory outcome, we will conduct the following sensitivity analyses to assess the robustness of the estimated impacts to alternative approaches.

a. **Matching-based estimation approaches.** Instead of weighting Wagner-Peyser participants by the inverse estimated probability of participating in RP services, we will alternatively explore estimating impacts by matching RP participants to Wagner-Peyser participants based on the propensity score. We will do this using caliper matching with bias correction and nearest-neighbor matching with replacement. Caliper matching is an approach that selects all Wagner-Peyser participants with a propensity score within a given distance ("caliper") from each RP participant to form a matched comparison group sample. Nearest-neighbor matching with replacement constructs a matched comparison group by selecting Wagner-Peyser participants who are "nearest" to each RP participant based on the values of the estimated propensity scores. These approaches may reduce bias relative to IPW in cases when there is relatively low overlap between the treatment and comparison group (Busso et al. 2014).
b. **Logistic regression models for binary outcomes.** We will estimate impacts on binary outcomes using logistic regression to assess the robustness of the estimates resulting from the linear probability model.
c. **Bayesian causal forests.** Our final sensitivity analysis will use a recently developed estimation method called Bayesian causal forests (BCF), which flexibly models both the propensity score and the outcome of interest to estimate the effect of RP participation. BCF does not impose strong restrictions on the functional form of the model and therefore is more flexible in a range of scenarios. BCF has also outperformed other predictive modeling techniques in simulations (Hahn et al. 2020).

d. **Stronger service contrast.** We will also conduct a sensitivity analysis in which we limit the comparison group to those that received only light-touch employment services – such as access to a computer lab to conduct job searches – rather than more intensive services such as one-on-one counseling. In theory, if some members of the comparison group receive similarly intensive employment services as the RP participants, then one might expect this to result in a smaller impact. This analysis will reveal whether a stronger service contrast between the treatment and comparison group would generate a larger impact. However, since the original comparison group should be more representative of the services that RP participants would have received if not for the RP program, our findings will emphasize the impact estimate from that comparison.

e. **Using zip code, rather than county.** If there are zip codes with large enough numbers of RP and Wagner-Peyser participants, we will estimate the impact for this subset using only within-zip code variation to estimate impacts. This could reveal if comparisons across zip codes in our main specification could be affecting the impact estimates.

f. **Using pre-program earnings for some participants.** While we will not have a large enough sample of participants that will have pre-program earnings available, there will be a smaller group of participants and comparison group members for whom pre-program earnings are available. This group consists of participants that enrolled in PY2020Q3 through PY 2021Q2. We will use this group to test whether the absence of these characteristics for the larger sample induces bias in the impact estimates.

If the impacts from the sensitivity analysis differ substantively from those of the main impact specification, we will explore and describe the reason for those differences. However, we will ultimately emphasize the findings from the main specification because we think it reflects the best balance between precision and accuracy.

Subgroups

We will additionally estimate the cross-grantee impacts of RP program services for subgroups defined by baseline demographic characteristics. We will focus on subgroups based on age, race/ethnicity, gender, level of prior justice involvement, whether individuals were mandated to participate by a court (if this distinction is observable in the criminal justice data we receive from states), characteristics of the grantees, and types of services received. For each subgroup, we will estimate subgroup impacts by adjusting the regression model to include terms interacting subgroup indicators with the treatment status indicator.

Interpretation of findings

We will assess whether impact estimates are statistically significant at the 0.05 level. Our presentation of the findings will emphasize the confirmatory research questions and hypothesis tests and describe the exploratory, or secondary, analyses as suggestive. Our interpretation will include an in-depth discussion of the commonalities and differences in service provision and

program models across RP programs and how these differ from the services available to Wagner-Peyser participants. For transparency, we will report p-values, rather than just indicating whether or not an estimate was statistically significant at the 0.05 level.

The Bayesian framework we will use for the intermediary-specific impacts does not support frequentist hypothesis testing. Instead, we will interpret these impacts by estimating the probabilities that the true impact of RP participation for a given intermediary exceeds a selected threshold. This will allow us to make statements such as "There is a 70 percent probability that RP increased earnings for AMIkids participants in the 9th and 10th quarters after enrollment." We will also present credible intervals for each impact estimate, providing an interval within which each impact would fall with a given probability. For example, a 95 percent credible interval would define the range of impacts between the 2.5 and 97.5 percentiles of the posterior distribution (Gelman et al. 2013).

We will aggregate summary statistics in all public reporting and omit small cell sizes that might risk participant disclosure. Otherwise, we do not anticipate removing excluding any data due to disclosure risk.

Item 7 – 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 report, publication dates, and other actions.*

The timeline for the impact study is in Table 7.1. The details of the various data collection steps are discussed under Items 3 and 4.

Table 7.1. Timeline for RP impact study

Milestone	Timing
Submit initial design report to DOL	September 2021
Technical working group meeting on initial design report	December 2021
Submit final design report to DOL	August 2022
Negotiate data-use agreements with state agencies and grantees	July 2021–June 2022
Obtain WIPS data for both RP and Wagner-Peyser participants	February – March 2022
Obtain PII from state agencies and grantees	February – May 2022
Submit PII to NDNH	June 2022
Retrieve employment and earnings data from NDNH	October–December 2023
Deliver draft report to DOL	June 2024
Technical working group meeting on findings	July 2024
Deliver final report to DOL and release to stakeholders	September 2024

RP = Reentry Project, PII = Personally Identifiable Information, NDNH = National Directory of New Hires, WIPS = Workforce Integrated Performance System, DOL = Department of Labor.

The potential risks to the project involve the timing of collecting the WIPS data, the PII from state workforce agencies and grantees, and the records from state criminal justice agencies. If we cannot submit PII to NDNH by June 2022, then we could not get employment and earnings data from Q2 2020, and our earliest quarter of labor market data would instead be Q3 2020 or later. This could also mean that for some proportion of the sample, the NDNH coverage period would not include their 9th and 10th quarters after enrollment, and thus they would not contribute to the impact analysis for the primary outcomes.

We have several strategies to mitigate this risk. The study team has already started obtaining data-use agreements from the relevant state agencies. Second, we could submit data to NDNH in multiple batches. That is, if we have the necessary PII for the treatment and comparison groups

for most states by June 2022, but not all states, we could submit data for the states we have and only lose the Q2 2020 data for the remaining states, for which we would submit data later. Most importantly, since we are unable to obtain criminal justice records before June 2022, we will submit the PII from the WIPS and state workforce agencies to NDNH. This will result in submitting information to NDNH on some individuals who will not ultimately be used in the analysis—because the submission will include participants with no criminal justice involvement. However, this will not affect the analysis because those individuals could later be omitted after we obtain their criminal justice records (or lack thereof).

Any changes to the timeline that materially affect the study, and any changes to the methods described in this report, will be documented in the form of a revised report submitted to DOL.

Item 8 – 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.*

To solicit independent and expert feedback on the evaluation design and the study's findings, we will convene a technical working group (TWG) of experts for two meetings. The first meeting will cover the impact and implementation design, along with the design for the outcome study, and will take place toward the end of 2021. We will hold the second meeting toward the end of the project and will use it to gather expert input on the interpretation of the study findings. We will send a draft of the final report to TWG members in advance of the second meeting. If necessary, we will also ask for written input from TWG members if important design or analysis issues arise during the study. The TWG members for this study are:

- Shawn Bushway, State University of New York (Albany) and RAND
- Harry Holzer, Georgetown University
- Debbie Mukamal, Stanford University
- Omari Swinton, Howard University
- Christy Visher, University of Delaware

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

The study team originally sought to design a randomized controlled trial (RCT) of the RP grants because experimental designs tend to generate unbiased impact estimates with fewer assumptions than a QED. We held discussions with grantees and DOL to determine the feasibility of an RCT. In collaboration with DOL, we ultimately determined an RCT would not be feasible. We then designed the QED described in this report. We also considered a QED that focused primarily on estimating separate impacts for each of the four intermediaries. However, we determined the sample size would not be sufficient to generate precise impact estimates for most of these intermediaries. As a result, in conjunction with DOL, we designed the QED to focus on the aggregate impact estimate that pools across the four intermediaries and other grantees operating in the same states.

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

Abadie, Alberto, and Guido W. Imbens. "Matching on the Estimated Propensity Score." *Econometrica*, vol. 84, no. 2, March 2016, pp. 781-807.

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

Anderson, T., D. Kuehn, L. Eyster, B. Barnow, and R.I. Lerman. "New Evidence on Integrated Career Pathways: Final Impact Report for Accelerating Opportunity." Washington, DC: Urban Institute, 2017.Barden, B., R. Juras, C. Redcross, M. Farrell, and D. Bloom. "New Perspectives on Creating Jobs: Final Impacts of the Next Generation of Subsidized Employment Programs." New York: MDRC, 2018.

Blakemore, Arthur E., Paul L. Burgess, Stuart A. Low, and Robert S. St. Louis. "Employer Tax Evasion in the Unemployment Insurance Program." *Journal of Labor Economics*, vol. 14, no. 2, 1996, pp. 210–230.

Busso, Matias, John DiNardo, and Justin McCrary. "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators." *The Review of Economics and Statistics*, vol. 96, no. 5, 2014, pp. 885–897.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. "BART: Bayesian Additive Regression Trees." *Annals of Applied Statistics*, vol. 4, no. 1, 2010, pp. 266–298.

Copson, E., K. Martinson, V. Benson, M. DiDomenico, J. Williams, K. Needels, and A. Mastri. "The Green Jobs and Health Care Impact Evaluation: Findings from the Implementation Study of Four Training Programs for Unemployed and Disadvantaged Workers." Bethesda, MD: Abt Associates, 2016.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. "Dealing with Limited Overlap in Estimation of Average Treatment Effects." *Biometrika,* vol. 96, no. 1, January 2009, pp. 187–199.

Czajka, John L., Ankita Patnaik, and Marian Negoita. "Data on Earnings: A Review of Resources for Research." Washington, DC: Mathematica, 2018.

D'Amico, R., and H. Kim. "An Evaluation of Seven Second Chance Act Adult Demonstration Programs: Impact Findings at 30 Months." Oakland, CA: Social Policy Research Associates, 2018.

Funk, Michele Jonsson, Daniel Westreich, Chris Wiesen, Til Stürmer, M. Alan Brookhart, and Marie Davidian. "Doubly Robust Estimation of Causal Effects." *American Journal of Epidemiology,* vol. 173, no. 7, 2011, pp. 761–767.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis, Third Edition.* Boca Raton, FL: CRC Press, 2013.

Hahn, P. Richard, Jared S. Murray, and Carlos M. Carvalho. "Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion)." *Bayesian Analysis,* vol. 15, no. 3, 2020, pp. 965–1056.

Hendra, R., D.H. Greenberg, G. Hamilton, A. Oppenheim, A. Pennington, K. Schaberg, and B.L. Tessler. "Encouraging Evidence on a Sector-Focused Advancement Strategy: Two-Year Impacts from the WorkAdvance Demonstration." New York: MDRC, 2016.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica*, vol. 71, no. 4, 2003, pp. 1161–1189.

Horvitz, Daniel G., and Donovan J. Thompson. "A Generalization of Sampling Without Replacement from a Finite Universe." *Journal of the American Statistical Association*, vol. 47, no. 260, 1952, pp. 663–685.

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

Houseman, Susan N. "Why Employers Use Flexible Staffing Arrangements: Evidence from an Establishment Survey." *Industrial and Labor Relations Review*, vol. 55, no. 1, October 2001, pp. 149–170.

Katz, Lawrence F., and Alan B. Krueger. "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, 2016.

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

Kornfeld, Robert, and Howard S. Bloom. "Measuring Program Impacts on Earnings and Employment: Do Unemployment Insurance Wage Reports from Employers Agree with Surveys of Individuals?" *Journal of Labor Economics*, vol. 17, no. 1, January 1999, pp. 168-197.

Lacoe, J., and H. Betesh. "Supporting Reentry Employment and Success: A Summary of the Evidence for Adults and Young Adults." Oakland, CA: Mathematica, 2019.

McCaffrey, Daniel F., Greg Ridgeway, and Andrew R. Morral. "Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies." *Psychological Methods*, vol. 9, no. 4, 2004, pp. 403–425.

Raghunathan, Trivellore E., James M. Lepkowski, John van Hoewyk, and Peter Solenberger. "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models*." Survey Methodology*, vol. 27, no. 1, June 2001, pp. 85–96.

Redcross, C., M. Millenky, T. Rudd, and V. Levshin. "More Than a Job: Final Results from the Evaluation of the Center for Employment Opportunities (CEO) Transitional Jobs Program." New York, NY: MDRC, 2012.

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

Stuart, Elizabeth A., Brian K. Lee, and Finbarr P. Leacy. "Prognostic Score–Based Balance Measures can be a Useful Diagnostic for Propensity Score Methods in Comparative Effectiveness Research." *Journal of Clinical Epidemiology*, vol. 66, no. 8, 2013, pp. S84-S90.

Tahamont, Sarah, Zubin Jelveh, Aaron Chalfin, Shi Yan, and Benjamin Hansen. "Dude, Where's My Treatment Effect? Errors in Administrative Data Linking and the Destruction of Statistical Power in Randomized Experiments." *Journal of Quantitative Criminology,* vol. 37, 2021, pp. 715–749.

Tibshirani, Robert. "Regression Shrinkage and Selection via the LASSO." *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, 1996, pp. 267–288.

Wiegand, A., and J. Sussell. "Evaluation of the Re-Integration of Ex-Offenders (RExO) Program: Final Impact Report." Oakland, CA: Social Policy Research Associates, 2015.

Zhang, Zhongheng, Hwa Jung Kim, Guillaume Lonjon, and Yibing Zhu. "Balance diagnostics after propensity score matching." *Annals of translational medicine*, vol. 7, no. 1, 2019.