Options for Building Evidence on RESEA Programs

Evaluation to Advance RESEA Program Evidence

April 2022

Call Order Number 1605DC-18-F-00382

Submitted to:
Chief Evaluation Office
Office of the Assistant Secretary for Policy
U.S. Department of Labor
Frances Perkins Building
200 Constitution Ave., NW
Washington, DC 20210

Submitted by:
Abt Associates
6130 Executive Boulevard
Rockville, MD 20852

Authors:
Jacob Alex Klerman, Abt Associates
Demetra Nightingale, Urban Institute
Andrew Clarkwest, Abt Associates
Zachary Epstein, Abt Associates

Chief Evaluation Office
U.S. DEPARTMENT OF LABOR

This report was prepared for the U.S. Department of Labor (DOL), Chief Evaluation Office by Abt Associates under Contract # 1605DC-18-A-0037. The views expressed are those of the authors and should not be attributed to DOL, nor does mention of trade names, commercial products, or organizations imply endorsement of same by the U.S. Government.

About This Report

This report is the capstone product of the Abt Associates team's work to develop options for building evidence on the effectiveness of RESEA programs, conducted as part of the *Evaluation to Advance Reemployment Services and Eligibility Assessments (RESEA) Program Evidence*. Other contract products include the *RESEA Evaluation Toolkit: Key Elements for State RESEA Programs* (Mills De La Rosa et al., 2021), the *Report on the State of Evidence for RESEA* (Epstein et al., 2022), and the report on *RESEA Program Strategies: State and Local Implementation* (Trutko et al., 2022). All project reports can be found on the "Completed Studies" page of the U.S. Department of Labor/Chief Evaluation Office (DOL/CEO) website: https://www.dol.gov/agencies/oasp/evaluation/completedstudies.

Acknowledgements

The authors acknowledge the contribution of a wide range of people who make this document possible. Five members of a technical expert panel—Thomas D. Cook (Northwestern University), John Deke (Mathematica Policy Research), Marta Lachowska (Upjohn Institute), Jeffrey Smith (University of Wisconsin-Madison), and Till von Wachter (University of California-Los Angeles)—provided early ideas on evaluation considerations and designs. Karin Martinson, Daniel Litwok, John Trutko, Yvette Chocolaad, and Julie Squire provided insightful review and feedback. Alec Wall provided data analysis and interpretation. Caroline Roddey provided additional research assistance. Bry Pollack edited the manuscript. David Dupree helped with graphics and formatting.

At DOL's Chief Evaluation Office, Megan Lizik and Chief Evaluation Officer Christina Yancey provided strong guidance and support. At DOL's Office of Unemployment Insurance, Lawrence Burns and former director Gay Gilbert provided very helpful input. The study team also acknowledges help from the following other staff at DOL: Michelle Beebe, Gloria Salas-Kos, and Ellen Wright.

Contents

Exhibits

Boxes

Executive Summary

The 2018 amendments to the Social Security Act (hereafter "the Statute") permanently authorized the Reemployment Services and Eligibility Assessment (RESEA) program, required that states' programs be supported by evidence, and allowed states to use up to 10 percent of their RESEA grant for evaluations. Developed as part of the *Evaluation to Advance RESEA Program Evidence*, this evidence-building options report aims to serve as a resource for decision makers to understand and weigh options for developing evidence of various types. Specifically, for a wide range of options, the document considers: *What* should be evaluated? *How* should that be evaluated? *Who* (states, consortia of states, or DOL) should initiate the particular evaluation? *When* should that evaluation occur? And *how* can the state workforce agency's capacity to conduct evaluations be strengthened?

ES.1 Research Questions and Specific Options

This document considers options related to four research questions (RQs):

RQ1. ***Whole Programs***: What is the impact of being selected for RESEA—relative to not being selected for RESEA?

RQ2. ***Subgroups***: How does the impact of RESEA vary with the characteristics of the claimant at initial claim?

RQ3. ***Components***: How does the impact of an RESEA program vary with service or component details?

RQ4. ***What Works Best for Whom***: How does the impact of changing a component vary with the characteristics of the claimant at initial claim?

Each of these RQs considers impacts; that is, *what difference did the intervention make?* This document includes options for impact evaluation designs, as well as options for non-impact analyses that are useful in helping to design interventions or to produce complementary evidence that enhances the value of evidence produced through impact analyses. Exhibit ES-1 lists the options considered.

Exhibit ES-1. Summary of Evidence-Building Options

Option	Who would lead?	When (to start and to finish)?
Estimating Impact of Whole Programs		
Option 2.1-1/Individual-Level Random Assignment	**Most states**: States that can randomize 30,000 to 50,000 RESEA-eligible UI claimants over 1-3 years	**To Start:** Approximately 6-12 months to prepare initial random assignment **To Finish:** Results available in 3-4 years, at a minimum
Option 2.1-2/Regression Discontinuity	**Few states**: States that use (or have used) profiling score in deciding which claimants to select for RESEA and have roughly 200,000 RESEA-eligible claimants "near" the profiling score cutoff	**To Start:** Immediately, if profiling model does not require revision **To Finish:** Results available in about 1 year if retrospective, and 4 or more years if prospective
Support to Estimate Impact of Whole Programs		
Option 2.2-1/Evaluation Technical Assistance	**DOL**	**To Start:** Can start immediately **To Finish:** Indefinite

Option	Who would lead?	When (to start and to finish)?
Option 2.2-2/Providing Common Analytic Tools	**DOL**	**To Start:** Can start immediately **To Finish:** Within 1 year
Option 2.2-3/Providing Common Data	**DOL**	**To Start:** Short term, to begin defining data elements to collect **To Finish:** Ongoing as states submit data in the future
Option 2.2-4/Cost-Benefit Analysis	**Most states**: States able to complete Option 2.1-1 or 2.1-2	**To Start:** After completion of state impact evaluation in 3-5 years **To Finish:** Within 1 year of start, assuming cost data were collected during the impact evaluation; otherwise, 2 years
Option 2.2-5/Develop a Template for Cost-Benefit Analysis	**DOL**	**To Start:** Can start immediately **To Finish:** Within 1 year
Option 2.2-6/Synthesis of the Impact of a Generic RESEA Program	**DOL**	**To Start:** After completion of at least 2-3 state impact evaluations in 3-5 years **To Finish:** Within 1 year after starting
Option 2.2-7/Sophisticated Synthetic Estimates of the Impact of a Specific State's RESEA Program	**DOL**	**To Start:** After completion of several impact evaluations in 4-7 years **To Finish:** 1-2 years after starting
Option 2.2-8/Coordinating Evaluations in Support of Synthesis	**DOL**	**To Start:** Can start immediately **To Finish:** Within 1 year
Subgroup Analysis		
Option 2.3-1/Coordinating Subgroup Analysis–Reporting Guidance	**DOL**	**To Start:** Can start immediately **To Finish:** Within 1 year
Option 2.3-2/Synthesis of Subgroup Impact Estimates	**DOL**	**To Start:** After completion of at least 2-3 state impact evaluations in 3-5 years **To Finish:** Within 1 year after starting
Option 2.3-3/Disparate Impact of Selection Strategies	**DOL**	**To Start:** Can start immediately **To Finish:** Within 6 months of start
Estimating Impact of Components		
Option 3.2-1/Individual-Level Random Assignment to Estimate Impact of a Component on RESEA Statutory Outcomes	**Few states**: States that can randomize perhaps 150,000 RESEA-eligible UI claimants over 1-3 years	**To Start:** Approximately 6-12 months to prepare initial random assignment **To Finish:** Results available in 3-4 years, at a minimum
Option 3.2-2/Individual-Level Random Assignment to Estimate Impact of a Component on Intermediate Outcomes Only	**Most states**: Feasibility depends on component of interest	**To Start:** Approximately 6-12 months to prepare initial random assignment **To Finish:** Results available in 2 years, at a minimum
Option 3.2-3/Cross-State Meta-Regression of Whole Program Evaluations	**DOL**: Analyze experimental estimates of impact to estimate how impacts vary with program characteristics	**To Start:** After completion of several impact evaluations in 4-7 years **To Finish:** 1-2 years after starting

Option	Who would lead?	When (to start and to finish)?
Option 3.2-4/Cross-State Interrupted Time Series of Observational Data	**DOL**: Observe how whole-program outcomes vary with changes in program design	**To Start:** After collection of 5-10 years of data **To Finish:** Within 1 year after starting
Support to Estimate Impact of Components		
Option 3.3-1/Consortia Building	**DOL and/or states**: States cooperate on evaluating the same program component	**To Start:** Can start immediately **To Finish:** Long term; results of experimental impact evaluations available in 3-4 years, at a minimum
Option 3.3-2/Deliberate Program Development for Pilots and Demonstrations	**DOL and/or states**: Consortia states specify details of program design	**To Start:** Can start development immediately **To Finish:** 1-3 years to complete
Option 3.3-3/Implementation Studies to Accompany Impact Evaluations	**All states**: Carefully document the components of a whole program	**To Start:** Can start development immediately **To Finish:** 1-2 years to complete
Option 3.3-4/Annual Survey of State Program Characteristics	**DOL**	**To Start:** Can start adapting existing RESEA Implementation Study Survey immediately **To Finish:** Indefinite; survey conducted annually for 5-10 years

ES.2 Short-Term Priorities (within the Next Three to Five Years)

For both DOL and the states, the highest short-term priority is to generate sufficient evidence to satisfy the statutory requirement that RESEA programs be demonstrated effective. That determination that existing evidence is sufficient to deem a generic[1] RESEA program effective will be made by DOL's Clearinghouse for Labor Evaluation and Research (CLEAR) program (https://clear.dol.gov/), once more evidence is available. Under the evidence criteria established in DOL guidance (Unemployment Insurance Training Letter 01-20), the most direct route to that determination requires at least two experimental evaluations that both: (1) meet CLEAR standards for study evidence quality[2] and (2) find statistically significant evidence of positive impacts (**Option 2.1-1/Individual-Level Random Assignment**). This

[1] Here "generic" means that CLEAR has found RESEA overall as effective, based on evidence from whichever states have completed evaluations. An alternative would be to deem a state's own program effective—and perhaps to require that.

[2] For *studies* that examine the causal effect (or "impact") of a labor-related intervention, CLEAR assigns to each study a rating that reflects the credibility of the evidence that the study presents—that is, how confident we can be that the findings presented by the study truly reflect the causal effect of that intervention, rather than some other factor that might also influence outcomes. For details on how CLEAR rates the credibility of evidence of causal studies, see CLEAR's Causal Evidence Guidelines at https://clear.dol.gov/about. Study ratings do not indicate whether the intervention itself is effective (i.e., these study ratings do not consider whether there is evidence that the program improves outcomes).

For reemployment-related *interventions*, CLEAR also provides a rating of evidence of effectiveness (i.e., whether the intervention improves outcomes). Those ratings take into account all sufficiently credible studies of the intervention (specifically, studies that received a High or Moderate CLEAR rating) to rate the extent of causal evidence that the intervention is effective. Under Social Security Act Section 306, interventions are required to have a High or Moderate rating to be eligible for funding. For details of how interventions are rated, see CLEAR's RESEA page at https://clear.dol.gov/reemployment-services-and-eligibility-assessments-resea.

determination might be reached in time for Fiscal Year (FY) 2025 or 2026 state RESEA Plans, but FY 2027 or even FY 2028 seem more likely.

DOL is providing two types of evaluation technical assistance (TA): (1) general evaluation TA to all states; and (2) additional customized evaluation TA to a select, small group of states. Together this evaluation TA will likely speed up when DOL CLEAR can deem a generic RESEA program effective. Continuing evaluation TA at least until results become available from the 2020 Cohort would substantially improve the likelihood of having high-quality evaluations that meet rigorous CLEAR standards for credibility of causal evidence (**Option 2.2-1/Evaluation Technical Assistance**). Option 2.2-1 therefore seems worthy of serious consideration, as does synthesis of those studies to make that determination **(Option 2.2-6/Synthesis of the Impact of a Generic RESEA Program**).

The activities of the previous paragraph will likely be sufficient to deem a *generic* RESEA program effective. Here "generic" means that CLEAR finds RESEA overall as effective, based on evidence from whichever states complete an evaluation. If an individual state (or DOL) wants to test a *specific state's* RESEA program to determine whether it is effective, additional support would be useful. Such additional support might start with whole-program evaluation TA (**Option 2.2-1/Evaluation Technical Assistance**) to more states and for longer than is currently funded. That assistance might also include steps to lower the costs to conduct evaluations and increase state agencies' capacity to conduct such evaluations (**Option 2.2-2/Providing Common Analytic Tools** for analysis; **Option 2.2-3/Providing Common Data** for data).

When a dozen or more state **RQ1/Whole Program** estimates become available,[3] **Option2.2-7/ Sophisticated Synthetic Estimates of the Impact of a Specific State's RESEA Program** would be feasible. By accumulating strong evidence from several evaluations over time and across states, the sophisticated synthesis of whole-program evaluations would help to address the lack of precision in single-year estimates of impact from evaluations conducted in smaller states that can only yield small sample sizes. This evidence synthesis option would also help to generate insights into how the impact of the RESEA program varies with the business cycle, and in particular, with changes in the state unemployment rate.

Finally, understanding how program impact varies with claimant characteristics (**RQ2/Subgroups**) is crucial for understanding equity-related questions and may produce useful insights as to who to select for RESEA. However, few single-state studies are likely to be large enough to estimate subgroup impacts with a useful level of precision. Pooling across multiple states' whole-program experimental impact evaluations (**Option 2.1-1/Individual-Level Random Assignment**) would address that limitation and provide insights about effective programs. To extract those insights, state whole-program experimental impact evaluations need to conduct their analyses in a common way and report a common set of outcomes. To make such analyses possible, DOL may want to consider publishing recommendations for how and what subgroups and outcomes the evaluations report (**Option 2.3-1/Coordinating Subgroup Analysis–Reporting Guidance**) and then fund a subgroup synthesis (**Option 2.3-2/Synthesis of Subgroup Impact Estimates**). Neither of these options would be high cost or methodologically complicated.

ES.3 Longer-Term Priorities (beyond Three to Five Years)

Though satisfying the RESEA statutory requirement is likely to be a pressing short-term priority, a longer-term priority is to generate evidence as to which programmatic strategies would improve outcomes for Unemployment Insurance (UI) claimants receiving RESEA or similar reemployment services. One

[3] Likely at least half a dozen.

approach to improving program outcomes—selecting for RESEA those claimants who will benefit more—is noted in the previous section.

The other approach to improving program outcomes is to conduct component experimental impact evaluations (**Option 3.2-1/Individual-Level Random Assignment to Estimate Impact of a Component on RESEA Statutory Outcomes**) that evaluate the change in outcomes if a state adds, deletes, or changes components. If primary funding for component evaluations will come from states, then the most important criterion for selecting which components to evaluate is that several states want to and agree to participate in and fund a coordinated evaluation of that component. With that crucial caveat, four component evaluations seem promising, based on a combination of expressed interest by DOL, extent of recent RESEA program implementation changes (Trutko et al., 2022), and REA impact evidence (Klerman et al., 2019):

- ***Remote Services***. States responded to the COVID-19 pandemic by shifting to a range of remote services. Specifically, most states switched from in-person one-on-one and group sessions to some combination of phone calls, live video calls, and recorded videos (Trutko et al., 2022). Limited available evidence suggests that states will not totally return to their pre-COVID strategies. These alternative modes of service delivery have advantages and disadvantages, and it is not clear which alternative produces better outcomes. This option would estimate the impact of various remote service models relative to a conventional—that is, all in-person—service delivery model.
- ***Intensive Reemployment Services***. Across federal staff, state leadership, and state line workers, RESEA is often viewed primarily as a casework program; that is, RESEA workers providing one-on-one reemployment services to individual UI claimants. If *some* casework is good, perhaps *more intensive* casework—in particular, more hours per claimant—would be better. This option would estimate the impact of a more intensive service delivery model relative to a conventional, less intensive service delivery model.
- ***Responses to Non-Attendance at the RESEA Meeting***. A recent evaluation of an RESEA-like program suggests a major role for non-attendance policy in driving the impact of reemployment programs—including RESEA—on UI weeks and on how that impact varies across states (Klerman, et al., 2019). This option would estimate the impact of "suspend until attend" relative to a less punitive and more cooperative approach.
- ***Eligibility Assessment***. Eligibility assessment appears in the RESEA program's name, but implementation research consistently finds staff reluctant to vigorously enforce ongoing eligibility requirement (Klerman, et al., 2019). This option would estimate the impact of vigorous assessment of ongoing eligibility—in particular, sufficiently intensive work search—relative to a less punitive and more cooperative approach.

For each of these options, the discussion above has only described a *general direction* for a component. An evaluation would estimate the impact of a *specific version* of the component. Proceeding would require carefully specifying the component to be evaluated. Ideally, the component to be evaluated would incorporate the field's sense of the most promising form of the component. Then the preliminary design for that component would be piloted and refined—before starting a large-scale, and more expensive impact evaluation (Epstein and Klerman, 2012).

Option 3.3-2/Deliberate Program Development for Pilots and Demonstrations describes a process to develop a candidate component for evaluation. For example, if the component to be studied were intensive services, the process might address questions such as: *When will intensive services be provided? To whom? What will be the content and intensity (e.g., minutes per claimant) of those services?* Then that candidate component would be refined through piloting and formative evaluation. How long such a process—from identifying a general direction through to a piloted concept ready for impact evaluation—

would take would vary—from less than a year to several years—depending on the number of piloting and program refinement cycles. Deliberate program development could start immediately, but it would need to finish before the impact evaluation starts.

Given these timelines, if DOL wants to be ready to start large component evaluations when the current round of whole-program evaluations ends, then starting them in calendar year 2022 appears worthy of serious consideration. Once the component evaluation starts, it can be accompanied by cost studies and a cost-benefit analysis (**Option 2.2-4**) and an implementation study (**Option 3.3-3**).

ES.4 Discussion

Based on evidence from evaluations of the Reemployment and Eligibility Assessment program, the Bipartisan Budget Act of 2018 sharply increased funding for RESEA, required states to show that their programs are evidence-based, and allocated substantial funds to the states for their own evaluations.

This document describes a wide range of **short-term evaluation options**; that is, studies whose primary goal is to allow states to show that either their own RESEA program or a generic RESEA program is evidence-based. Some of those options are for the evaluations themselves; others would support those evaluations.

There is some urgency to this short-term goal of allowing states to demonstrate that their programs are evidence-based. Some of the statutory requirements related to showing that a state's program is evidence-based are already in effect, and others will be in place in FY2023. However, evaluation timelines imply that new RESEA-specific impact results will not be available until well after that. It is possible that CLEAR will be able to deem a generic RESEA program effective in time for state FY 2025 or 2026 RESEA Plans, but in time for state FY 2027 or 2028 RESEA Plans seems more likely. Evidence for more than a few states will likely not become available until later in this window.

This document also describes a wide range of **longer-term evaluation options**; that is, studies whose primary goal is to identify programmatic changes that would improve the impact of RESEA programs on claimants' outcomes. As with short-term options, some of these longer-term options concern the evaluations themselves; others would support those evaluations (e.g., provide data or software, synthesize the results).

For several reasons, timelines for such longer-term evaluations to identify ways to improve the impact of RESEA programs are longer. Each of the steps in the evaluation takes longer: accumulating enough sample, building multi-state consortia to achieve even larger samples, and developing and piloting program models that are worth subjecting to impact evaluation. It appears that some states' short-term evaluations will include some effort to identify specific component designs that improve the impact of RESEA programs. More sustained effort in that direction will probably not occur until after the first round of evaluations publish final reports in the late 2020s. The second round of evaluations is thus not likely to finish until the early 2030s.

Even this timeline assumes that the programs to be subjected to impact evaluation have already been developed and piloted. Done right, such program development/piloting takes a year or more. Starting in 2022 would ensure that the specific component designs exist when states consider—and form consortia for—the second round of impact evaluations.

Throughout, the report considers **what entity is best positioned to fund and direct each of the options**: a single state, a consortium of states, or DOL. Because states administer the RESEA programs, and the Statute gives them control of most of the evaluation funds, they naturally have a leading role. However, this leading role for individual states brings with it two challenges. First, most of the pertinent research questions need samples larger than available from most states alone. Second, research always has a "free

rider" problem. When one state runs the evaluation, all states benefit. As a result, each state has an incentive to "free ride" on another state's evaluation and not conduct its own evaluation. That free rider problem is intrinsic to research; it can be exacerbated when states control the funds and choose whether and what to evaluate. Both of these challenges imply that coordination of state efforts would be useful. States on their own could collaborate to produce some of that collaboration. But DOL has capabilities that put it in a uniquely strong position to support and catalyze some types of inter-state coordination that may be unlikely to occur otherwise. Specifically: DOL has national responsibilities and perspective; DOL has funds to stimulate coordination; DOL writes the RESEA guidance; and DOL reviews and approves state RESEA Plans.

1. Introduction

To better serve participants, federal and state governments have increasingly emphasized the use of rigorous evidence to inform policymaking and program design decisions. For example, the Foundations for Evidence-Based Policymaking Act of 2018 (hereafter "Evidence Act") requires federal agencies to develop evidence-building agendas and evaluation plans. In this spirit, the Unemployment Insurance system (UI), operated by the U.S. Department of Labor (DOL), is developing and conducting evaluations to build evidence as to what activities and policies improve UI claimants' labor market outcomes, thereby improving program outcomes.

The **Reemployment Services and Eligibility Assessment (RESEA)** program is the latest in a series of DOL-funded state programs that verify eligibility for UI and provide reemployment services to a subset of UI claimants. The 2018 amendments to the Social Security Act (hereafter "the Statute") required that states' RESEA programs be supported by evidence. The Statute also allowed states to use their RESEA grant monies to fund evaluations to help generate that kind of supporting evidence. In response, DOL sponsored the *Evaluation to Advance RESEA Program Evidence* study being conducted by Abt Associates, in partnership with the Urban Institute. Under that same contract, Abt Associates and the Urban Institute produced this report on options for building evidence on RESEA programs.

Granting this role for evaluation in the legislation and the subsequent study, research questions that remain include these: *What* should be evaluated? *How* should that be evaluated? *Who*—states or DOL—should initiate a specific evaluation? *When* should that evaluation occur? And *How* can the state workforce agency's capacity to conduct evaluations be strengthened?

Some of the options presented here would generate estimates of the impact of the RESEA program. Other options would support generating estimates of impact (e.g., providing states with tools for conducting their own evaluations). Still others (e.g., "meta-analysis") would synthesize estimates from across new impact studies to produce additional insights. What entity seems best positioned to conduct these options varies: sometimes states, sometimes consortia of states, and sometimes DOL.

This report describes each option and its purpose. The discussion of each then considers the entity best positioned to initiate the study, when the study would occur (including actions that must occur before the corresponding impact evaluation study could start and actions that cannot occur until the corresponding impact evaluation is completed), feasibility considerations (including how much the option would cost), and recommendations of which options are highest priority.

The remainder of this opening chapter proceeds as follows. Section 1.1 describes the RESEA program and the Statute's evidence requirements. Section 1.2 summarizes the evidence base developed from prior evaluations of predecessor programs and the status of evaluations of RESEA. Finally, Section 1.3 describes the structure of the balance of the report; that is, how the report organizes its description of options to build on that existing evidence base to meet the various learning needs of state workforce agencies and DOL.

1.1. The RESEA Program

This section presents a brief overview of program goals, program content, and evidence-based features of the RESEA program.

Program Goals. As specified in the Social Security Act (Section 306(b)); (emphasis added), RESEA's aims are:

*To improve **employment** outcomes of individuals who receive Unemployment Compensation (UC)[4] and to reduce the average **duration of receipt** of such compensation through employment;*

To strengthen program integrity and reduce improper payments of Unemployment Compensation by states by detecting and preventing such payments to individuals who are not eligible for them;

To promote alignment with the broader vision of the Workforce Innovation and Opportunity Act (WIOA) of increased program integration and service delivery for job seekers, including claimants for Unemployment Compensation; and

To establish RESEA as an entry point into other workforce system partner programs for individuals receiving UC.

Program Design. As of this writing in July 2021, the required components of a state RESEA program are the following:[5]

- UI claimants selected for RESEA are scheduled for a mandatory initial RESEA meeting with appropriate staff in the local American Job Center (AJC).
- During the meeting, a staff person reviews the claimant's work search efforts to confirm that they are meeting eligibility requirements, ensures they are registered in the Wagner-Peyser employment services system, and provides a range of services to help them become reemployed.[6] The staff person may also refer the claimant to relevant services available through other WIOA partner programs. Prior to the COVID pandemic when in-person meetings were required, non-attendance at the initial meeting appears to have been common. Nearly a third of those selected never attend,[7] meaning they do not receive employment services through the RESEA program.[8]
- Consistent with the RESEA program's mandatory nature, if a claimant does not attend as scheduled, states are to respond by initiating the non-compliance process.
- Beyond the initial meeting, at state option, selected claimants may be required to attend up to two subsequent meetings.

Evidence-Based Features. The Statute specifies that RESEA interventions must be either demonstrated effective or under evaluation. Starting in FY 2023, increasing shares of each state's RESEA grant monies must be allocated to demonstrated effective interventions, irrespective of whether the intervention is

---

4 For our purposes, UC is equivalent to Unemployment Insurance (UI) payments.

5 From Unemployment Insurance Program Letter No. 13-21 (USDOL, 2021), the most current DOL RESEA program guidance at the time this report was written. Guidance is subject to change.

6 The minimum required elements of an initial RESEA meeting are a one-on-one UC eligibility review; customized labor market and career information; enrollment in the employment services program; development of an individual reemployment plan; and information and referral to additional reemployment services, as appropriate (USDOL, 2021).

7 For REA, see Klerman et al. (2019). Available evidence suggests similar no-show rates for in-person RESEA meetings (Trutko et al., 2022).

8 With COVID, many states have shifted to virtual meetings (phone and perhaps videoconference). Interviews with state staff suggest that COVID-era virtual meetings have much higher attendance rates. As of this writing, it is unclear whether and when states will return to in-person RESEA meetings and whether attendance rates at virtual meetings will remain high.

under evaluation. Finally, states are allowed to spend up to 10 percent of their RESEA grant on evaluation.

To be *demonstrated effective,* an RESEA intervention—either a whole-program model or a component of a program model—must

- have a sufficient number of evaluations, that
- meet standards for evidence quality established by DOL's Clearinghouse for Labor Evaluation and Research (CLEAR; https://clear.dol.gov/), and
- show that the intervention decreased UI weeks claimed and increased employment in the second quarter after the start of the UI claim (i.e., in Q2).

The text box (opposite) provides more detail on those requirements and CLEAR's role in rating whether an intervention has been demonstrated effective. Unemployment Insurance Training Letter No. 01-20 (USDOL 2019) provides more detail on these provisions, as of this writing. Some specific rating standards could change in the future.

CLEAR's Role in Deeming an RESEA Intervention Demonstrated Effective

CLEAR has a dual role in deeming an RESEA intervention *demonstrated effective*.

First, as is true for all CLEAR reviews, CLEAR will rate ***study evidence credibility*** and assign one of three ratings: High, Moderate, or Low. Random assignment studies of reemployment interventions have usually received a High evidence credibility rating. Most non-experimental (that is, non-random assignment) designs can at most receive a Moderate evidence credibility rating. In practice, many do not even receive that.

Second, considering evidence across all sufficiently credible studies of an RESEA intervention, CLEAR will assign an ***intervention effectiveness*** evidence rating. Intervention effectiveness ratings are based on evidence from studies that received at least a Moderate rating for evidence credibility. Like the study credibility ratings, CLEAR's RESEA intervention effectiveness ratings also use a set of categories that include High and Moderate. Interventions not rated High or Moderate are categorized as either Potentially Promising or Not Rated. For an intervention to be considered demonstrated effective, the intervention must receive a High or Moderate rating. To earn a High rating, CLEAR needs two studies of the intervention that meet standards and detect impacts on what this document refers to as the statutory outcomes: employment, UC duration. To earn a Moderate intervention rating, CLEAR needs one study of the intervention that meets standards and detects impacts on employment and one that detects impacts on UC duration. The two findings can come from the same study or from two different studies.

To receive a High rating, impacts must be detected with a statistical significance of $p<.05$. For a Moderate rating, impacts must be detected with a statistical significance of $p<.10$.

1.2. Completed Studies and Studies in Process

This section describes completed studies for RESEA's immediate predecessor program (Reemployment and Eligibility Assessment [REA]) or earlier reemployment programs. Differences between RESEA and REA are subtle. RESEA retains REA's core, required eligibility assessment and reemployment services elements. But under RESEA, DOL has emphasized that services should be more individually tailored to claimants' reemployment needs. How that individualization occurs is left up to states and is likely to vary across states.[9]

After discussing existing evidence, the section then considers studies of RESEA starting in late 2020 and 2021 that are not yet completed.

Completed Studies. RESEA shares a number of features with reemployment programs that came before it—REA and the Worker Profiling and Reemployment Services (WPRS) system in particular. A large

[9] Several other small differences between REA and RESEA exist. Staffing options also remained fundamentally similar between the two programs, but RESEA has included a move toward having both the eligibility assessment and reemployment services functions being conducted in a single session by a single staff member. Finally, claimant selection has changed. States are permitted to select WPRS claimants for RESEA. Consequently, more states use profiling models to select RESEA claimants.

evidence base exists on the effects of those programs (studies of 19 state programs as of this writing). Appendix A reviews the REA studies. Klerman et al. (2019) and CLEAR (2018) provide more detailed reviews of a broader set of reemployment interventions. The section below provides a high-level overview of what that research has found.

Most of the previous studies use an experimental research design in which eligible UI claimants are randomly selected for the RESEA-like program or not selected. Claimants who are not selected are not required to attend RESEA meetings (and thus do not receive its employment services), but they can access the usual reemployment services available to any job seeker at an AJC. The studies then use administrative data—either state UI wage data or National Directory of New Hires (NDNH) data—to measure UI duration and employment outcomes. Comparing the average outcome for the group of claimants selected for RESEA versus the average outcome for the group not selected yields the (causal) impact of RESEA. Appendix B discusses this design further.

In brief, studies consistently find that programs are effective in reducing UI duration—by half a week to a week, across a range of program models.[10] Studies of the impact of reemployment programs on employment or earnings are less common. Those that do study impact on those labor market outcomes often do not detect an impact. In part, this appears because sample sizes are too small: True impacts are in the range of 2 percentage points on Q2 employment and 2 percent on earnings. To reliably detect impacts of that magnitude, however, appropriate sample sizes would be large, in the range of 20,000 to 50,000 study participants.

Although there are many studies of the impact of reemployment programs, most estimate the impact of being selected for the program relative to not being selected. Few studies estimate the impact of components of those programs; that is, how impact varies with variation in the details of the intervention. One evaluation that has such a large sample, the REA Impact Study (Klerman et al. 2019), reports mixed evidence on a particular component: whether programs that require subsequent meetings have larger impacts than those that have just a single, initial meeting.

Similarly, few studies have large enough samples to explore how impacts vary with claimant characteristics at initial claim. Klerman et al. (2019) found evidence that REA programs' impacts were larger for claimants with lower earnings and smaller weekly benefit amounts at the time of program enrollment. Like earlier research on WPRS by Black et al. (2003), the REA Impact Study did not find evidence of larger impacts for those with higher risk of benefit exhaustion as measured by WPRS profiling scores.[11]

This review of the existing literature and the more detailed discussion in Appendix A suggest that sufficient sample size to reliably detect impacts is a fundamental challenge for studies of reemployment interventions. The discussion in Chapters 2 and 3 therefore focuses on addressing this sample size challenge.

Studies in Progress. To address the need to show that state RESEA programs are evidence-based, DOL's Office of Unemployment Insurance (OUI) has been encouraging states to launch experimental evaluations of their RESEA programs. OUI has also provided both general and state-specific evaluation technical assistance (TA) as part of the *Evaluation to Advance RESEA Program Evidence* producing this document.

---

[10] See Appendix A for support for the arguments in this and the next two paragraphs.

[11] Profiling scores are produced by statistical models that use a new claimant's pre-claim characteristics (and sometimes local labor market factors) to estimate the probability that the claimant will exhaust UC benefits before the end of their claim.

In addition, starting in September 2020, DOL formed a Learning Cohort of five states that are receiving more intensive evaluation TA, and may include more states as resources allow.

Nevertheless, as of mid-2021, states were in various stages of planning for such evaluations (see Appendix Section C.2 for detail on state evaluations), but no state had begun random assignment.[12] COVID and the challenges it brought to the UI system substantially pushed back evaluation timelines. As a result, few evaluations are likely to start before the summer of 2021. Combined with evaluation timelines that require several years for a study to be completed, this implies that CLEAR might deem a generic[13] RESEA program *demonstrated effective* in time for FY 2025 state RESEA Plans, but FY 2026 or even FY 2027 seem more likely. (See Appendix Section C.1 for discussion of timing.) The balance of this report refers to options toward this initial determination that a generic RESEA program is demonstrated effective as ***short-term options***. Options toward estimating the impact of components, which have even longer timelines, are referred to as ***longer-term options***.

1.3. Plan for the Balance of This Report

The balance of this report considers options for research that satisfies statutory requirements and in different ways produces evidence that can help states strengthen their RESEA program designs.

Chapter 2 considers evaluation options to build evidence on the impact of whole programs (RQ1) and how the impact of whole programs varies among subgroups of participants (RQ2).

Chapter 3 considers evaluation options to build evidence on the impacts of program components (RQ3) and how those component impacts vary among subgroups of participants.

Each of these RQs considers *impact*; that is, what difference the intervention makes. *Non-impact* activities are also useful in helping to design interventions or to produce complementary evidence that enhances the value of evidence produced through impact analyses. Specifically, Chapters 2 and 3 also consider these:

Report Research Questions

RQ1/Whole programs: What is the impact of being selected for RESEA versus not being selected for RESEA?

RQ2/Subgroups: How does the impact of RESEA vary with the characteristics of the claimant at initial claim?

RQ3/Components: How does the impact of an RESEA program vary with the components included and how they are provided?

RQ4/What works best for whom: How does the impact of components included and how they are provided vary with the characteristics of the claimant at initial claim?

- *How might estimates of impact help states improve policies and program designs?*
- *What are the options for evaluation designs that directly address the RQ by estimating impact?*
- *What are the options for non-impact research activities that would support answering the RQ?* Non-impact activities that this report considers include building consensus as to what to evaluate, developing data and software infrastructure to support the estimation of impact, evaluation TA, implementation studies, cost-benefit analysis, and synthesizing results across studies.

---

[12] In practice, state efforts in that direction were delayed by the need to focus UC staff efforts on dealing with the major statutory changes to the UC program in the Coronavirus Aid, Relief, and Economic Security (CARES) Act and processing the huge surge in UC claims with the onset of the COVID-19 pandemic.

[13] By "generic" we mean that CLEAR has found RESEA overall as effective, based on evidence from whichever states have completed evaluations. It does not imply effectiveness of specific RESEA program models that may vary in their particular characteristics.

Chapter 4 summarizes the findings, discusses broader goals, and considers what options seem most feasible and appropriate in the short term and longer term.

Throughout, this report focuses on the key issues for RESEA evaluations including: posing the right research question, selecting and properly implementing a design that will accurately estimate impact, and building sufficiently large samples to detect impacts and provide insights on changes to program design that could improve claimant outcomes.

For each of options identified in Chapter 2 and Chapter 3, the report provides:

- ***What?*** *This is a description of the option*. The description covers the basic elements of what activities the option entails.
- ***Who, when, how much?*** *This is a discussion of what is involved in carrying out the option.* This includes: (1) what entity would likely lead it (individual states, consortia of states, or DOL)? (2) when would it likely occur (in the short term or longer term)? (3) is it a higher-cost, medium-, or lower-cost option? While how much an option would cost would depend in large part on specifics of the statement of work for the project, we roughly define the cost categories as lower-cost: less than $250,000; medium-cost: between $250,000 and $999,999; higher-cost: $1 million or more.
- ***Why?*** *This is a description of the value of doing the option*. The description of how the option would produce or help produce information that is useful for states and DOL and considerations (such as logistical feasibility, timeliness, data quality, etc.) that may make the option attractive.

For options that have a defined end point, the cost is inclusive of all years of the project. A few activities, such as repeated annual surveys, are indefinite in length; for those, costs categorization reflects expected costs over a five-year period.

Appendix A surveys completed studies of the REA program. Appendix B describes the canonical design: random assignment with outcomes measured in administrative data. Appendix C discusses the status of state plans for studies of RESEA. Finally, Appendix D considers non-experimental designs.

2. Options for RQ1/Whole Programs and RQ2/Subgroups

The chapter discusses options for evaluating the impact of a whole program and options to support and build on findings from those whole-program evaluations. This approach can be used to address RQ1 and RQ2.

- **RQ1/Whole Programs: What is the impact of being selected for RESEA—relative to not being selected for RESEA?**
- **RQ2/Subgroups: How does the impact of RESEA vary with the characteristics of the claimant at initial claim?**

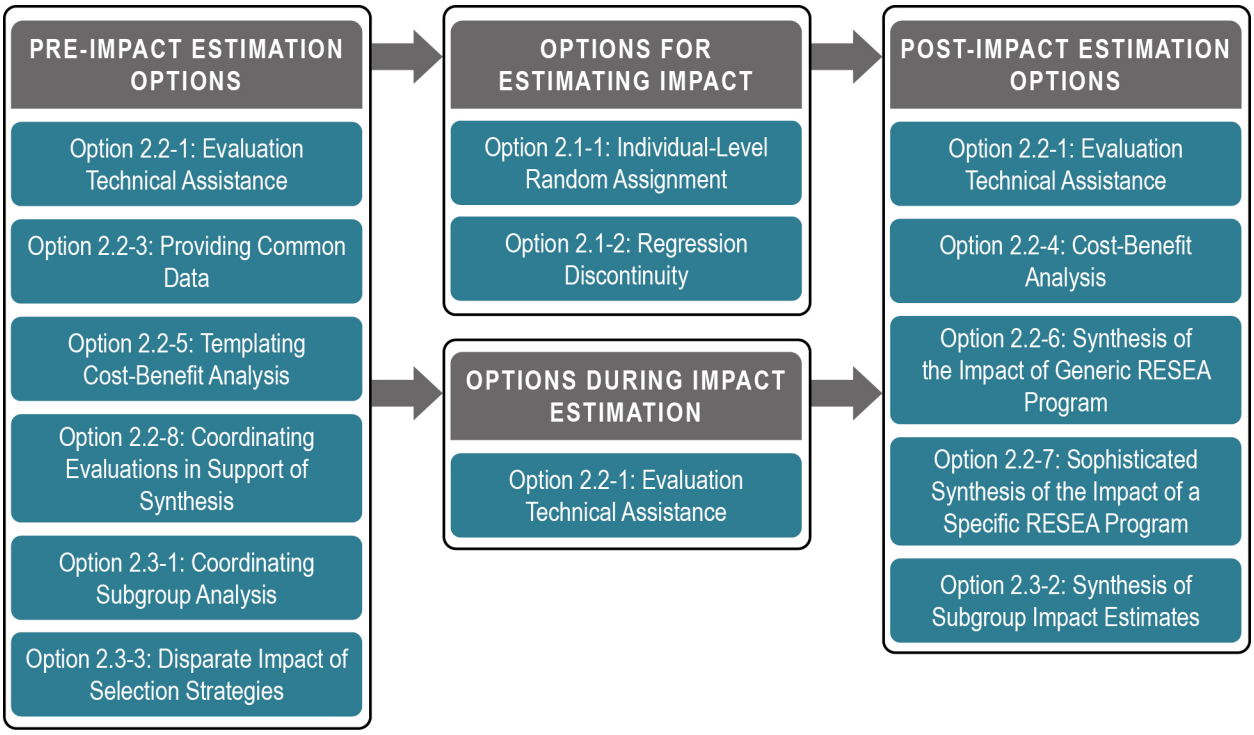
Both of these RQs are relevant for meeting RESEA statutory requirements for eligibility for funding; they also are relevant for program management purposes beyond meeting the statutory requirement. Regarding the statute, demonstrating that whole programs are effective (RQ1) is likely the most straightforward way to meet the statutory requirement to demonstrate the effectiveness of states' RESEA interventions. The evidence from a study that answers RQ1 concerns all of a program's intervention components collectively. Beyond the statutory motivation, most programs have an interest in understanding how effective their services are, and potentially providing information on areas for improvement. Given that states are likely implementing improvements to their programs on an ongoing basis, estimates of program overall effectiveness can provide an important baseline of information to which findings from later whole-program evaluations can be compared.

Findings from studies examining RQ2 may help inform a state's decisions regarding which claimants to select for RESEA. If a state has shown that its program has a larger impact for some subset of claimants than for others (e.g., by industry, replacement rate, profiling score, etc.) and the state focuses selection on that subset of claimants, then that may serve to demonstrate that the state's approach to selecting claimants is effective for statutory purposes. Findings from studies of RQ2 by gender or race/ethnicity may also indicate possible inequities that the state could subsequently understand and address.

Exhibit 2-1 shows graphically how the options for answering RQ1 and RQ2 fit together. On the left are options to be implemented *prior* to impact analysis. In the middle are options to conduct impact analysis or that would be implemented in parallel with the impact analysis. On the right are options to be implemented *after* impact analysis.

The balance of this chapter is organized as follows. Section 2.1 presents options for estimating the impact of being selected for RESEA relative to not being selected for RESEA (**RQ1/Whole Programs**). Section 2.2 presents options that do not directly estimate impact for either RQ1 or RQ2, but instead support estimating impacts (some options would occur before impact estimates, some in parallel with impact estimates, and some after impact estimates). Section 2.3 presents options in support of estimating impact for **RQ2/Subgroups**. For each option, we discuss what, who, when, how much, and why. Appendix B provides more detail on experimental (random assignment) study designs, and Appendix D provides more detail on non-experimental study designs.

Exhibit 2-1. Inter-Relation of Options for RQ1/Whole Programs and RQ2/Subgroups



Source: Abt Associates.

2.1. Options for Estimating Impact of Whole Programs

This section describes evaluation options for addressing **RQ1/Whole Programs**: *What is the impact of being selected for RESEA—relative to not being selected RESEA?* Specifically, it discusses individual-level random assignment (**Option 2.1-1**) and regression discontinuity (**Option 2.1-2**) study designs. The section ends with a brief discussion of several other designs for whole-program research and explains why those other options do not appear appropriate.

Addressing RQ1 can potentially justify funding for the program as a whole. Almost all previous evaluations of REA and other reemployment programs have focused on this whole-program research question of the impact of being selected for the program compared to not being selected. For example, OUI used whole-program evaluations of the REA program to justify the REA program and its successor RESEA (Poe-Yamagata et al., 2011; Michaelides & Mueser, 2016). In addition, 2018 amendments to the Social Security Act[14] requires states to show that increasing percentages of their RESEA grant are spent on demonstrated effective programs.

The easiest way to meet the Act's requirement appears to be to show that the whole RESEA program is demonstrated effective—through one or more whole-program evaluations. Given that decisions about the need for sufficient evidence to meet grant requirements have not been made and that evidence standards could shift over time, this chapter takes a broad view of options for addressing this whole-program RQ.

[14] The new evaluation and evidence standards are contained in Section 306 of the Act.

Option 2.1-1 Individual-Level Random Assignment (RQ1/Whole Program)

What: This is the canonical design (Appendix B) used by most previous evaluations of reemployment programs (see Appendix A). Under the canonical design, rather than systemically selecting UI claimants for RESEA, UI claimants are selected randomly.

This option would estimate the impact of a state's RESEA program on the statutory outcomes of employment UI claim duration. These outcomes would be measured using administrative data. For this research question, the only data required are randomization status (intervention/control),[15] weeks of benefits (probably also dollars of benefits paid), and quarterly earnings (from which employment is imputed from positive earnings). Additional data on claimants' background characteristics, RESEA meeting attendance, and reemployment services received would provide additional insight but are not strictly necessary.

Who, When, and How Much: To reliably detect impacts of the size that might be expected on both earnings and UI duration outcomes, an evaluation would want a sample of 30,000 to 50,000 RESEA-eligible UI claimants, with roughly half selected for RESEA and the other half not selected for RESEA (see Appendix B). For example, a state with 30,000 RESEA-eligible UI claimants per year would randomly choose which 15,000 of them to select for RESEA; the other 15,000 would serve as the control group. Many states can achieve these samples sizes in about a year of conducting random assignment. Most smaller states with smaller UI caseloads or RESEA programs can achieve these sample sizes by conducting random assignment for two or three years.

Random assignment is a prospective evaluation design.[16] Only after random assignment is conducted, outcomes occur, and data measuring those outcomes become available can analysis begin. For states that need only a year of random assignment, results can likely be available three to four years from the decision to implement this option and putting in place an evaluator. For states needing two or three years of random assignment, timelines would be one or two years longer. (For more on timeline issues see Appendix Section C.1.)

This is a higher-cost option. Costs include finalizing the design and writing a design report, setting up and monitoring random assignment, acquiring and processing appropriate data files, conducting the analysis, and writing up and presenting the results.

Why: Because the assignment of UI claimants is random, any systematic differences in outcomes must be due to selection to RESEA. Such individual-level random assignment provides the strongest evidence of program effectiveness. Furthermore, for a given sample size, individual-level random assignment yields the most precise information. A more "precise" estimate is one with a smaller **confidence interval**; sometimes called "margin of error." That individual-level random assignment produces estimates that are more precise means that the sample sizes needed to detect impacts are smaller than the sample sizes needed for other study designs. Given that sufficient precision will often be the key challenge, smaller required sample sizes are a major advantage. In addition, this option is relatively certain to yield results that will meet CLEAR standards for a study's evidence credibility. Finally, other options—including those discussed in the balance of this section—are more challenging to complete successfully.

---

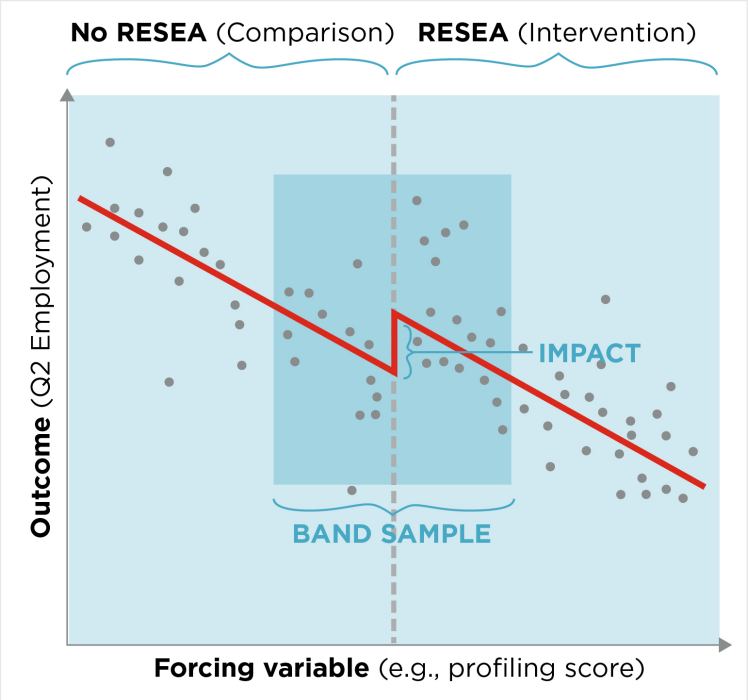
[15] Randomization assigns claimants either to a group selected for RESEA ("intervention group) or to a group that cannot access RESEA ("control" group).

[16] Such prospective evaluation designs are in contrast to retrospective designs considered below. Retrospective designs do not require changing anything about how the program runs. Thus, analysis can be conducted immediately—on outcomes that occurred in the recent past (perhaps since the start of RESEA in 2015). See Appendix Section D.1 for more on the prospective design/retrospective design distinction.

Option 2.1-2 Regression Discontinuity (RQ1/Whole Program)

What: This design requires that claimants be selected for the program based on whether they score above or below some cutoff value. It is then plausible to assume that those who are just on either side of the cutoff are otherwise similar and any differences can be controlled for by the score.[17] In a regression discontinuity design, all RESEA-eligible claimants with UI profiling scores above the cutoff are selected for RESEA. Those with profiling scores below the cutoff are not selected. As shown on Exhibit 2-2, this design then compares outcomes of claimants just on either side of the cutoff to provide an estimate of RESEA's impact. Information from the initial claim is likely easily available and would support standard specification tests (McCrary, 2008).

Exhibit 2-2. Regression Discontinuity Illustration



Source: Abt Associates.

Who, When, and How Much: Most states that use a profiling score as part of their strategy for deciding whom to select for RESEA can use this study design (Trutko, et al., 2022). If the profiling score is the only criterion used for RESEA selection, applying regression discontinuity is relatively straightforward. If—as is often true—other criteria are used in addition to the profiling score—such as AJC-specific capacity limitations—applying regression discontinuity is more difficult and might not be possible.

States that have used a profiling score in the recent past might be able to apply regression discontinuity retrospectively; that is, to claimants selected or not selected in the past few years. In that case, the timeline for impact estimates might be one year. Alternatively, states might apply regression discontinuity prospectively—that is, on those selected in future years. In that case, the timeline is likely to be similar or longer than for random assignment—four or more years.

The major challenge of applying regression discontinuity is sample size. To attain sufficient precision, a state would need roughly 200,000 RESEA-eligible claimants "near" the profiling score cutoff.[18] On their own, only a few states will have those samples, even over multiple years; however, consortia of states may be able to attain such sample sizes.

Although CLEAR has not yet published standards to rate studies that use regression discontinuity designs, CLEAR has indicated that it considers regression discontinuity to be a design that can produce strong

---

[17] The appropriate cutoff value could vary with time (e.g., calendar week) or place (e.g., America's Job Center).

[18] The estimate follows from recent analyses suggesting that sample sizes for regression discontinuity are four or more times larger than for the equivalent random assignment design (Deke and Dragoset 2012).

evidence. In practice, applying regression discontinuity requires subtle, complex decisions that depend on the characteristics of the study's sample. The appropriate technical level and experience of the evaluator is moderately higher than required for random assignment designs. As a result, compared to individual-level random assignment designs, regression discontinuity designs are less likely to receive a Moderate or higher CLEAR study evidence credibility rating.

This is a medium-cost option. With two exceptions, tasks for this option are similar to those for Option **2.1-1/Individual-Level Random Assignment**. On one side, costs would be lower because there is no need to implement and monitor random assignment. On the other side, costs would be higher because analysis is more complicated and time intensive.

Why: When feasible, regression discontinuity provides highly credible evidence—though not as credible as random assignment—without the inconvenience of random assignment. Furthermore, if the state used profiling scores in the past, results could be available much sooner than for random assignment.

Other Potential Whole-Program Designs, Considered

This section discusses three more options for estimating whole-program impact, based on a scan of a larger set of possible designs. (See Appendix D for more discussion of these and other designs.) These three designs are feasible under certain conditions; but because RESEA evaluations seem unlikely to meet these conditions, they are not discussed in detail.

One option is **group random assignment**, where outcomes for a group of *offices* (at which RESEA meetings are held; usually American Job Centers) randomly chosen to provide RESEA are compared to outcomes for the other offices that do not provide RESEA. This group random assignment design requires samples several times larger than does individual random assignment. Samples of that size are likely infeasible almost everywhere.

A second option (sometimes called **regression adjustment** or **propensity score matching**) compares outcomes of claimants who are selected for RESEA to outcomes for claimants who are not selected for RESEA (for some non-random reason). These approaches can be used to estimate impact when we it is possible to match claimants who were selected for RESEA to other claimants who are "otherwise like them" but were not selected.

Given that RESEA selects claimants with particular characteristics and does not select claimants who lack those characteristics, it will usually not be possible to find people who are "otherwise like them." Some claimants are not selected because they are union attached or have a definite recall date, but there are no claimants who were selected who are union attached or have a definite recall date. Some claimants are not selected because they have a low profiling score, but there are no claimants with similarly low profiling scores who were selected (at least in this office, in this week).

Finally, if a state implemented RESEA on a rolling basis across the offices in the state in a purposive way, an **interrupted time series (ITS)** design might be possible. This approach would compare the change in outcomes immediately before and after implementation of RESEA. Pre-COVID, we are unaware of any state that has recently rolled out RESEA to enough offices to make this design feasible. It may be possible that states will relaunch their RESEA programs post-COVID across the state in a systematic way making this decision possible.

Regardless, CLEAR's requirements for an ITS can be difficult to satisfy. Among other things, to satisfy CLEAR's standards for ITS, the RESEA rollout must be *intentionally* staggered by the evaluator such that the order in which the intervention is implemented in different parts of the state is unrelated to characteristics of those areas that might be related to outcomes (e.g., local labor market condition, composition of the workers in that area, capacity of AJCs). That is, the rollout must be effectively, if not

explicitly, random. As this is seldom the case (and most states have a strategy for how they implement RESEA across the state), most studies using an ITS design are unlikely to pass CLEAR evidence credibility review.[19]

For readers who are interested in further discussion of these three designs, the *RESEA Evaluation Toolkit* (Mills De La Rosa et al., 2021) provides additional detail. However, because of the considerations discussed here, ITS should only be considered if other options are not feasible. ITS designs are likely to not receive a High or even Moderate study rating.

2.2. Options to Support Estimating Impact of Whole Programs

This section describes options that, though not directly estimating impact, would *support* estimating impact or would *make use of* individual studies' impact estimates. Options considered are evaluation TA (**2.2-1**), building evaluation infrastructure (**2.2-2** and **2.2-3**), cost-benefit analysis (**2.2-4** and **2.2-5**), and synthesizing impact results (**2.2-6**, **2.2-7**, and **2.2-8**).

Option 2.2-1 Evaluation Technical Assistance (RQ1/Whole Programs)

What: This option would provide group and one-on-one evaluation TA to state RESEA evaluations. Experience with RESEA and with other tier-based evidence programs[20] suggests that, when unassisted, states and their evaluators often find it challenging to design, implement, and report the results of evaluations in ways that will meet CLEAR standards. Since 2018, DOL has funded both group evaluation TA—*RESEA Evaluation Tool Kit* (Mills De La Rosa et al., 2021), other written products, webinars—and one-on-one TA.

Who, When, and How Much: If it pursued this option, DOL would continue to offer group and/or one-on-one evaluation TA to be used by states at their discretion. The RESEA evaluation funded general evaluation TA through September 2021 and some additional customized evaluation TA for a small number of states through September 2023. In contrast, the initial round of RESEA evaluations likely will not have findings until 2025 or later. **Option 2.2-1** would provide similar activities past current funding, until the first round of whole-program evaluations concludes. Of course, the expectation is that evaluations will continue beyond that as a regular part of RESEA program operations.

Total cost would vary with the number of states supported, but over the course of five years, this would be a medium- to higher-cost option. More than most activities, such evaluation TA can be scaled to the available funds. More funds allow more group activities and more intensive one-on-one assistance to more states.

---

[19] *CLEAR Causal Evidence Guidelines* says this explicitly: "Although ITS designs can receive High or Moderate ratings for causal evidence, CLEAR leadership anticipates that only a small number of these studies in the topic areas examined by CLEAR will receive a Moderate rating, and few (if any) studies will be highly rated" (2015, p. 12, ftn. 8). https://clear.dol.gov/reference-documents/causal-evidence-guidelines-version-21

[20] Tier-based grant programs provide more funding to grantees that adopt evidence-based program designs. Examples of tier-based grant programs include U.S. Department of Education's programs Striving Readers (https://www.ed.gov/category/program/striving-readers), Investing in Innovation (https://www.ed.gov/open/plan/investing-innovation-i3), and Education Innovation Research (https://oese.ed.gov/offices/office-of-discretionary-grants-support-services/innovation-early-learning/education-innovation-and-research-eir/); U.S. Department of Health and Human Services' programs HomeVEE (https://homvee.acf.hhs.gov/) and Teen Pregnancy Prevention (https://opa.hhs.gov/grant-programs/teen-pregnancy-prevention-program-tpp/about-tpp); and DOL's Workforce Innovation Fund (https://www.doleta.gov/workforce_innovation/).

Why: Several recent evaluation TA experiences suggest a key role for evaluation TA. Unless evaluations are highly experienced, they often find it difficult to produce impact evaluations that will meet CLEAR standards. Meeting standards is most likely if the selected evaluator has previously conducted several similar evaluations that have met standards. Preliminary discussions with some states as part of ongoing evaluation TA suggest that some of them do not expect to contract with evaluators with that level of experience. They instead intend to contract with local (but less experienced) private evaluation firms or some part of the state's university system, or else use a research office within their own agencies.

For evaluators lacking relevant experience, a combination of general evaluation TA and robust state-specific evaluation TA make it substantially more likely that a state-sponsored evaluation produces findings that the field can consider credible and that can meet CLEAR standards. Other evaluation TA efforts for tiered evidence programs suggest that in the absence of ongoing evaluation TA, fewer than half of the targeted evaluations will meet standards. With the moderate intensity, but voluntary, evaluation TA being provided to the Learning Cohort (see Appendix Section C.2), that rate likely rises to above half. With robust and mandatory evaluation TA, that rate probably rises to three-quarters (see Epstein, 2022).

Option 2.2-2 Providing Common Analytic Tools (RQ1/Whole Programs)

What: Whole-program random assignment evaluations (**Option 2.1-1**) are common but take time and commitment to implement properly. This option would develop a statistical software program to estimate impact and precision. In practice, this program would likely build on some existing statistical software package (e.g., macros in SAS, Stata, or R). The option would also generate a user guide for the software program.

Who, When, and How Much: This would be a DOL-directed activity. Ideally, it would be completed before states start to analyze the data from **Option 2.1-1**. Considering both the cost to develop initial software and documentation, to maintain the software over a period of years (e.g., to address issues that arise and to add requested features), and to provide support to states and evaluators using the software, this is a lower to medium-cost option.

Why: The basic analysis for experimental study designs is routine. Asking each evaluation to develop its own analytic software is therefore inefficient. Furthermore, although the basic analysis is straightforward, there are typical errors (e.g., improperly weighting for varying randomization fractions). Providing common analytic tools would likely lower costs for state evaluations and—more important—eliminate hard-to-detect errors.

Option 2.2-3 Providing Common Data (RQ1/Whole Programs)

What: The standard approach to state-specific evaluations uses state-specific data. This option would collect state data and make them available for evaluation in a common format. The text box (opposite) lists possible common data items. These items provide enough information to estimate impact using NDNH data on outcomes (i.e., quarterly UI benefits paid, employment, and earnings). A broader vision of the common data set might also include outcomes (e.g., UI weeks, UI dollars, employment, and earnings by quarter).

Details would vary by data element. Similar to its Participant Individual Record Layout for workforce data, DOL might develop a common individual-level RESEA record format for

Candidate Data Items for Common Data File (*Option 2.2-3*)

- Social Security number
- UI initial claim start date
- Eligibility for RESEA selection (yes/no)
- Selected for RESEA (yes/no)
- Profiling score
- Randomization status (selected, not selected, not randomized)
- [Optional] Attended initial RESEA meeting (yes/no)

UI claimant data. States might report those data to DOL quarterly, in the specified format, as a grantee responsibility.

Who, When, and How Much: This would be a DOL-directed activity. The actual task might be performed by DOL staff or by a contractor. Ideally, DOL would maintain this activity such that state evaluations could rely on these data. This is a higher-cost option. Cost would vary with contractor role and scope. Costs are driven by the need to work separately with each state and every state has a different data system.

Why: Much of the cost of **Option 2.1-1/Individual-Level Random Assignment** is induced by the effort required to understand and process multiple data systems. Usually those costs are moderate, but evaluators need to budget against the possibility that something might go very wrong (e.g., data are maintained in a format, such that the state has trouble providing data to an outside evaluator). **Option 2.2-3** would create common databases and software programs to process them. Doing so would radically cut the cost of state evaluations, but moderately increase state costs for ongoing reporting.

Option 2.2-4 Cost-Benefit Analysis (RQ1/Whole Programs)

What: This option would estimate the incremental cost and benefit per claimant selected for RESEA. Costs considered would include not only the cost of RESEA services, but also services provided by the broader workforce system (e.g., Wagner-Peyser funds), and savings due to lower UI benefits paid and higher UI taxes paid. Such analyses consider multiple perspectives: the claimant, the state, the federal government, all government, society as a whole. Benefits are derived from the impact estimates, where outcomes considered are expanded to include dollars of benefits paid and earnings.[21]

Who, When, and How Much: This would be a state-directed activity. Ideally, cost data would be collected while or shortly after services are provided to the randomly assigned claimants. Final cost-benefit analysis results require the final impact estimates. A cost-benefit analysis is usually paired with the impact analysis. Based on past studies, states can anticipate this to be lower cost as an add-on to an impact study. This cost estimate includes collecting cost data, doing cost-benefit analysis, and writing up and briefing the results.

Why: Justifying a program often requires not merely showing that the program has impacts, but also showing that the value of the implied benefits exceeds the value of the costs. Cost-benefit analysis measures both and then compares them.

Option 2.2-5 Develop a Template for Cost-Benefit Analysis (RQ1/Whole Programs)

What: This option would explore the details of how to conduct a cost-benefit analysis for an RESEA whole-program evaluation. This option would generate a "How-to Guide" and evaluation TA materials (i.e., a template) to support state cost-benefit analyses (**Option 2.2-4/Cost-Benefit Analysis**).

Who, When, and How Much: This would be a DOL-directed activity. Ideally, the TA materials would be available so that states can collect the appropriate cost data while they are doing random assignment (**Option 2.1-1/Individual-Level Random Assignment**). This is a lower-cost option.

---

[21] Ideally, benefits would include changes in job characteristics beyond earnings. No other job characteristics appear to be recorded in administrative data. Given appropriate sample sizes, measuring other job characteristics through a survey seems cost prohibitive.

Why: There are few cost-benefit analyses of reemployment programs. The general principles of cost-benefit analysis are well understood (Boardman et al., 2017); however, given imperfect data, applying those general principles to a particular intervention is something of an art. It can be a particular challenge in the context of the workforce system, because many different programs serve participants and funds are to some degree fungible across programs for state and local agencies.

A template and associated guide would provide specific guidance for cost-benefit analysis for RESEA programs. In other fields, there are how-to guides or at least worked examples—for welfare-to-work programs (Greenberg & Appenzeller, 1998) and for job training programs (Mastri & McCutcheon, 2018; Schaberg & Greenberg, 2020). Those guides make it feasible for less experienced evaluators to conduct high-quality cost-benefit analyses. Inasmuch as cost-benefit analysis is seen as valuable, creating such a template/guide seems worthwhile.

Option 2.2-6 Synthesis of the Impact of a Generic RESEA Program (RQ1/Whole Programs)

What: **Options 2.1-1** and **2.1-2** estimate the impact of a single RESEA program. This option would synthesize the results of RESEA and related reemployment evaluations that have met CLEAR study credibility standards and effectiveness standards (which presumably would be a part of regular CLEAR activities). This option would combine information across evaluations.

Consistent with the standards for rating RESEA intervention effectiveness as of the time of writing (see USDOL, 2019), the simplest version of this option would be a count: How many studies meet standards and find impact—at varying levels of study credibility and strength of evidence? A more sophisticated version of this option would be to perform meta-analysis (see Appendix Exhibits A-1 and A-2).

Who, When, and How Much: This may be a DOL-directed activity. Ideally, it would begin as studies of the impact of RESEA are completed—likely starting about 2025—and be updated annually. This is a lower-cost option. Costs include one-time costs to set up the basic analysis, plus per-study costs to extract required information.

Why: Such meta-analyses are a more informative way to combine individual study results for program decisions and operations. The simple "vote counting" approach is straightforward to implement. When there are multiple studies, however, vote counting makes inefficient use of the available evidence. Meta-analyses make more efficient use, giving more weight to estimates from more precise evaluations, but also using the estimates from less precise (and even not statistically significant) evaluations.[22]

Meta-analysis might also suggest an alternative way to deem a generic RESEA program effective. DOL's current planned approach involves vote counting; that is, deems RESEA to be effective once two studies of sufficiently high quality find clear statistical evidence of impact. An alternative would be to combine all available studies of sufficiently high quality using meta-analysis.

Such a meta-analytic approach is potentially important because it is possible that a meta-analytic approach would yield clear statistical evidence of favorable impact of a generic RESEA program *before* the vote counting strategy—perhaps after only a few months of random assignment. This alternative strategy would benefit from states publishing—or at least providing to DOL and to CLEAR—incremental results, perhaps quarterly or semi-annually. Given that some states seem likely to start random assignment

[22] On these issues see Valentine et al. (2017).

in late 2021 (see Appendix Section C.3), the first quarterly results might be available as soon as early 2023.[23]

Option 2.2-7 Sophisticated Synthetic Estimates of the Impact of a Specific State's RESEA Program (RQ1/Whole Programs)

What: This approach would combine information from all completed evaluations (that satisfy CLEAR standards for evidence quality) to estimate the current impact of a given state's RESEA program. This approach is particularly helpful to address sample size limitations in smaller states. Specifically, this approach would use the estimates from other states to generate better estimates for the given state. Thus, some small state estimates that would not be statistically significant when considered alone might be statistically significant using these statistical methods.

Such an effort would do so by "borrowing strength" across states and over time. Estimates that are for the state under study and that are more recent would get more weight. Estimates for other states and less recent estimates would get some, but less weight.

In the short term, the feasibility of such sophisticated synthesis methods could be explored using estimates from REA evaluations. In particular, open questions exist about the number of completed studies needed to apply these methods. Those questions could be explored through applying these methods to REA studies.[24]

Technical Terminology Sidebar:
Possible methods to create sophisticated synthetic estimates of the impact of a specific state's RESEA program (*Option 2.2-7*)

Small area estimation

Empirical Bayes

Kalman filtering

See: Molina & Rao (2010), Rao (2014)

Another, longer-term use of these meta-analysis approaches is to consider changes in program impact size over time. Under this scenario, states run evaluations based on those claimants selected for RESEA in one or more years. States and CLEAR then use that evidence as the basis for establishing whether their RESEA program is considered effective for many years into the future. These methods can be used to explore how program impacts are likely to vary over time.

Four sources of variation over time appear relevant: (1) business cycle variation; (2) explicit changes in program model (e.g., adding a second meeting); (3) changes in the characteristics of claimants; and (4) otherwise unexplained changes in the details of a state's program (including general program improvement for reasons not explained by observed program model characteristics, such as stronger management procedures). State year-by-year estimates from multi-year evaluations are enough to apply simpler versions of these methods. Once such methods are applied, the results estimate can be used to predict the impact of RESEA programs in years in which there were no evaluations. These methods can also be used to assess how precise those predictions are. That information could then be used to formally assess how often DOL might want states to re-estimate the impact of their RESEA programs.

---

[23] See Appendix Section C.1's discussion of state evaluation timelines. Suppose a state started random assignment on October 1, 2021. Then, Q2 employment for those people would occur in 2022Q2. State quarterly wage data on 2022Q2 would become available in 2022Q4. Prompt data process, analysis, and write up might yield reports on that first quarter of random assignment in 2023Q1. Details matter, but a meta-analysis might be able to show clear statistical evidence of impact as early as mid-2023.

[24] Analyses could then use the resulting estimates to do Monte Carlo simulations to explore the properties of the resulting estimates

Who, When, and How Much: This would be a DOL-directed activity. It could be piloted using REA studies.

An analysis for RESEA studies would not be appropriate until there are estimates from approximately six evaluations and those evaluations have been reviewed by Option 2.2-6. Ideally, such an effort would issue guidance about what individual evaluations should report (e.g., annual estimates, even when not statistically significant; properly computed standard errors; see **Option 2.2-8/ Coordinating Evaluations in Support of Synthesis**). This is a lower-cost option for a single round, but a medium-cost option if multiple yearly updates are included. Costs cover building/updating the database, running the analysis, and writing up and briefing the results.

Why: As other evidence-based programs have done, DOL strategically set the RESEA evidence standards low, such that fulfillment of the statutory requirement would be achievable for states. In particular, the current evidence standards implicitly assume that all RESEA programs are identical and aim to show that such a generic RESEA program is effective.

For the purposes of identifying the most effective programs and program strategies—and moving the field toward those strategies—this approach is not ideal. DOL recognizes that issue and has indicated that it may raise intervention rating thresholds in the future as more evidence becomes available. Existing evidence suggests that the impact of RESEA programs likely varies across states and across time. This means that there is value to establishing effectiveness for individual states on an updated basis. Furthermore, for the smallest states, there simply is not enough sample to reliably estimate impact annually. Combining information across states has the potential to address these challenges.

Option 2.2-8 Coordinating Evaluations in Support of Synthesis (RQ1/Whole Programs)

What: Develop and disseminate standards for what to estimate and report for impact evaluations. Then encourage (perhaps require) states to adopt those standards.

Who, When, and How Much: This would be a DOL-directed activity. Ideally, such an effort would be completed before states begin to generate impact estimates.

This is a lower-cost option. Costs cover developing the standards and creating materials to help states conduct their evaluations consistent with the standards. This option may involve developing a draft set of standards, circulating them for comment, and then revising into a detailed document and a quick reference for evaluators.

Why: Cross-evaluation syntheses are stronger if estimates are computed and reported in common ways. In addition, cross-evaluation syntheses can often use information that is of little interest to the evaluation itself. For example, a cross-evaluation synthesis would be interested in annual estimates, even if a state needs to pool estimates from several years in order to detect an impact. By pooling annual estimates across states, such a synthesis could explore how quickly impacts change over time and how they vary with local economic conditions (e.g., the state unemployment rate).

2.3. Options for Subgroup Analyses

This section considers options related to **RQ2/Subgroups**: *How does the impact of RESEA vary with the characteristics of the claimant at initial claim?* Groups of claimants that might be considered include higher (versus lower) profiling score, weekly benefit amount, education, and age. There might also be interest in racial and ethnic groups.

Unlike whole-program impact evaluations, subgroup impact evaluations yield insights to improve outcomes. Specifically, if a state knows that the program does more to improve the employment outcomes of one group of claimants than it does for other group of claimants, the state can use those subgroup

findings to focus RESEA selection more on the claimants who benefit most from the program. Doing would increase the existing program's average impacts.

Alternatively, subgroup findings may help states identify and address inequities. For example, suppose a state's evaluation found that its program did less to improve the employment outcomes of Hispanic workers than it did those of non-Hispanic workers with similar employment backgrounds. That evidence may suggest an area for state and local areas to investigate further to better understand the source of the disparities and corresponding changes that might be made to improve services to Hispanic participants. Those efforts may include evidence-building approaches described later in this report (**Option 3.3-3/Implementation Studies** and **Option 3.3-2/Deliberate Program Development**).

This discussion of subgroup options is included in the same chapter as options for addressing **RQ1/Whole Programs** because most of the issues are identical or quite similar for both.

- Random assignment is the preferred design for RQ2, just as it was for RQ1 (**Option 2.1-1/Individual-Level Random Assignment**). Random assignment for a subgroup impact evaluation and a whole-program impact evaluation are performed identically.[25] The only difference is estimation of differential subgroup impacts in the resulting data.
- RQ2/Subgroups can also be explored with **Option 2.1-2/Regression Discontinuity** (RD). That approach is less preferred for RQ2, however, because appropriate sample sizes are larger than for a whole-program evaluation and because RD cannot be used to explore whether impacts vary with profiling score. The RD sample only includes (and the estimates only apply to) those near the cutoff.
- Data requirements are nearly identical. The only difference is that information is needed on which claimants are in which subgroup. That information should be readily available in administrative data—from the claim (*age, education*), from the earnings history used to review the claim (*recent employment, recent earnings*), and from the disposition of the claim (*weekly benefit amount, maximum weeks*).
- At the state level, analysis is similar. Rather than estimating impact for the entire sample (the whole program), the evaluation estimates impact within each subgroup—and then tests for differences of impacts between the subgroups.

Given this similarity between addressing RQ1 and RQ2, most of the options to *support* estimating impact of whole programs (discussed in Section 2.2) carry over nearly directly for **RQ2/Subgroups**. The only difference is that the options need to make the obvious and simple generalization to consider subgroups. Options requiring no or only obvious and simple generalizations are these:

- **Option 2.2-1/Evaluation Technical Assistance**—where the TA needs to consider subgroups, but the changes are minor and easy to implement.
- **Option 2.2-2/Providing Analytic Tools**—where the tools need to consider subgroups, but the changes are minor and easy to implement.
- **Option 2.2-3/Providing Common Data**—where the data needs to be expanded to include subgroup identifiers, but the changes are minor and easy to implement.

---

[25] States often never select some (otherwise eligible) for RESEA. If a state wants to estimate how impacts vary, between usually included and excluded groups, then random assignment must include these otherwise excluded groups.

- **Option 2.2-4/Cost-Benefit Analysis**—where the analysis needs to be conducted separately for each subgroup, but the changes are minor and easy to implement.
- **Option 2.2-5/Develop a Template for Cost-Benefit Analysis**—where the template needs to be expanded to include subgroups, but the changes are minor and easy to implement.

Approximate Sample Sizes that Are Appropriate for Different Kinds of Studies

Random Assignment impact evaluations:

Whole programs: 30,000 – 50,000 (roughly half program participants and half non-participants).

Subgroups: 150,000 – 500,000

Quasi-experimental impact designs: 2 – 12 times larger than for a random assignment evaluation of the same research question.

Note: The estimates above are for studies of the impact of RESEA on employment outcomes. They are based on past studies of REA. Exactly how large of a sample a study needs to detect impacts, depends on how large the program's impacts actually are. The larger the impact, the smaller the sample needed.

Nevertheless, in a crucial way, addressing RQ2 is different from addressing RQ1. Appropriate sample sizes are much larger—four, 16, or more times larger.[26] Thus, whereas a whole-program experimental impact evaluation (**Option 2.1-1**) is likely to need samples of 30,000 to 50,000 to detect statistically significant impacts on employment, a subgroup experimental impact evaluation needs samples that are multiple times larger. How much larger depends on the size of the subgroups. At best, if the study is examining two similarly sized subgroups that each constitute about half of the sample, then the needed sample sizes are four times larger than for a whole program evaluation.[27] For smaller subgroups, the sample sizes needed can be 10 or more times larger.

---

[26] The other key issue for sample sizes is the size of the differential impact. The four-times-as-large case arises when the two subgroups are equally sized and the differential impact is as large as the pooled (overall) impact. So, for example, suppose that the overall impact on Q2 employment is 2 percentage points (approximately the average for REA; see Appendix Exhibit A-2). One might explore the differential impact of *weekly benefit amount* on Q2 employment by dividing those selected in half. Doing so ensures that the two subgroups selected into the study's intervention group are the same size (the best case/smallest sample size). Then suppose for one subgroup (say those with *weekly benefit amount* below the median) the estimated impact is 3.0 percentage points, whereas in the other subgroup (say those with *weekly benefit amount* above the median) the impact is 1.0 percentage points. In this example, the differential impact is 2.0 percentage points (= 3.0 − 1.0), which equals the pooled impact (also 2 percentage points). In this case, the appropriate sample size is four times that for the pooled impact (100,000 vs. 25,000). This would be a large differential impact. Suppose instead that the differential impact is half the pooled impact; that is, maybe 1.5 versus 2.5, for a differential impact of 1.0, which is half the pooled impact of 2 percentage points. Now the appropriate sample size is 16 times as large (400,000 vs. 25,000).

[27] The sample size in the text is to detect a differential impact of the same size as the overall impact. Consider a study with a pooled impact of 2 percentage points (and the number of men and women in the sample is roughly equal). Then this is the sample required to detect an impact of 3 percentage points in one group and 1 percentage point in the other group—that is a differential impact of 2 percentage points (3 percentage points – 1 percentage point) would be four times as large as the sample required simply to detect a pooled impact of 2 percentage points. Note also that this sample size would detect an impact in the group with the larger estimated impact, but no impact in the group with the smaller impact.

A state might be interested in smaller differential impacts. For example, a state might want to detect an impact equal to half the overall impact. Continuing our example, this might be a pooled impact of 2 weeks and impacts in the two subgroups of 2.5 and 1.5 weeks. That would require a sample 16 times as large as the sample to detect the impact of 2 weeks in the pooled sample.

As noted, those are estimated sample sizes for detecting impacts on employment. Impacts on UI duration can typically be detected with smaller samples, maybe fewer than 10,000 or 15,000. But most evaluations will want to be able to detect impacts, not only on UI duration, but also on employment.

It follows that most states can do a whole-program experimental impact evaluation without pooling estimates—perhaps after randomizing for a few years—but few states can do a subgroup experimental impact evaluation without pooling—even after randomizing for a few years. Instead, pooling estimates across state-specific evaluations will be useful and perhaps required to address RQ2.[28]

As described below, these larger samples sizes and the need for pooling that they induce suggest two options to support subgroup impact analyses.[29]

Option 2.3-1 Coordinating Subgroup Analysis–Reporting Guidelines (RQ2/Subgroups)

What: Pooling is stronger when states define and collect data for subgroups in a consistent manner. This option would generate reporting guidelines to ensure states report common subgroup identifiers and do their analyses in (sufficiently) similar ways. With the reporting guidance would come evaluation TA materials and state-specific evaluation TA for using the materials.

This is the Subgroups equivalent of **Option 2.2-8/Coordinating Evaluations in Support of Synthesis**. For whole-program evaluations, such coordination was nice, but not necessary. Because subgroup analyses will almost always involve pooling, such coordination is crucial for subgroup evaluations.

Who, When, and How Much: This would be a DOL-directed activity. Ideally, such an effort would be completed before states begin to generate impact estimates.

This is a lower-cost option. The costs include developing guidance (i.e., standards) and to creating materials to help states conduct their evaluations consistent with those standards. This option would involve developing a draft set of standards, circulating them for comment, and then revising into a detailed document and a quick reference for evaluators.

Why: Analyses pool estimates that, taken alone, are not precisely estimated, in order to yield pooled estimates that *are* precisely estimated. Doing so requires estimates that are easy to generate and report. However, those estimates are imprecisely estimated at the state level (and thus yield essentially no insights for this state). Thus, each state on its own has little incentive to produce them. Reporting guidelines are intended to induce states to do their analyses (sufficiently) similarly and to report the results needed for pooling.

Option 2.3-2 Synthesis of Subgroup Impact Estimates (RQ2/Subgroups)

What: This option would combine the state-specific subgroup estimates—which alone have insufficient samples—to yield insights into differential impact across subgroups. Given the fact that impacts likely vary among states, the appropriate approach is to use a random-effects meta-analysis. The goal here is not to establish that there is any differential impact. Instead, the goal is to estimate the size of any differential impact. For that purpose, simple counting strategies will not be appropriate.

---

[28] Of course, if a state is conducting subgroup analyses as an exploratory add-on to a whole-program evaluation, not as a means of conclusively demonstrating impacts for statutory purposes, then it might find suggestive evidence that falls somewhat short of statistical significance thresholds practically useful for making targeting decisions, even if the evidence is not definitive.

[29] There is likely to be less interest in **Option 2.2-7/ Sophisticated Synthetic Estimates of the Impact of a Specific State's RESEA Program**.

This is the subgroup equivalent of **Option 2.2-6/Synthesis of the Impact of a Generic RESEA Program**. For whole-program evaluations, such synthesis is nice but not necessary, because most states can produce sufficiently powered estimates on their own. Because subgroup analyses will almost always involve pooling, such cross-state synthesis is crucial.

Who, When, and How Much: This would be a DOL-directed activity. It would not be appropriate until there are estimates from perhaps half a dozen evaluations and those evaluations have been reviewed by Option 2.2-6. This is a lower-cost option. These costs cover developing the basic model for pooling results, creating a database with the results to be pooled, writing up the results, and briefing DOL on those results

Why: This option would generate the pooled estimates. Single state samples are too small. Thus, without this option, there are no insights toward **RQ2/Subgroups**.

Option 2.3-3 Disparate Impact of Selection Strategies (RQ2/Subgroups)

What: States use various methods to choose whom to select for RESEA. In particular, RESEA survey evidence suggests that the overwhelming share of states continue to select those claimants with high probability of exhausting benefits. It is possible that those methods have disparate impact; that is, some racial or ethnic groups are more likely to be selected. This study would use REA or RESEA to explore disparate impact on selection for REA/RESEA.

Who, When, and How Much: This could be either a state-directed or a DOL-directed activity. It could be done immediately with REA data from earlier evaluations. It could be done soon on current RESEA data. This is a lower-cost option.

Why: This option would provide evidence on the extent to which, in practice, selection for RESEA is biased with respect to racial and ethnic groups. This option would also allow policymakers to explore how differential selection would vary with different selection rules.

3. Options for RQ3/Components and RQ4/What Works Best for Whom

Whereas Chapter 2 focused on options for evaluations of whole programs, this chapter examines options related to evaluating service components of RESEA programs and what works best for whom. Specifically, the chapter considers options to understand how adding, dropping, or changing a component would change program outcomes. Such evaluations would address or support addressing:

- **RQ3/ Components: How does the impact of an RESEA program vary with service or component details?**
- **RQ4/What Works Best for Whom: How does the impact of changing a component vary with the characteristics of the claimant at initial claim?**

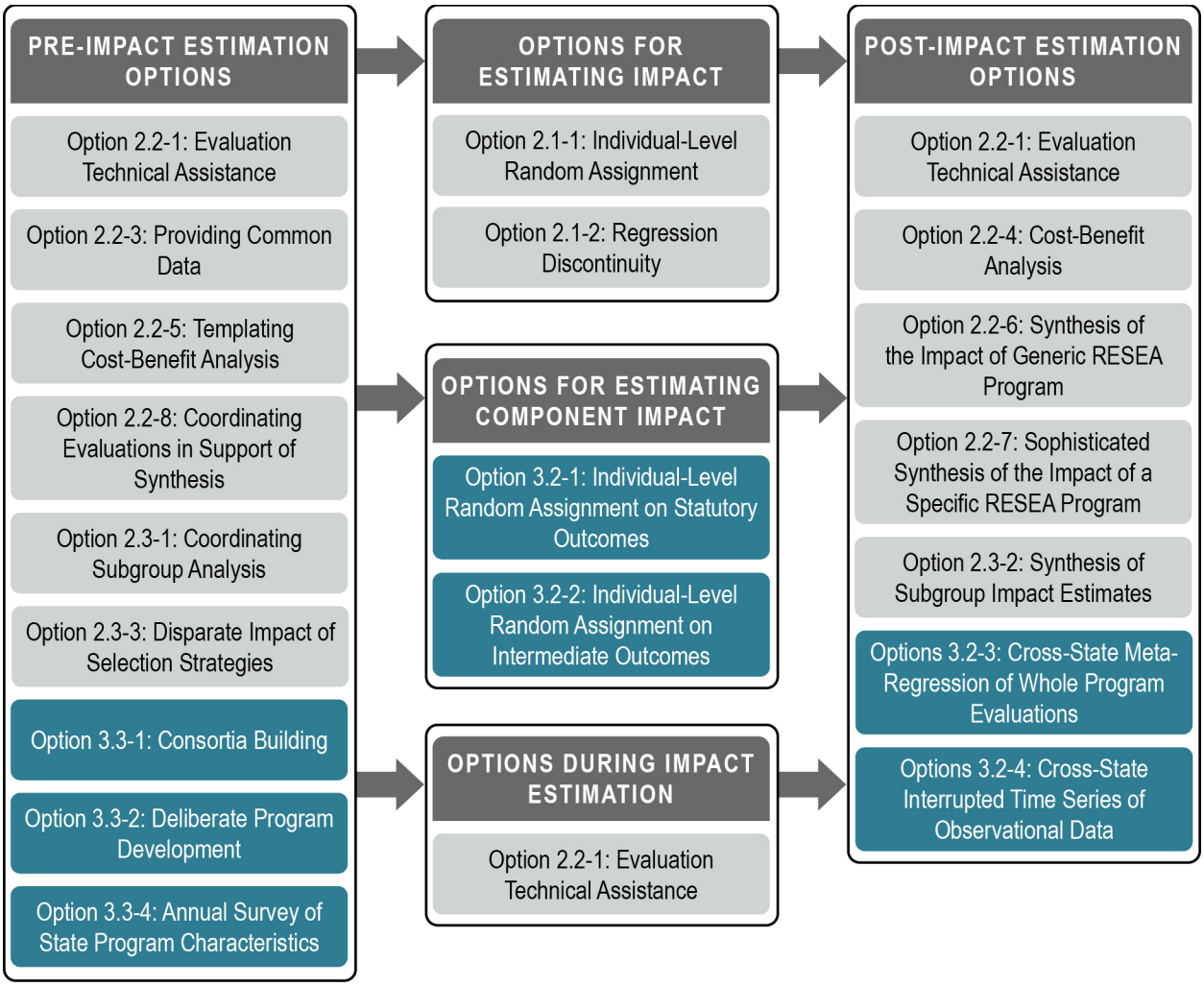
Evaluations focusing on these research questions are important for understanding how to *improve* program design overall and for specific subsets of claimants. Evidence of effectiveness for a set of components could, in theory, also cumulate to meet the Statutory stipulations for FY 2023 and beyond that require that certain percentages of RESEA program funding go toward interventions that are demonstrated to be effective.

To address RQ3 requires specifying the component in detail during the design phase of the evaluation, specifically which service component is to be dropped, (precisely) what service component is to be added, or which service component is to be varied (and how). Section 3.1 discusses specific components of RESEA whose impacts might be evaluated. Section 3.2 presents options for estimating impact of the components of interest (**RQ3/Components**). Section 3.3 presents options for activities that would support an impact analysis to address RQ3, but do not directly estimate impact. There is strong overlap between these support activities in Section 3.3 and those in Section 2.2 to support **RQ1/Whole Programs** and those in Section 2.3 to address **RQ2/Subgroups**. Nevertheless, Section 3.3 adds new options, related to choosing which component(s) to evaluate and building consortia to do so.

Evidence on program effectiveness for claimants with different characteristics will provide information states can use to develop and offer different program models for different groups of UI claimants. Each group of claimants then receives program that is most effective (or perhaps cost-effective) for them. RQ4 considers options for addressing **RQ4/What Works Best for Whom**.

Exhibit 3-1 shows graphically how the RQ1, RQ2, RQ3, and RQ4 options relate. To the left are options to be implemented *prior* to impact analysis. In the middle are options to conduct impact analysis or that would be implemented in parallel with the impact analysis. To the right are options to be implemented *after* impact analysis. Options that were discussed in Chapter 2 are greyed out; those that are new in this chapter are not.

Exhibit 3-1. Inter-Relation of Options for RQ1/Whole Programs, RQ2/Subgroups, RQ3/Components, and RQ4/What Works Best for Whom



Source: Abt Associates.

3.1. Components That Might Be Evaluated

This section considers RESEA service components that might be evaluated. These component tests are presented only as examples, not as the definitive list.

Some potential tests of service components refer to *what* is delivered (e.g., reemployment assistance vs. no reemployment assistance); others refer to *how* the service is delivered (e.g., remotely vs. in person). Some are tests of components typically included in programs—such as the *current core set* of reemployment services as part of the initial RESEA meeting. Others are tests of possible *additions to a program's typical features* (e.g., regular case manager contacts, subsequent RESEA meetings) or *changes to the design of existing features* (e.g., holding meetings remotely rather than in person; changing who staffs RESEA meetings).

The section discusses five broad categories of components of RESEA programs: (1) activities before the RESEA meeting; (2) mode of the meeting; (3) meeting design, including content and number of meetings; (4) services outside the meeting; and (5) responses to non-compliance. Later sections describe

options for how to evaluate components and return to some of the component tests listed below to illustrate studies that could be performed.

Exhibit 3-2 provides examples in each category that might be tested. Each specific component test described refers to design decisions that program administrators face in implementing their program. For example, initial meetings must be scheduled, but how the scheduling happens (e.g., by the program or by the claimant) and at what point in the claim the meetings occur (e.g., within the first few weeks or in later weeks) is to be determined.

The reasons for inclusion of the specific components outlined in Exhibit 3-2 vary. In some cases, the component test is an area where state programs currently differ. This suggests that additional evidence might affect and/or assist state policy choices. In other cases, the component is hypothesized to improve outcomes but formal evaluation is needed to confirm—or disprove—that hypothesis. Finally, in some cases, the component is an area where DOL or multiple states have expressed a policy interest.

Exhibit 3-2. Examples of RESEA Components That Can Be Evaluated

Component category	Component test	Why the component change might affect claimants' outcomes
Activities before the RESEA meeting	(a) Timing: Week for which the meeting is scheduled (earlier vs. later in the claim)	• Because some claimants will quickly find a job on their own, holding the initial meeting later may mean that services go to claimants on whom the services have the greatest impact. • Holding the initial meeting later reduces the potential impact that RESEA can have on UI duration and can delay receipt of services that might help some claimants become reemployed.
	(b) Scheduling: Meeting self-scheduling vs. scheduling by staff	• Self-scheduling might increase attendance rates, and in turn receipt of services. • Self-scheduling might delay the timing of the first meeting. • Some claimants might not self-schedule the meeting at all.
	(c) Reminder frequency: Adding meeting reminders or increasing the frequency of meeting reminders	• Adding reminders might increase meeting attendance rates, and in turn receipt of services.
	(d) Reminder mode: Contact by phone, email, etc.	• Claimants might respond more to contacts perceived as personalized (e.g., a phone call) than to automated ones.
Mode of the RESEA meeting	(e) Venue: Remote vs. in person	• Offering remote meetings might increase attendance rates. • Services provided in the meeting might be more effective if delivered in person.
	(f) Location: At AJC vs. at other physical location	• Conducting the meetings at an AJC puts claimants in closer proximity to services available through WIOA and other partner programs and may lead to more use of reemployment services. • Offering meetings at a physical location other than an AJC might increase attendance by offering locations that are more convenient.

Component category	Component test	Why the component change might affect claimants' outcomes
RESEA meeting design	(g) Content: Reemployment services offered in the RESEA meeting vs. excluding those services	• RESEA offers reemployment services in the RESEA meeting with the expectation that they help with claimants' job search, and in turn might improve claimants' outcomes.
	(h) Number of RESEA meetings	• Subsequent RESEA meetings might help claimants connect to and benefit from reemployment services received.
	(i) Staffing the RESEA meeting (WIOA staff vs. UI staff vs. both)	• Staffing might influence the extent to which claimants connect to individualized career services and potentially better jobs more quickly by making different referrals, consistent with differences among staff in their familiarity with particular employment programs. • Having WIOA and UI staff, respectively, split the reemployment services and eligibility assessment responsibilities between them, rather than having one staff member handle both, might increase claimants' trust in the staff member who provides reemployment services and, in turn, their willingness to engage with those services.
Reemployment services outside the RESEA meeting	(j) Case management contact requirements (none vs. weekly vs. monthly)	• Assigning a staff member to each RESEA participant and requiring the staff member to make contact with each claimant regularly might more effectively connect claimants to services or job openings that get them back to work more quickly.
	(k) Requirements for claimants to participate in RESEA and some set of AJC reemployment services	• If claimants, on their own, do not fully understand and take advantage of AJC services that are available to them, requiring that they co-enroll and engage in some number of workshops or other services outside the RESEA meeting might help them return to work more quickly and to a better job.
Responses to non-compliance	(l) Suspending claimants' benefits for failure to attend the RESEA meeting until they attend	• Suspending benefits immediately after a claimant fails to report provides greater incentive to attend the RESEA meeting(s) at which they can receive services that might help them return to work more quickly.

If primary funding for component evaluations will come from states, then the most important criterion for selecting which components to evaluate is that several states want to and agree to participate in and fund a coordinated evaluation of that component. This is because successful component evaluations need large samples, often more than a tenth of the national RESEA caseload (see Appendix Section B.3). Achieving samples of that size will usually require the participation of multiple states, including several of the larger ones.

With that crucial caveat, four component evaluations are promising candidates for evaluation, based on a combination of expressed interest by DOL, extent of recent RESEA program implementation changes (Trutko et al., 2022), and REA impact evidence (Klerman et al., 2019):

1. ***Remote Services.*** States have responded to the COVID-19 pandemic by shifting to a range of remote services—switching from in-person one-on-one and group sessions to some combination of phone calls, live video calls, and recorded videos. These alternative modes of service delivery

have advantages and disadvantages.[30] As the pandemic recedes, in-person meetings will again be feasible and states' RESEA programs will face a decision regarding how widely and often they should offer remote meetings.

This option would estimate the impact of various remote service models relative to a conventional, in-person service model. Specifically, participating states would jointly develop one or perhaps two remote service models to test. Those models seem likely to involve some combination of pre-recorded videos (perhaps with built-in quizzes) and video-conference one-on-one meetings. Given variation in state approaches to remote services it might be worthwhile to consider both a model with a video-conference meeting of similar length to an in-person meeting and a second model with a much shorter video-conference meeting. Because remote services make access easier, in addition to Q2 employment and UI weeks, outcomes of interest may include meeting attendance and subsequent use of AJC services.

2. ***Intensive Reemployment Services.*** Assistance with reemployment is a key component of the RESEA program. If some assistance is effective, perhaps more-intensive assistance—longer meetings with the caseworker, inclusion of a wider range of reemployment services—would be better.

   This option would estimate the impact of a more intensive service model relative to a conventional, less-intensive service model. Specifically, participating states would jointly develop one or perhaps more intensive service models to test. Those models seem likely to involve some combination of more and longer meetings, a broader range of services provided in those meetings, and more concerted efforts to induce claimants to use appropriate AJC services outside of RESEA meetings.

3. ***Responses to Non-Attendance at the RESEA Meeting.*** The REA Impact Study (Klerman et al., 2019) findings are consistent with a major role for non-attendance policy in driving the impact of reemployment programs—including RESEA—on UI weeks and variation in impacts on UI weeks across states.

   This option would estimate the impact of "suspend until attend" relative to an approach with consequences that are weaker, less certain, or less immediate. Specifically, participating states would jointly develop one or perhaps more intensive responses to non-attendance. Those models seem likely to involve some combination of "sure, swift, and substantial" response to non-attendance; that is, benefits would be suspended immediately and until attendance, with supervisors' review to ensure compliance by caseworkers. This model might include total loss of those weeks of benefits (vs. the current policy, which makes them available later in the spell of unemployment). This model might also include additional—perhaps monthly—meetings and more intensive verification of ongoing eligibility (see the next bullet).

4. ***Eligibility Assessment.*** Eligibility assessment—that is, review of compliance with ongoing eligibility requirements (beyond the requirement to attend the meeting)—appears in the RESEA program's name, but implementation research consistently finds that staff are reluctant to enforce those ongoing eligibility requirements rigorously (Minzer et al., 2017; Trutko et al., 2022).

   This option would estimate the impact of a rigorous assessment of ongoing eligibility—in particular, sufficiently intensive work search—relative to a less punitive and more cooperative

[30] For example, conducting RESEA meetings remotely may have the benefit of making it easier for claimants to attend. However, conducting meetings in person may have the benefit of allowing staff to develop better rapport with and understanding of the RESEA participant. In-person meetings also put participants in physical proximity to AJC partner services and may make participants more likely to engage those services.

approach. Specifically, participating states would jointly develop one or perhaps two models for intensive eligibility assessment. Those models would start with requiring filling out an on-line work search log, for all weeks since the initial claim—prior to the RESEA meeting. Then at the meeting, the RESEA worker would review the logs and suspend UI benefits for every week out of compliance—including that the log was completed on time (in the week following when the work search occurred, not immediately before the RESEA meeting). Furthermore, non-compliant individuals would be scheduled for more frequent (perhaps weekly) RESEA meetings, until they were consistently compliant for some period (perhaps a month). Finally, supervisors would review logs and case files to verify that RESEA workers were implementing the policy. The model might also include verification of a sample of claims on the log; that is contacting employers.

Evaluating a component will require carefully specifying the component to be evaluated beyond the general parameters discussed above. Ideally, the evaluated component would reflect the field's assessment of the most promising form of the component. Then the preliminary design for that component could be piloted and refined—before starting a large and more expensive impact evaluation.

Option 3.3-2/Deliberate Program Development for Pilots and Demonstrations, described below, includes a process to develop a candidate component. For example, if the component was intensive services, the steps of the process might address the following questions: *When will intensive services be provided? To whom? What will be the content and intensity (e.g., minutes per claimant) of those services? Then that candidate component would be refined through piloting and formative evaluation.*

3.2. Options for Estimating Impact of Components

This section describes two experimental options and two non-experimental options for addressing **RQ3/Components**: *How does the impact of a RESEA program vary with the components included and how they are provided?* Specifically, it discusses individual-level random assignment (**Options 3.2-1** and **3.2-2**) and cross-state analyses of data from different sources (**Options 3.2-3** and **3.2-4**). The section briefly discusses other potential evaluation options not considered in detail here.

All of the evaluation designs in this section would be appropriate nearly regardless of specific component being evaluated. **Option 3.3-1/Consortia Building** in the Section 3.3 describes a process for selecting the component(s).

Option 3.2-1 Individual-Level Random Assignment to Estimate Impact of a Component on RESEA Statutory Outcomes (RQ3/Components)

What: With this design, some of the claimants *selected* for RESEA are randomly assigned to a program with the component being evaluated; the others *selected* for RESEA are assigned to an alternative program without the component or that varies the component in some way.[31] This design produces estimates of the impact of the component that distinguishes one program version from the other. Both UI duration and employment outcomes (RESEA's statutory objectives) would be measured using administrative data.

The evaluation could be used to demonstrate effectiveness for statutory purposes (e.g., estimating the impact of the reemployment services offered in the meeting; see Example g in Exhibit 3-2) or testing the impact of adding an ongoing case management component (Example j). But a larger goal is to provide

[31] In the simplest version of this design, there is no control group; that is, random assignment only determines the program; selection for RESEA itself is not random.

insight toward program refinement and better program outcomes by showing how much ultimate outcomes of interest change with the change in the component.

In practice, this design is usually implemented using three-way random assignment, with claimants assigned to one of three groups receiving (1) the current RESEA program; (2) the alternative RESEA program; or (3) no RESEA program, meaning claimants would search for work on their own.[32] Because each of the program approaches (groups 1 and 2) can be compared to the no-service group (group 3), this design also provides a whole-program evaluation of each program approach. It also estimates the effectiveness of each program approach compared to the other.

Who, When, and How Much: To reliably detect impacts on both UI duration and employment outcomes, samples of roughly 150,000 UI claimants *selected* for RESEA would be required, with roughly half assigned to the current program and the other half to the alternative program. If the assignment ratio differs sharply from 1:1, an appropriate sample would be larger.

A sample size of 150,000 claimants is more than a tenth of the annual national RESEA caseload. Few states can generate samples of this size in a year or even in several years. Thus, all but the very largest states would likely need to pool their estimates in a multi-state effort to conduct this type of evaluation.

The multi-state efforts or consortia likely required to implement this option can be difficult to undertake and launch. They require agreement across states about the design of components being tested and consistent implementation, presenting a number of challenges (see **Options 3.3-1** and **3.3-2**, which address those). After agreement is reached, timelines are similar to those for **Option 2.1-1/Individual-Level Random Assignment**. Depending on how many years of random assignment are needed, it will be three to six years from the start of random assignment to a results report. Given the required sample sizes, time to the report is likely to be toward the upper end of that range.

This is a higher-cost option. The costs include work setting up and monitoring random assignment, processing the data systems, doing the analysis, and writing up and briefing the results. Note also that this cost does not include any incremental cost of delivering the component being tested.[33]

Why: Individual-level random assignment provides the strongest evidence of component effectiveness (see **Option 2.1-1** for a discussion of why). Furthermore, for a given sample size, individual-level random assignment needs the smallest possible sample relative to non-random assignment designs. Because sufficient precision will often be a key challenge for state RESEA evaluations, random assignment designs offer a major advantage.

Furthermore, this option is relatively certain to yield results that will meet CLEAR effectiveness standards for Moderate or High evidence. Other non-experimental designs (discussed in Section 2.1) are more challenging to complete successfully.

Option 3.2-2 Individual-Level Random Assignment to Estimate Impact of a Component on Intermediate Outcomes Only (RQ3/Components)

What: The ultimate goal of an RESEA RQ3/Components evaluation is to show that a component improves UI weeks and Q2 employment. Toward that goal, an impact evaluation of a component might

---

[32] The REA Impact Study's analysis of the impact of multiple meetings is an example of such a study (Klerman et al., 2019).

[33] With any given RESEA grant size, testing a more expensive model would reduce the number of claimants who can be served by RESEA. Apart from the programmatic ramifications, a smaller number of claimants who can receive the program model means a smaller sample for the evaluation.

start by showing that it improves some intermediate outcome(s)—for example, meeting attendance. Showing an improvement in meeting attendance is promising but not conclusive for impacts on UI weeks and Q2 earnings. Thus, a follow-on study of impact on UI weeks and Q2 employment should follow.[34]. This option would defer to a follow-on evaluation showing that the component improves Q2 employment.

This option considers evaluations are designed to estimate intermediate outcomes and not scaled to detect net impacts on Q2 employment (i.e., the employment difference between the intervention and control groups). **Option 3.2-1** discussed above requires large appropriate sample sizes because it would scale the evaluation to detect impacts on Q2 employment. For a component test, impacts are likely to be modest (because it compares to two service components rather than to a no-services or limited-services group) and therefore hard to detect. In contrast, because this **Option 3.2-2** would estimate impact on intermediate outcomes (e.g., attendance at the RESEA meeting), and thus appropriate sample sizes are much smaller.[35] Smaller samples increase feasibility, because they decrease the need for states to pool estimates.

Who, When, and How Much: Studies that are *not* designed to reliably detect differential impact on Q2 employment can use smaller samples. That implies that most states could do their own **RQ3/Components** evaluation if the impact of interest is UI weeks, and that nearly all states could do a **RQ3/Components** evaluation if the impact of interest is an intermediate outcome (e.g., meeting attendance).

Because these likely evaluations could be conducted by single states, the time and effort to develop multi-state consortia would not be required. A single-state evaluation could be started as soon as the RESEA component of interest was identified, defined well enough to be implemented, and then launched. A report can be available 18 months to two years after randomization begins. This is a lower- to medium-cost option, depending on the type of component that is being evaluated. Similar to **Option 3.2-1**, the cost of **Option 3.2-2** includes analysis and reporting of the results. Note also that this option does not include any incremental cost to states delivering the component being tested, which might be an enhanced approach with higher costs than the alternative.[36]

Why: Using individual-level random assignment to estimate impact provides insights into the effectiveness of a component—and with *much* smaller appropriate samples than to detect impacts on employment. However, showing impacts on intermediate outcomes does *not* necessarily imply impacts in the desired direction on ultimate outcomes (i.e., employment, UI duration).[37] Furthermore, even if the

---

[34] See Epstein and Klerman (2012) for a discussion of the role of intermediate outcomes.

[35] For example, with a total sample of 30,000, the REA Impact Study in one state detected differential impact of multiple meetings on UC weeks, but not on Q2 employment; in another state, it did not detect even impacts on UC weeks. In contrast, Darling et al. (2017) detected differential impact of messaging on meeting attendance with a sample of roughly 1,000.

[36] With any given RESEA grant size, testing a more expensive model would reduce the number of claimants who can be served by RESEA. Apart from the programmatic ramifications, this reduces the sample size that is available to the study.

[37] An example helps to make the issues vivid. Inasmuch as RESEA achieves its impacts by denying UC benefits to claimants who do not attend the meeting, better noticing could lead to increased meeting attendance. That increased attendance could lead to higher attendance and less "suspend until attend"—which would lead to higher UC weeks. If a loss of UC benefits causes claimants to search more intensively and accept more job offers, Q2 employment and earnings might increase. Outcomes would improve only if the impact of the additional assistance delivered at the RESEA meeting outweighs these impacts of less "suspend until attend." See Klerman et al. (2019) for more discussion of these issues.

impacts on ultimate outcomes are in the desired direction, they might be trivially small.[38] This option may be part of **Deliberate Program Development (Option 3.3-2)** efforts. The fact that outcomes are measured quickly, and analyses can be conducted with smaller sample sizes, means that such studies can be used as part of rapid cycle evaluations wherein findings are used to inform repeated program refinements on relatively short timelines.

But findings should only be viewed as *suggesting* conjectures to be evaluated in a follow-on evaluation with samples large enough to detect impacts on Q2 employment (i.e., **Option 3.2-1/Individual-Level Random Assessment to Estimate Impact of a Component on RESEA Statutory Outcomes**).

Option 3.2-3 Cross-State Meta-Regression of Whole Program Evaluations (RQ3/Components)

What: This design would build on state-specific whole program evaluations (usually **Option 2.1-1/Individual-Level Random Assignment**) to explore which components are associated with larger impacts. Although the underlying studies are experimental, the variation in program features is observational. Thus, this should be viewed as a non-experimental approach.[39] There are no known applications of this design to reemployment programs. This contract tried to apply these ideas to reemployment programs but was not successful. This lack of success in applying these ideas appears to have been because there were not sufficient studies, each with sufficient precision.

For this meta-analytic approach to succeed, key components of programs that distinguish one program from another must be well documented. Components are such as those listed in Exhibit 3-2. For example, how many meetings were required, what strategies were used to promote meeting attendance, how meetings were staffed, what services were provided in meetings, what services were provided or required outside of meetings, and what were the consequences of failing to attend meetings or comply with work search requirements.

In practice, prior research has often not carefully documented such programmatic details, making it difficult to clearly understand how programs with larger impacts differ from programs with smaller impacts.[40] **Option 3.3-4/Annual Survey of State Program Characteristics** discusses use of future implementation studies to help address that problem and make meta-regressions a more feasible option for identifying component impacts.

Who, When, and How Much: This would be a DOL-sponsored study. However, it would not be feasible until there are several (e.g., a dozen) completed state-specific whole-program evaluations. In practice, this option is probably feasible only if most states move toward continuous whole-program experimental impact evaluations (i.e., wide application of **Option 2.1-1/Individual-Level Random Assignment**). This is a medium-cost option.

---

38 For example, the experimental literature on job training consistently shows that the offer of the program increases training received. Impacts on earnings are much rarer (e.g., the WIA Impact Evaluation; Fortson et al., 2017). One interpretation of this pattern of results is the magnitude of the impact on training receipt is detectable but too small to lead to a detectable impact on earnings. (See Weiss et al. [2015] for a similar argument in the community college literature.)

39 Formally, this is a random-effects meta-regression.

40 For example, the *Employment Strategies for Low-Income Adults Evidence Review* (ESER) included reemployment programs as part of its application of these ideas to all employment strategies (Vollmer et al., 2017), but the results were inconclusive. In part this might be because the "strategies" examined were very broadly defined and with extremely wide variation in the programs grouped together. Studies applying this design in other domains include Greenberg et al. (2003) on job training programs and Bloom et al. (2003) on welfare-to-work programs.

Why: Relative to conducing multiple large sample evaluations to directly estimate the impacts of various components, cross-state meta-regression of whole program impact results is a less costly and less disruptive way to generate insights about effective components. Any estimates emerging using this design should be viewed as exploratory.[41] For instance, the option can be used to identify promising components on which to focus additional evaluation (**Option 3.2.1/Individual-Level Random Assignment/Statutory Outcomes**).

Option 3.2-4 Cross-State Interrupted Time Series of Observational Data (RQ3/Components)

What: Cross-state ITS design would explore which program characteristics of state RESEA programs are associated with better outcomes (see discussion in Chapter 2). A non-experimental approach, this study design uses statistical methods to extract state-specific impacts from information on outcomes. It is thus much weaker than **Option 3.2-3/Cross-State Meta-Regression**, which analyzes experimental estimates of impact from state-specific whole-program evaluations.

In practice, **Option 3.2-4** yields plausible estimates of impact only when there is variation in program components over time. This is because the design explores how outcomes vary as a single state changes its program model. By exploring how outcomes vary with changes in a state's program model, the design implicitly uses each state as its own control. In addition, this design uses other states to control for national time-varying factors and to model changes in outcomes with changes in local economic conditions.

This ITS option is sometimes called **difference-in-differences**. Toohey (2017) uses this design to estimate the impact of state UI work search requirements. Klerman and Danielson (2011) apply these ideas to the Supplemental Nutrition Assistance Program caseload. See Mayer (1995) and Bloom (2003) for more discussion of this design.

Who, When, and How Much: Such cross-state ITS would likely be a DOL-sponsored study. This design requires aggregate state-level data for many years (enough years to include multiple changes in program design), from multiple states, on both outcomes and program details. Some data on outcomes are currently available, but the quality of those data is unclear. The current project's state survey collected information on program details; such a survey would need to be repeated annually for many years to implement this design. Thus, this design is not feasible in the near future. This is a medium-cost option.

Why: Such cross-state ITS is a relatively low-cost and undisruptive way to generate insights to be formally tested using the canonical design (**Option 3.2-1/Individual-Level Random Assignment to Estimate Impact of a Component on RESEA Statutory Outcomes**). Its inferences are considerably weaker than those of **Option 3.2-3/Cross-State Meta-Regression**. Similar to **Option 3.2.3**, any estimates emerging using this design should be viewed as exploratory. Such studies could find it challenging to satisfy CLEAR's standards for a Moderate rating. In particular, it often might not be possible to "demonstrate equivalent *trends* between the intervention and comparison groups being analyzed" (CLEAR, 2015, p. 4; emphasis in the original). This "equivalent trends" condition is particularly challenging because it can only be established well into the analysis.

[41] As an approach to cross-study evidence synthesis, as opposed to a primary study of an intervention, it likely would not be reviewed by CLEAR.

Other Potential Components Designs, Considered

This section has considered four options for estimating component impact, based on a scan of a larger set of possible options. This sub-section briefly notes several other designs. All have substantial feasibility challenges due to sample size requirements.

First, similar to whole-program **group random assignment**, components group random assignment—e.g., some randomly chosen *AJCs* administer the component, others do not—requires samples several times larger than does individual-level random assignment. Because sample size issues are already a challenge, an evaluation design that required much larger samples is likely infeasible for all but the largest states.[42]

A second option (sometimes called **regression adjustment** or **propensity score matching**) compares outcomes of claimants (not randomly) selected to RESEA versus outcomes for claimants not selected and controlling for observable differences This design is extremely unlikely to estimate true impact because the reasons that claimants do or do not receive services[43] are likely to be strongly related to their employment outcomes. Such an evaluation is therefore unlikely to satisfy CLEAR's standards for a Moderate rating of as study's evidence credibility.

Third, if a state rolled out RESEA to offices in some purposive way, **interrupted time series** might be possible. Two studies of reemployment interventions have used ITS designs; however, they achieved only a low rating in CLEAR. Impact studies of RESEA interventions need to be able to earn at least a Moderate rating from CLEAR in order to be accepted as evidence demonstrating the intervention's effectiveness. With the help of an evaluator with relevant expertise, it might be possible for an ITS study to achieve at least a Moderate CLEAR rating. However, an ITS study design requires very large sample sizes and involves complicated operational logistics to implement the intervention on the pre-specified, staggered timeline.

Fourth, precision and logistical issues make the Components equivalent of **regression discontinuity design** (**Option 2.1-2**) a less desirable option in most cases.

3.3. Options to Support Estimating Impact of Components

Most of the options discussed in Section 2.2 to *support* **RQ1/Whole Programs** evaluations carry over nearly directly to *supporting* **RQ3/Components** evaluations, with at most minor adjustments. Section 2.2 options requiring minor adjustments (or less) are:

- **Option 2.2-1/Evaluation Technical Assistance**—where the TA needs to consider components, but the changes are minor and easy to implement.
- **Option 2.2-2/Providing Common Analytic Tools**—where the tools need to consider components, but the changes are minor and easy to implement.

---

[42] This is particularly unfortunate in the Components context. A component impact evaluation often requires a state to run two program models in parallel (e.g., the current model and a model with intensive services). Effective evaluation requires that the two models stay separate. It is problematic if the existing model borrows "good ideas" from the alternative model. AJC-level randomization is one way to prevent such contamination. However, as noted in the body, AJC-level randomization does not appear to be feasible due to sample size issues.

[43] The most likely reasons that one claimant receives services and another does not are differences in factors such as how much a claimant needs the services in order to find a job (e.g., if you already know you can find a job, you are less likely to use services) or the claimant's motivation or capacity to find a job.

- **Option 2.2-3/Providing Common Data**—no change.
- **Option 2.2-4/Cost-Benefit Analysis**—where the analysis needs to be conducted for components, but the changes are minor and easy to implement.
- **Option 2.2-5/Develop a Template for Cost-Benefit Analysis**—where the template needs to be expanded to include components, but the changes are minor and easy to implement.

Similarly, as for **RQ2/Subgroup** impact analyses, almost all of the **RQ3/Components** impact analyses involve pooling of estimates across states. As a result, the pooling options discussed in Section 2.3 carry over nearly directly for **RQ3/Components**, as well. The only change is that the options need to consider the content of components, rather than whole programs. For example, under **Option 2.3-1/Coordinating Subgroup Analysis–Reporting Guidance,** the reporting guidelines need to consider components, but the changes are minor and easy to implement; and under **Option 2.3-2/Synthesis of Subgroup Impact Estimates,** the synthesis needs to consider components, but the changes are minor and easy to implement.

As described below, issues related to coordination suggest two options, and issues related to data suggest another two options, all in support of component impact analyses.

Option 3.3-1 Consortia Building (RQ3/Components)

What: Given that such evaluations will typically require participation of multiple states in a pooled evaluation, the main challenge in implementing **Option 3.2-1/Individual-Level Random Assignment to Estimate Impact of a Component on RESEA Statutory Outcomes** is getting enough states to agree to evaluate the same component (e.g., "intensive services" or "suspend until attend"). In some cases, states might form such consortia on their own, however the coordination and implementation challenges are substantial. This option would support a process to build consortia of states committed to evaluating the same component.

One vision is for a cooperative process that induces states to volunteer to participate. Such a process might unfold as follows:

1. A coordinating entity solicits expressions of interest in various **RQ3/Components** topics. Likely this would involve constructing a list and circulating it to all states or holding phone discussions with states to assess their interest in various topics. The list would include a space for states to suggest topics not on the initial list.

2. The coordinating entity would review the list, looking for topics that are most likely to get enough states to participate. That list would be recirculated to the states, asking for tentative commitments. At this stage, DOL might indicate its preference for some topics and its willingness to provide funds toward evaluating them.

3. From those states providing tentative commitments, the coordinating entity would work to get formal agreements to join consortia. Again, at this stage, DOL might indicate its preference for some topics and its willingness to provide funds toward evaluating them.

Who, When, and How Much: If states can form sufficiently large consortia on their own, this option is unnecessary. Because doing so is likely to be challenging, as part of this option, DOL would fund efforts to spur the formation of consortia. This option could start immediately and would need to start well before states begin **Option 3.2-1/Individual-Level Random Assignment to Estimate Impact of a Component on RESEA Statutory Outcomes (RQ3/Components).** This is a lower- to medium-cost option.

Why: Such consortia may be needed to execute **Option 3.2-1/Individual-Level Random Assignment to Estimate Impact of a Component on RESEA Statutory Outcomes (RQ3/Components).** This option would increase the chances of consortia forming.

Option 3.3-2 Deliberate Program Development for Pilots and Demonstrations (RQ3/Components)

What: A pilot or demonstration project would involve testing a *well-defined intervention* of high interest. Thus, for example, even if multiple states are interested in "intensive services" for **Option 3.2-1/Individual-Level Random Assignment to Estimate Impact of a Component on RESEA Statutory Outcomes,** the states would need to agree on at least the basic details of the intervention.

Say states were interested in testing "ongoing case management" (*Item j* in Exhibit 3-2): Because all claimants already have potential access to some kind of ongoing case management through their AJC, an RESEA case management intervention for testing would need to be a new approach, specific to RESEA participants, that involves a particular frequency, type, and content of interactions by an assigned case manager—with RESEA participants only. Those details of interaction frequency/type/content would need to be collaboratively developed and agreed upon by states. This option would help the states in the consortium to reach consensus on the details of the test.

Specific activities might include:

- Implementation studies to explore various ways in which the component of interest is currently implemented (see **Option 3.3-3/Implementation Studies to Accompany Impact Evaluations** for a discussion).
- Literature reviews and expert panels to explore theoretical and programmatic insights from within and outside the reemployment services field.
- Developing materials to train line staff on the new approach and for line staff to use in implementing the new approach.
- Piloting the new approach, formative evaluation to assess implementation, and potentially an experimental evaluation to examine whether the approach improves intermediate outcomes, (e.g., meeting attendance rate, frequency of contacts; **Option 3.2-2Individual-Level Random Assignment to Estimate Impact of a Component on Intermediate Outcomes Only**) to refine the design of the intervention.

Who, When, and How Much: A group of states might fund and jointly direct such an activity or DOL could do so. This option would occur between **Option 3.3-1/Consortia Building** and **Option 3.2-1/ Individual-Level Random Assignment to Estimate Impact of a Component on RESEA Statutory Outcomes**. Depending on the number of rounds of piloting and refinement, such efforts could take less than a year or several years. This is a medium-cost option.

Why: Because evaluation of a component requires operating two separate versions of a program, impact studies of a component can be costly and challenging to conduct. Developing a demonstration is likely to refine the program model. A refined program model is likely to have larger impacts. Such larger impacts are more likely to be detected. This option would refine the program model—prior to formal impact evaluation on the statutory outcomes.

Option 3.3-3 Implementation Studies to Accompany Impact Evaluations (RQ3/Components)

What: Prior research has shown that apparently similar reemployment programs vary in their impacts (Klerman et al., 2019, Appendix A). That variation is likely, at least in part, due to variation in the

program models themselves.[44] This option would conduct implementation studies that document the components of a whole program being evaluated, in order to make it possible to identify what key components distinguish those programs that have larger impacts. This approach is important if conducting meta-analyses is a goal. Specifically, if whole-program impact evaluations consistently are accompanied by high-quality implementation studies that document the program's components, **Option 3.2-3/Cross-State Meta-Regression of Whole Program Evaluations** would become more feasible.

Beyond specific use to support component mega-regressions, implementation studies are more broadly valuable for purposes such as generating hypotheses for why a program was or was not as effective as hoped, identifying possible changes to components that might strengthen a program, and allowing other states to replicate programs that have been found effective.

Implementation studies typically involve interviews with program staff members at different levels to understand how the program is designed and carried out. Such implementation studies might also involve staff surveys, analysis of formal guidance on program administration and other program documentation, analysis of administrative data on services provided, and interviews with program participants. Data collection often occurs at more than one point in time, to document changes that occur, and would be important if random assignment continues for long periods.

Who, When, and How Much: Implementation studies are usually conducted by the same entity conducting the impact study. By providing general guidance, DOL could support consistency in the level and type of implementation data reported. Data collection for these studies typically occurs during random assignment to capture the experiences of sample members. This is a lower-cost option.

Why: Implementation studies are important to understanding how programs operated. They are also crucial for implementing **Option 3.2-3 Cross-State Meta-Regression of Whole Program Evaluations**.

Option 3.3-4 Annual Survey of State Program Characteristics (RQ3/Components)

What: This option would repeat annually some version of this project's survey of state programs (see Trutko et al., 2022). This option would then generate a database of state program characteristics through time, as well as documentation for that database. This would be similar to products produced for other programs; for example, DOL's *Comparison of State Unemployment Laws*.[45]

Who, When, and How Much: This option would likely be a DOL-directed activity. It could start immediately. To be useful in support of **Option 3.2-4/Cross-State Interrupted Time Series of Observational Data** to test the impact of components, the survey would need to be conducted approximately annually for five or more years. As a repeated survey this is a medium-cost option.

Why: The data produced by the survey are crucial for implementing **Option 3.2-4**. That data can also serve useful descriptive functions, as they did for *Comparison of State Unemployment Laws*.

3.4. Options to Address What Works Best for Whom

This section considers options related to **RQ4/What Works Best for Whom:** *How does the impact of components included and how they are provided vary with the characteristics of the claimant at initial claim?* Understanding whether program services are more effective for certain subgroups than others is important for program design and potentially individualizing services.

---

[44] Though other factors, such as differences in local labor market conditions at the time the evaluation was conducted, could also matter.

[45] https://oui.doleta.gov/unemploy/comparison/2020-2029/comparison2020.asp

Individualized service delivery presumes that what services are appropriate to provide is different for different claimants, depending at least in part on their characteristics at the initial claim. To provide evidence-based support for individualized service delivery, a program needs to know how the impact of components varies with claimant characteristics. Conversely, if some component is better for all claimants, there is no need for individualized service delivery (see Klerman 2017 for more on individualized service delivery).

Two examples help to clarify the issues. It seems plausible that claimants who will become reemployed quickly will not benefit much from intensive services often provided over longer periods. They will likely be reemployed before the intensive services can be delivered. It also seems plausible that we can identify those claimants based on information from their initial claim. Similarly, it seems plausible that some claimants should not be called in for the RESEA meeting earlier because they could find a job on their own. Options to address RQ4 would explore whether some claimant subgroups have large impacts from some program models and other subgroups have large impacts from other program models.

This research question can be viewed as a combination of **RQ2/Subgroups** (how do impacts vary with claimant characteristics) and **RQ3/Components** (how do impacts vary with components). The appropriate basic design is **Option 3.2-1/Individual-Level Random Assignment to Estimate Impact of a Component on RESEA Statutory Outcomes**.

However, this Option 3.2-1 requires larger samples, because it is trying to detect differential impacts simultaneously for both program models and subgroups. Given that **RQ2/Subgroups** evaluations likely require samples of roughly 250,000 claimants selected for RESEA to detect impacts on employment, sample sizes for this option would be even larger. Only the very largest studies are likely to detect even large differential impacts for subgroups. Detecting impacts on UI duration alone will require smaller, but still large samples. Detecting impacts on intermediate outcomes will be more feasible because sample size requirements are much lower. As noted earlier (see **Option 3.2-2/Individual-Level Random Assignment to Estimate Impact of a Component on Intermediate Outcomes Only**), impacts on intermediate outcomes do not always translate to impacts on final outcomes.

Options discussed in Section 3.3 to support estimating impact of components are also relevant for RQ4.

4. Conclusion

The previous two chapters of this report enumerated multiple evidence-building options for RESEA programs. This chapter looks across the options with twin goals. First, it identifies relatively high priority options, both in the short term and in the longer term. Second, the chapter explains how the options fit together—logically and sequentially. Exhibit 4-1 summarizes the key features of each of evaluation options related to whole program evaluations (gray), subgroups (red), and components (purple). Specifically, the exhibit shows whether the option is likely to be led by states or by DOL, how soon the activity might start and end, and a rough estimate of cost.

Exhibit 4-1. Summary of Options Discussed

Option	Who would lead?	When (to start and to finish)?	How much?
Estimating Impact of Whole Programs			
Option 2.1-1/Individual-Level Random Assignment	**Most states**: States that can randomize 30,000 to 50,000 RESEA-eligible UI claimants over 1-3 years	**To Start:** Approximately 6-12 months to prepare initial random assignment **To Finish:** Results available in 3-4 years, at a minimum	Higher cost
Option 2.1-2/Regression Discontinuity	**Few states**: States that use (or have used) profiling score in deciding which claimants to select for RESEA and have roughly 200,000 RESEA-eligible claimants "near" the profiling score cutoff	**To Start:** Immediately, if profiling model does not require revision **To Finish:** Results available in about 1 year if retrospective, and 4 or more years if prospective	Medium cost
Support to Estimate Impact of Whole Programs			
Option 2.2-1/Evaluation Technical Assistance	**DOL**	**To Start:** Can start immediately **To Finish:** Indefinite	Medium to higher cost
Option 2.2-2/Providing Common Analytic Tools	**DOL**	**To Start:** Can start immediately **To Finish:** Within 1 year	Lower to medium cost
Option 2.2-3/Providing Common Data	**DOL**	**To Start:** Short term, to begin defining data elements to collect **To Finish:** Ongoing as states submit data in the future	Higher cost
Option 2.2-4/Cost-Benefit Analysis	**Most states**: States able to complete Option 2.1-1 or 2.1-2	**To Start:** After completion of state impact evaluation in 3-5 years **To Finish:** Within 1 year of start, assuming cost data were collected during the impact evaluation; otherwise, 2 years	Medium cost
Option 2.2-5/Develop a Template for Cost-Benefit Analysis	**DOL**	**To Start:** Can start immediately **To Finish:** Within 1 year	Lower cost
Option 2.2-6/Synthesis of the Impact of a Generic RESEA Program	**DOL**	**To Start:** After completion of at least 2-3 state impact evaluations in 3-5 years **To Finish:** Within 1 year after starting	Lower cost

Option	Who would lead?	When (to start and to finish)?	How much?
Option 2.2-7/Sophisticated Synthetic Estimates of the Impact of a Specific State's RESEA Program	**DOL**	**To Start:** After completion of several impact evaluations in 4-7 years **To Finish:** 1-2 years after starting	Lower cost for a single round Medium cost if yearly updates are included
Option 2.2-8/Coordinating Evaluations in Support of Synthesis	**DOL**	**To Start:** Can start immediately **To Finish:** Within 1 year	Lower cost
Subgroup Analysis			
Option 2.3-1/Coordinating Subgroup Analysis–Reporting Guidance	**DOL**	**To Start:** Can start immediately **To Finish:** Within 1 year	Lower cost
Option 2.3-2/Synthesis of Subgroup Impact Estimates	**DOL**	**To Start:** After completion of at least 2-3 state impact evaluations in 3-5 years **To Finish:** Within 1 year after starting	Lower cost
Option 2.3-3/Disparate Impact of Selection Strategies	**DOL**	**To Start:** Can start immediately **To Finish:** Within 6 months of start	Lower cost
Estimating Impact of Components			
Option 3.2-1/Individual-Level Random Assignment to Estimate Impact of a Component on RESEA Statutory Outcomes	**Few states**: States that can randomize perhaps 150,000 RESEA-eligible UI claimants over 1-3 years	**To Start:** Approximately 6-12 months to prepare initial random assignment **To Finish:** Results available in 3-4 years, at a minimum	High cost
Option 3.2-2/Individual-Level Random Assignment to Estimate Impact of a Component on Intermediate Outcomes Only	**Most states**: Feasibility depends on component of interest	**To Start:** Approximately 6-12 months to prepare initial random assignment **To Finish:** Results available in 2 years, at a minimum	Low to medium cost
Option 3.2-3/Cross-State Meta-Regression of Whole Program Evaluations	**DOL**: Analyze experimental estimates of impact to estimate how impacts vary with program characteristics	**To Start:** After completion of several impact evaluations in 4-7 years **To Finish:** 1-2 years after starting	Medium cost
Option 3.2-4/Cross-State Interrupted Time Series of Observational Data	**DOL**: Observe how whole-program outcomes vary with changes in program design	**To Start:** After collection of 5-10 years of data **To Finish:** Within 1 year after starting	Medium cost
Support to Estimate Impact of Components			
Option 3.3-1/Consortia Building	**DOL and/or states**: States cooperate on evaluating the same program component	**To Start:** Can start immediately **To Finish:** Long term; results of experimental impact evaluations available in 3-4 years, at a minimum	Low to medium cost
Option 3.3-2/Deliberate Program Development for Pilots and Demonstrations	**DOL and/or states**: Consortia states specify details of program design	**To Start:** Can start development immediately **To Finish:** 1-3 years to complete	Medium cost

Option	Who would lead?	When (to start and to finish)?	How much?
Option 3.3-3/Implementation Studies to Accompany Impact Evaluations	**All states**: Carefully document the components of a whole program	**To Start:** Can start development immediately **To Finish:** 1-2 years to complete	Lower cost
Option 3.3-4/Annual Survey of State Program Characteristics	**DOL**	**To Start:** Can start adapting existing RESEA Implementation Study Survey immediately **To Finish:** Indefinite; survey conducted annually for 5-10 years	Medium cost

Exhibit 4-2. Broad Considerations for Evaluation of Whole Programs, Subgroups, and Components

What is being evaluated?	Positives	Negatives and challenges
Whole Program	• Provides information on how successfully the program is meeting its ultimate objectives to improve claimants' employment outcomes and reduce UI claim duration. • Relative to a component evaluation, a whole-program evaluation is logistically easier and can be done with smaller samples.	• On their own, the quantitative impact findings provide little information on how programs might be improved.
Subgroup	• Provides information that may help improve program impacts by identifying which UI claimants benefit most from being selected for RESEA. • Can provide information on disparities that may indicate areas to focus program improvements. • Can typically be produced at very low cost as an add-on to a whole-program or component evaluations.	• On their own, the quantitative findings may provide little information on the reasons behind subgroup differences. • Larger samples are needed to estimate impacts for subgroups than for participants as a whole.
Component	• Provides information on which components contribute most to program impacts. • Can provide information on changes to a program's design that may improve its effectiveness, which can have great value for program design decisions.	• Evaluations are logistically more challenging because they require operating more than one version of a program at the same site. • Relative to a whole-program evaluation, needs larger (often very large) samples to estimate impacts of the component on final outcomes (employment and UI duration). • Many components aim to affect an intermediate outcome (e.g., meeting attendance, use of AJC services). Impacts on intermediate outcomes can often be evaluated with much smaller samples, but those estimates leave uncertain whether the observed intermediate impacts will translate into impacts on final outcomes (employment and UI duration).

Exhibit 4-1 organizes the options by whether they apply to their corresponding research question: whole programs, subgroups, or components. Exhibit 4-2 briefly summarizes the advantages "positives") and disadvantages ("negatives and challenges") of addressing each of these research questions.

Addressing each of these three research questions would provide value and the three research questions are potentially mutually complementary. Similarly, descriptive evidence (e.g., from implementation research) and causal impact evidence are also potentially mutually complementary. As such, it is useful to think through an evidence-building strategy in a sequential fashion. This is particularly important for RESEA programs, which have some short-term requirements to meet, along with ongoing, longer-term learning interests.

The remainder chapter discusses how the options discussed in Chapters 2 and 3 and summarized in Exhibit 4-1 might fit into short-term and long-term strategies. The chapter proceeds in four sections. The first section considers short-term evaluation priorities; that is, studies that would address the short-term goal of showing the effectiveness of a generic RESEA program—or even of the RESEA program in a specific state. Given impact evaluation timelines, many of these activities could be started in 2021, but results are likely several years away. The second section considers longer-term priorities and their implications; that is, studies that would address the longer-term goal of providing insights into how to improve RESEA program impacts through adding, deleting, or changing components, Section 4.4 provides some closing thoughts.

4.1. Short-Term Priorities

For both DOL and the states, the highest short-term priority is likely to generate sufficient evidence to satisfy the statutory requirement that RESEA programs be demonstrated effective. That determination that existing evidence is sufficient to deem a generic[46] RESEA program effective will be made by DOL's Clearinghouse for Labor Evaluation and Research (CLEAR) effort (www.clear.dol.gov). Under the evidence criteria established in DOL guidance (Unemployment Insurance Training Letter 01-20), the most direct route to that determination requires at least two experimental evaluations that both: (1) meet CLEAR standards for study evidence quality[47] and (2) find statistically significant evidence of positive impacts (**Option 2.1-1/Individual-Level Random Assignment**). COVID has pushed back state evaluation timelines such that CLEAR might have enough studies in time for Fiscal Year (FY) 2025 or 2026 state RESEA Plans, but FY 2027 or even FY 2028 seem more likely (see Appendix Section C.3).

---

[46] Here "generic" means that CLEAR has found RESEA overall as effective, based on evidence from whichever states have completed evaluations. An alternative would be to deem a state's own program effective—and perhaps to require that.

[47] For *studies* that examine the causal effect (or "impact") of a labor-related intervention, CLEAR assigns to each study a rating that reflects the credibility of the evidence that the study presents—that is, how confident we can be that the findings presented by the study truly reflect the causal effect of that intervention, rather than some other factor that might also influence outcomes. For details on how CLEAR rates the credibility of evidence of causal studies, see CLEAR's Causal Evidence Guidelines at https://clear.dol.gov/about. Study ratings do not indicate whether the intervention itself is effective (i.e., these study ratings do not consider whether there is evidence that the program improves outcomes).

For reemployment-related *interventions*, CLEAR also provides a rating of evidence of effectiveness (i.e., whether the intervention improves outcomes). Those ratings take into account all sufficiently credible studies of the intervention (specifically, studies that received a High or Moderate CLEAR rating) to rate the extent of causal evidence that the intervention is effective. Under Social Security Act Section 306, interventions are required to have a High or Moderate rating to be eligible for funding. For details of how interventions are rated, see CLEAR's RESEA page at https://clear.dol.gov/reemployment-services-and-eligibility-assessments-resea.

DOL is providing two types of evaluation technical assistance (TA): (1) general evaluation TA to all states; and (2) additional customized evaluation TA to a select, small group of states. Together this evaluation TA will likely speed when DOL CLEAR can deem a generic RESEA program effective. Continuing evaluation TA at least until results become available from the 2020 Cohort would substantially improve the likelihood of having high-quality evaluations that meet rigorous CLEAR standards for credibility of causal evidence (**Option 2.2-1/Evaluation Technical Assistance**). Option 2.2-1 therefore seems worthy of serious consideration, as does synthesis of those studies to make that determination **(Option 2.2-6/Synthesis of the Impact of a Generic RESEA Program**).

These evaluation TA and synthesis activities would likely be sufficient to deem a *generic* RESEA program effective.[48] If individual states (or DOL) want to demonstrate that a *specific* state's RESEA program is effective, additional support would likely be useful. Such additional support would start with whole-program evaluation TA (**Option 2.1-1/Individual-Level Random Assignment**) to more states and for longer than is implied by current funding. Evaluation TA might also include steps to lower the cost and the skill level needed by state evaluators to conduct such evaluations (**Option 2.2-2/Providing Common Analytic Tools** for analysis; **Option 2.2-3/Providing Common Data** for data).

When a dozen or more state **RQ1/Whole Program** estimates become available,[49] **Option 2.2-7/Sophisticated Synthetic Estimates of the Impact of a Specific State's RESEA Program** would be feasible. By accumulating strong evidence from evaluations through time and across states, sophisticated synthesis of whole-program evaluations would help to address the lack of precision in single-year estimates of impact from evaluations in smaller states. This evidence synthesis option could also help to generate insights into how the impact of the RESEA program varies with the business cycle.

Finally, understanding how program impact varies with claimant characteristics (**RQ2/Subgroups**) would produce useful insights for program design. However, few single-state studies are likely to be large enough to estimate subgroup impacts with a useful level of precision. Pooling across multiple states' whole-program experimental impact evaluations (**Option 2.1-1**) can address that limitation and provide insights about effective programs. To extract those insights, state whole-program experimental impact evaluations need to conduct their analyses in a common way and report a common set of outcomes. Those outcomes are not difficult to report, but they are likely not worth reporting for each state alone. This is because, considered alone, most states will not have sample sizes large enough to statistically detect impacts. To make possible such analyses, DOL may want to consider funding the development of subgroup reporting standards (**Option 2.3-1/Coordinating Subgroup Analysis–Reporting Guidance**) and then a subgroup synthesis (**Option 2.3-2/Synthesis of Subgroup Impact Estimates**). Neither of these options would be very expensive or methodologically complicated.

4.2. Longer-Term Priorities

Though satisfying the RESEA statutory requirement is likely to be a pressing short-term priority, the longer-term priority is to generate evidence to help *improve* outcomes for UI claimants participating in RESEA or similar reemployment services. One approach to improving program outcomes—selecting for RESEA those claimants who will benefit more—has been noted in the previous section (and Section 2.3 on subgroup analysis).

---

[48] Here "generic" means that CLEAR has found RESEA overall as effective, based on evidence from whichever states have completed evaluations. An alternative would be to deem a state's own program effective—and perhaps to require that.

[49] Likely at least half a dozen.

The other approach to improving program outcomes is to conduct **RQ3/Components** experimental impact evaluations (**Option 3.2-1/Individual-Level Random Assignment to Estimate Impact of a Component on RESEA Statutory Outcomes**) that evaluate the change in outcomes as a state adds, deletes, or changes a component. Exhibit 3-2 provided a long list of possible components to evaluate. Section 3.1 called out four as particularly promising for evaluation: (1) remote services; (2) intensive reemployment services; (3) responses to non-attendance at the RESEA meeting; and (4) enforcement of the ongoing eligibility requirement.

Pursuing a pilot or demonstration is an important long-term strategy. "Intensive services" is a useful label for a *general direction* for RESEA interventions. An evaluation would estimate the impact of a *specific version* of intensive services. Ideally, the component to be evaluated would incorporate the field's sense of the most promising form of the component. Then the preliminary design for that component would be piloted and refined—before starting a large-scale, and more expensive impact evaluation (Epstein and Klerman, 2012). **Option 3.3-2/Deliberate Program Development for Pilots and Demonstrations** describes a process to develop a candidate component. Taking intensive services as an example, details to be agreed upon might include: *When would intensive services be provided? To whom? What would be the content and intensity (e.g., minutes per claimant) of those services?* Once initial agreement was reached, the candidate component could be refined through piloting and formative evaluation. Such efforts could start immediately. Depending on the number of rounds of piloting and refinement, such efforts could take less than a year or several years. That process of piloting and refinement would need to finish before an impact study starts.

Once the impact evaluation starts under any of the options, it could be accompanied by cost studies and a cost-benefit study (**Option 2.2-4/Cost-Benefit Analysis**) and an implementation study (**Option 3.3-3/Implementation Studies to Accompany Impact Evaluations**).

4.3. Discussion

Consistent with the RESEA program's federal structure, the Social Security Act gives primary responsibility for evaluation of RESEA and most evaluation funding to states. Chapters 2 and 3 have assumed that states or consortia of states would conduct the impact evaluations. Those impact evaluations are the high-cost options. Without them, the options to *support* impact evaluations have limited value.

Nevertheless, three considerations imply a potential role that DOL might play in inter-state coordination for RESEA evaluations. First, as discussed in detail in Appendix Section B.4, almost all states can provide ample sample to address **RQ1/Whole Program** alone, perhaps with several years of random assignment. In contrast, few states alone can address the other research questions—**RQ2/Subgroups**, **RQ3/Components**, or **RQ4/What Works Best for Whom**. With only one year of randomization, appropriate sample sizes would be a 10th or more of nationwide RESEA claimants. Instead, addressing these other—and crucial—research questions is likely to require multi-state consortia. Building such multi-state consortia is fundamentally a coordination problem. **Option 3.3-1/Consortia Building** discussed issues in forming and stimulating the formation of such consortia.

Second, some common tasks in support of impact evaluations would benefit from coordination. For **RQ2/Subgroups**, such common tasks include developing reporting standards (**Option 2.3-1/Coordinating Subgroup Analysis–Reporting Guidance**) and synthesizing the estimates (**Option 2.3-2/Synthesis of Subgroup Impact Estimates**). For **RQ3/Components** or **RQ4/What Works Best for Whom**, such common tasks also involve developing the program model(s) to be evaluated (**Option 3.3-2/Deliberate Program Development for Pilots and Demonstrations**). Because these tasks need to happen only once across multiple evaluations, this is a substantial coordination challenge.

Third, there is a "free rider" problem. When one state runs an evaluation, all states benefit. As a result, each state has an incentive to free ride on another state's evaluation and not conduct its own evaluation.

That free rider problem is intrinsic to research; it is exacerbated when states control the funds and choose whether and what to evaluate. Addressing the free rider problem would also benefit from inter-state coordination.

These three considerations suggest that coordination would be useful. For several reasons, DOL or an entity working on its behalf is well positioned to stimulate such coordination and evidence-building that may not otherwise occur: DOL has national responsibilities and perspective; DOL has funds to stimulate coordination; DOL writes the guidance; and DOL reviews and approves state RESEA Plans.

Appendix A: Completed Studies and Appropriate Sample Sizes

A large evidence base exists on the effects of U.S. pre-RESEA reemployment programs. The REA Impact Study (Klerman et al. 2019; Sections 2.3 and 2.4) and CLEAR; https://clear.dol.gov/synthesis-report/reemployment-synthesis) have reviewed that literature—both the pre-REA studies and the REA studies. This appendix provides a brief discussion, focused on appropriate sample sizes.

Overall Pattern of Impacts. Exhibits A-1 and A-2 report the results of an exploratory random-effects meta-analysis of the experimental evaluations of REA in 10 states[50]—all using the canonical design described in Appendix B. We included impact estimates on UI duration and Q2 employment outcomes from evaluations of REA that have been reviewed by CLEAR and assigned a High or Moderate causal evidence rating.

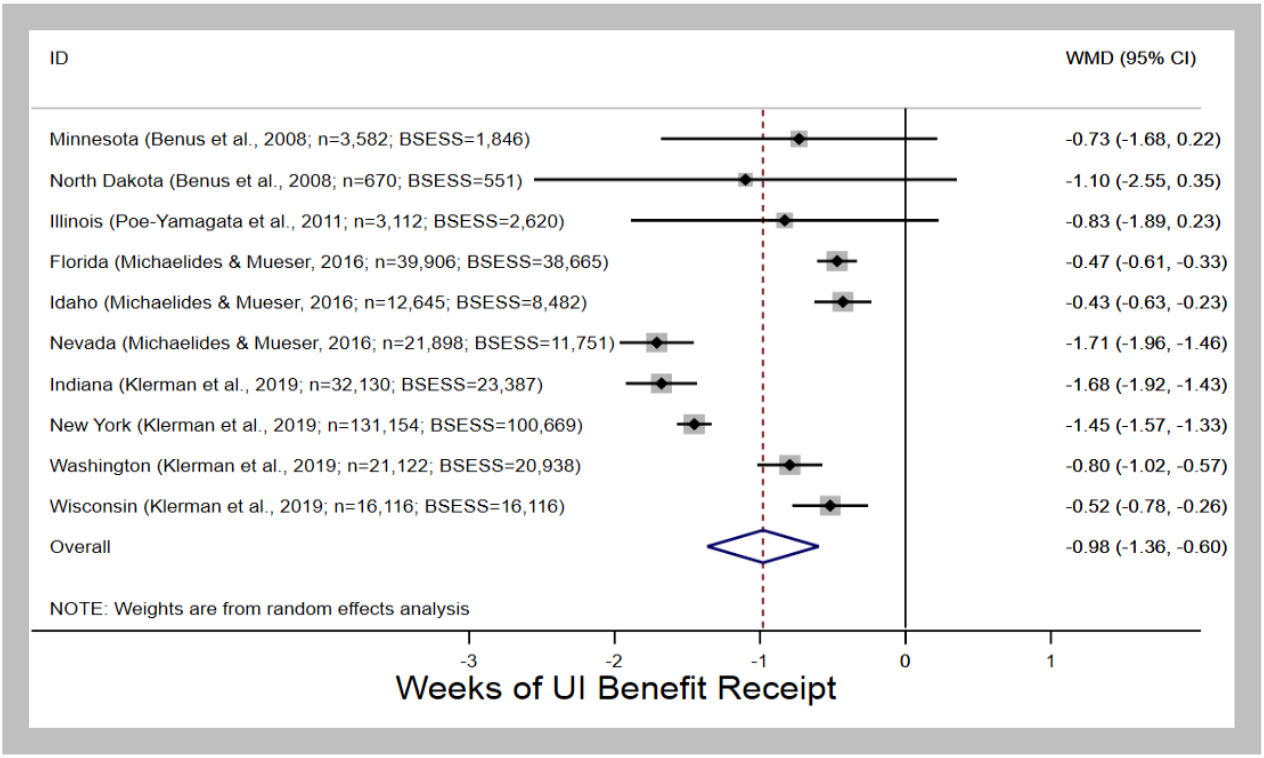
Rows of the exhibits give the results for each state in each study. The dots give the estimate of impact; the bars give the range of plausible values (formally the 95 percent confidence interval); and the size of the shaded squares correspond to the weight of each study in determining the overall estimate. The final row provides a statistical summary of the studies. The center of the diamond, which aligns with the vertical dotted line, gives the meta-analysis's best estimate of the average impact of REA programs (combining across studies); the edges of the diamond give the range of possible average impacts.

The labels give two sample sizes. The first sample size is the actual reported number of claimants in the analysis. This number can be misleading. For a given sample size, unbalanced treatment/control counts lead to less precision (i.e., larger standard errors). To address that, the labels also give the equivalent sample size with exact treatment/control balance (labelled as BSESS; that is, Balance Sample Equivalent Sample Size).

[50] Meta-analysis provides a statistical approach to combining the results from many studies to estimate an average impact of the intervention across all studies. Studies are weighted by the precision of their estimates. As a result, large studies are usually weighted more heavily.

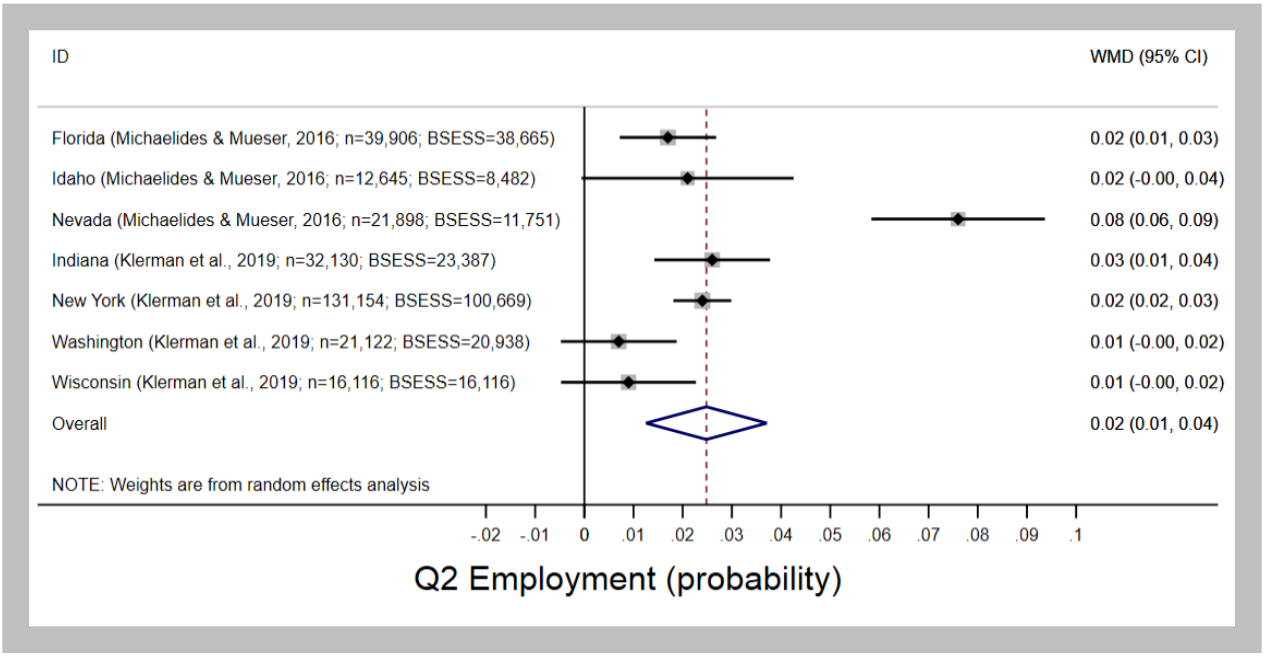
Random-effects (RE) meta-analyses assume that there are two potential sources of variation in impact estimates across studies: differences due to sampling error and differences in the true impacts. Alternative approaches to meta-analysis assume only the former. RE meta-analysis allows for better out-of-sample predictions (e.g., predictions about what would happen in a new state that implemented a program). However, the RE method does not work well when there are only a small number of studies, because when there are few studies, it can be difficult to estimate the variability in true effects.

Exhibit A-1. Meta-Analysis of Weeks of UI Benefit Receipt



Heterogeneity chi-squared =220.81, *p* = .000
I^2 =95.94%
Test of WMD=0, *p* = .0000
Notes:
We show estimates of heterogeneity and tests of significance of the overall analysis. These are a chi-squared test statistic for a test of heterogeneity and the associated p-value from that test, I^2 as the proportion of total variability in effects due to between-study heterogeneity, and the p-value of a test of significance for the weighted mean difference (WMD) effect size.
BSESS = Balanced Sample Equivalent Sample Size. This is the sample size that would have yielded the same precision if the sample were balanced (i.e., if there was a 50-50 allocation to intervention and control).
WMD = weighted mean difference between each point estimate and the null hypothesis (i.e., that the true impact is zero). Horizontal bars indicate the 95% confidence interval for each point estimate.

Exhibit A-2. Meta-Analysis of Q2 Employment



Heterogeneity chi-squared = 47.55 (df=6), p = .000
I^2 = 87.4%
Test of WMD=0, p = .0000
Notes:
We show estimates of heterogeneity and tests of significance of the overall analysis. These are a chi-squared test statistic for a test of heterogeneity and the associated p-value from that test, I^2 as the proportion of total variability in effects due to between-study heterogeneity, and the p-value of a test of significance for the weighted mean difference (WMD) effect size.
BSESS = Balanced Sample Equivalent Sample Size. This is the sample size that would have yielded the same precision if the sample were balanced (i.e., if there was a 50-50 allocation to intervention and control).
WMD = weighted mean difference between each point estimate and the null hypothesis (i.e., that the true impact is zero). Horizontal bars indicate the 95% confidence interval (CI) for each point estimate.

That analysis finds that REA consistently lowered UI durations—on average by about a week. The exceptions are three very small studies, each with unadjusted sample sizes of less than 4,000 individuals; adjusted sample sizes under 3,000).[51] There is also evidence that RESEA increases employment and earnings—on average by about 2.5 percentage points—though the effect is not as consistent. Four of seven studies show clear evidence of impact; the other three do not.

Appropriate Sample Sizes. As just alluded to, sample size issues appear to explain much of the lack of consistent impact on employment and earnings. Both formal statistical analysis and inspection of Exhibits A-1 and A-2 suggest that a random assignment study that randomized 10,000 REA-eligible claimants—half selected for REA, the other half not selected—*seems likely* to detect impacts on *UI weeks* given mean outcome for those not selected (control group) of about 15 weeks and such programs' typical impacts of 0.5 to 1.5 weeks.

[51] The adjusted sample size (BSES) controls for unbalanced randomization. It is the sample size with equal intervention/control group split that would yield the same precision.

The situation with respect to Q2 Employment is different. Given mean outcome for those not selected of about three-quarters employed in Q2 and impacts of about 2 percentage points,[52] much larger sample sizes are needed to detect impacts on employment, given the size of impacts that seem to be typical. Even a sample of 30,000 claimants *does not seem likely* to consistently detect impacts on Q2 employment. These large sample sizes for consistently detecting impacts on employment drive evaluation design for impact studies of RESEA (see the discussion throughout the body of this report).

Variation in True Program Impact. Formal tests (see the "heterogeneity chi-squared" immediately following the exhibit) show strong evidence of variation in impact across states. Even given moderate sample sizes, considerable inter-state variation in impacts should be expected simply due to random factors. Nevertheless, the random-effects meta-analysis suggest that there is considerable state-to-state variation in true program impacts (above and beyond simple sampling variability).

This strong evidence of variation in true program impacts has two important implications. First, impacts estimated in one state might not always apply to another state. Second, variation in true impacts implies that to be assured of detecting impacts, samples should probably be *larger* than the earlier 30,000 claimant figure. This is because the study design needs to consider the possibility that the true impact in this state is smaller than the average impact (and larger samples are needed to detect smaller impacts).

Differential Impacts by Claimant Characteristics. The REA Impact Study (Klerman et al., 2019) provides some evidence on *for whom* impacts are likely to be larger. Impacts on UI weeks were larger for claimants with smaller pre-unemployment earnings and smaller UI weekly benefit amounts. Impacts were not consistently larger for claimants with higher profiling scores.[53] Note that these differential impact estimates are for UI weeks. Even the REA Impact Study's huge samples—for these purposes about 150,000 claimants—were not sufficient to detect differential impacts on employment.

Differential Impact by Program Components. There is only limited evidence on what *program features* are likely to lead to larger impacts. DOL has developed standards to rate the effectiveness of reemployment interventions. Applying those standards requires defining interventions, for the purposes of summarizing evidence on those interventions. CLEAR has done this for evidence available through roughly early 2019. Ongoing consideration of those groupings suggests that, in practice, CLEAR's approach primarily groups studies by funding source (e.g., REA, WPRS, WPRS-predecessor programs). Unfortunately, those categorizations do not appear to correspond to well-defined components or even distinct program models. That is, the interventions are neither highly consistent within a given intervention category nor clearly distinct between intervention categories.

The REA Impact Study implemented multi-armed random assignment to estimate the impact of multiple versus single REA meetings. The study showed that in one state, multiple REA meetings cut UI weeks; in the other state, such multiple meetings did not cut UI weeks. (Again, this means there was variation in impact findings—in this case for the impact of components—across states.) The variation might be related to details of the states' implementation of multiple REA meetings. This suggests that inferences about components will benefit from careful definition and implementation of the component being evaluated. Broad categories—e.g., "multiple meetings"—might not be sufficient. Furthermore, note that even samples of more than 30,000 were too small to detect differential impacts of multiple meetings on earnings.

---

[52] Control group mean is about 70 percent.

[53] Black et al. (2003) does not find evidence of differential impact by profiling score, but their study's power to detect such differential impact is weak.

Causal Pathways. The REA Impact Study also included analyses to apportion impact between (1) eligibility assessment, (2) enforcement of the requirement to attend the REA meeting, and (3) reemployment services provided at and following the initial REA meeting. The study assesses the totality of evidence as suggesting:

- Only a minimal role for *eligibility assessment*. For several reasons, few eligibility issues are detected. Furthermore, the consequences of a detected issue appear to be relatively small. In net, the impact is perhaps a 10th of a week.
- A large role for *enforcement of the requirement to attend the meeting*. Non-attendance is common.[54] Inasmuch as states suspend benefits until attendance, the implied impact on UI weeks is large—more than 1 week. An impact of 1 week is very roughly the total impact of most REA and pre-REA programs.
- A moderate role for *reemployment services*. In net, the impact is perhaps half a week per claimant selected for REA (recall that more than a third never attend the meeting; they cannot benefit from its services/assistance).[55]

Note that these impacts are all with respect to UI weeks. Even the pooled samples—for these purposes, approximately 100,000 REA-eligible UI claimants—were insufficient to estimate the separate impact of these causal pathways on employment.

[54] Two caveats to this statement seem relevant. First, attendance rates seem to be higher for remote meetings (Trutko, et al, 2022). This might suggest smaller impacts on UI weeks for states not returning to predominantly in-person meetings. Second, there is some evidence that non-attendance is related to failure to contact (Darling et al, 2017; Trutko, et al, 2022). This might suggest that more intensive efforts to contact claimants (including better contact information) or some required activity before the RESEA meeting will cut the impact of RESEA despite likely being a positive overall.

[55] In contrast to the discussion in the previous footnote, higher attendance—through remote meetings or better contact information—might suggest larger impacts of assistance.

Appendix B: Canonical Design

Almost all impact studies of reemployment interventions use what this report calls the "canonical design." This appendix describes that design. Specifically, Appendix Section B.1 sketches the design itself. Then Appendix Sections B.2 and B.3 consider issues related to sample size for 0/1 Tests (i.e., no RESEA vs. RESEA) and for A/B Tests (i.e., one version/component of RESEA vs. some other version/component of RESEA), respectively. Appendix Section B.4 considers strategies to achieve appropriate sample sizes. Finally, Appendix Section B.5 discusses the implications of the analysis for design.

B.1. The Design Itself

The basics of the canonical design are common for both 0/1 Tests and A/B Tests. The evaluation specifies a pool of eligible claimants who are randomly assigned to one of the two treatment arms, according to some pre-specified randomization fraction, by the RESEA program's statewide computerized scheduling system.

Given concerns about sufficient sample size (see below), the optimal strategy will usually be to randomly assign every RESEA-eligible UI claimant. In particular, the available budget sets the number of claimants to be scheduled. The randomization fraction is then set to meet the budget.[56]

Outcomes are measured using administrative data. Underlying data come from state UI systems. The federal National Directory of New Hires (NDNH) aggregates those data into a single national database. When access to NDNH is possible, using it will usually have much lower cost than other sources or having to compile the data in other ways—especially for multi-state studies.

Statutorily, the two key outcomes are UI weeks and Q2 employment.

- ***UI Weeks***. State UI benefit payment systems record UI weeks as well as dollars of benefits paid. That information is reported to NDNH, which reports quarterly UI benefits paid. That quarterly data is probably sufficient for the statutory purpose. The state's own records have weekly detail and are thus preferable.
- ***Q2 Employment***. Q2 employment in a state can be inferred from positive earnings in state quarterly UI wage filings.[57] Those data should provide employment and earnings for any calendar quarter. These state-specific data miss out-of-state earnings. Such state quarterly earnings data are probably sufficient for the statutory purpose. NDNH aggregates state-specific data across states and augments

---

[56] An example will make the suggested approach clear. Suppose a state has 6,000 new RESEA-eligible UC claims per month, a budget to conduct 1,000 initial RESEA meetings per month, and about a third of claimants scheduled for an initial RESEA meeting ever attend. Given these assumptions, the state might set the randomization fraction to 25 percent; that is, one in four eligible claimants (1,500 per month) are randomly selected for RESEA, of whom about 1,000 would be expected to attend the initial meeting—reaching the state's budgeted goal and exhausting its available funds.

More general schemes would vary randomization fractions. For example, a state might adjust randomization fractions so that even claimants with low profiling scores have a substantial probability of being selected, but those with higher profiling scores have an even higher probability of being selected. This document is not the place for an extended discussion of such schemes.

[57] State UC wage records are known to miss self-employment and employment in government and in the informal sector (i.e., for which UC taxes are not paid). These issues are usually considered to be second order—and therefore ignorable.

them with information on federal employment. Thus, unlike for UI weeks, for Q2 employment, NDNH data are preferable to state-specific data.

Two other ultimate outcomes are likely to be of considerable interest. For cost-benefit analyses, impacts on total UI *benefits paid* are of more interest than impacts on UI *weeks* (as total benefits incorporates weeks receiving the benefits). Impacts on *Q2 earnings* are a better—but apparently harder to estimate—proxy for the magnitude of a program's impact on employment. Both of these outcomes are measured both in state administrative data and in the NDNH.

Information on intermediate outcomes—whether UI was claimed in a week (in particular, when UI was not paid), attendance at RESEA meetings, attendance at other reemployment services activities, responses to non-attendance—are also of interest and usually available in other state administrative data systems.

B.2. Sample Sizes for 0/1 Tests

To demonstrate effectiveness, the Statute and OUI's guidance require showing impact on both UI *weeks* and *Q2 employment*. Examination of existing estimates (see Appendix A.1) and formal analysis suggest that samples appropriate to detect impacts on Q2 employment are much larger—perhaps 10 times—than samples appropriate to detect impact on UI weeks. Thus, discussion of sample sizes can focus on Q2 employment. Any sample large enough to detect impact on Q2 employment is easily large enough to detect impact on UI weeks.

A simple calculation gives a rough sense of sample sizes. The RESEA program is the next generation of REA programs. Across all studies, the meta-analysis presented in Section A.1 suggests an average impact of REA (*no REA vs. REA*) on Q2 employment of about 2 percentage points, relative to a control group level of about 23 percent (see Exhibit A-2). Noting that Q2 employment is a binary outcome (employed *yes/no*) and applying the standard formula (two-sided test at 5 percent; $1\text{-}\alpha$=80 percent) implies a sample size of about 18,000. Details matter a little. Considering covariates will cut the sample slightly. Unbalanced randomization will increase the sample slightly. A 10 percent test will decrease the sample slightly. An evaluator designing a specific evaluation should do a more precise calculation. For the purposes of this document, this rough estimate of 18,000 is sufficient.

More fundamentally, setting sample size for average impact is probably not the right strategy. If the appropriate impact for this state is slightly larger, little is lost. If the appropriate impact turns out to be a little bit smaller, using the average impact is likely to lead to not detecting an impact. Given the considerable investment of time and resources to conduct a study well, those designing the studies probably want to set samples larger than implied by the simple calculation—that is, not 18,000 but perhaps 30,000, and arguably several times larger than that.

B.3. Sample Sizes for A/B Tests

Differential impacts of different RESEA programs (i.e., *Program Model A vs. Program Model B*) are likely to be much smaller than impacts of *no RESEA vs. RESEA*. It follows that appropriate sample sizes will be much larger. If, as is plausible, the differential impact is half of the overall impact, the appropriate sample is four times as large. If the differential impact is (slightly less than) a third of the overall impact, the appropriate sample is 10 times as large. If the differential impact is a fifth of the overall impact, the appropriate sample is 25 times as large. The previous section discussed samples for a 0/1 Test of 30,000 or more, which implies plausible appropriate sample sizes for A/B Tests starting at 120,000 and often several times larger. In practice, an A/B Test study might accept smaller samples—perhaps 50,000, understanding that that leaves a substantial chance of not detecting smaller impacts on Q2 Employment. How acceptable that risk is depends on the costly of the study and whether the components analysis is an add-on to a whole program evaluation (as part of a multi-arm study) or is the sole focus of the evaluation. If the components analysis is an add-on, then the smaller sample size and corresponding greater risk of failing to detect impacts may be more acceptable.

The sample size challenge for A/B tests is considerably more daunting than even this simple comparison of sample sizes would suggest. For 0/1 Testing, potential sample is every claimant *not exempt* from RESEA, whether or not selected. For A/B Testing, potential sample is every claimant *selected*. The number of non-exempt claimants is often much larger than the number of claimants actually selected. We return to this issue in the next appendix section.

B.4. Sample Strategies

The previous appendix sections have suggested sample sizes ranging from 18,000 RESEA-eligible claimants to considerably more than 120,000 RESEA-selected claimants. The feasibility of these sample sizes varies widely across states. Exhibit B-1 gives estimates of the number of UI claimants selected for RESEA in FY 2019.[58] Estimates of the number of non-exempt claimants not selected are not available. Limited available evidence suggests that most states are selecting well less than half of their non-exempt claimants. Inasmuch as that is right, then *in a year,* two-thirds of the states can get the 30,000 cases. Most other states could get to 30,000 in two to three years.

Exhibit B-1. RESEA-Selected Claimants by State (FY 2019)

N selected	N	State
100,000+	2	CA, NY
50,000–100,000	5	MA, NC, PA, TX, WA
25,000–50,000	7	AZ, FL, IN, MN, NJ, OR, WI
15,000–25,000	6	LA, MD, MI, MO, PR, TN
10,000–15,000	9	CT, IA, IL, KS, KY, OH, OK, UT, VA
5,000–10,000	12	AL, AR, CO, DC, GA, ID, NE, NH, NM, NV, SC, WV
0–5,000	10	AK, DE, HI, MS, MT, ND, RI, SD, VI, VT

Source: ETA 9128 reports. Accessed January 19, 2021. https://oui.doleta.gov/unemploy/DataDownloads.asp.

A/B Tests require larger samples. For a nearly best case—that is, a large component (one with a projected differential impact of half the overall impact)—appropriate sample size is roughly 120,000 RESEA-selected claimants. Only the two states have that many cases in a single year. Perhaps an additional five states would have that large of a sample in two years and an additional five states would in three years. Sample sizes for components with projected differential impacts smaller than half the overall impact are several times larger. A/B Tests would not be feasible for any state alone randomizing for one year and often not feasible randomizing over several years.

There are at least three strategies for achieving these sample sizes. The previous paragraph has alluded to two of them: randomizing for more than one year and "large components." With respect to larger RESEA program components, these sample size considerations suggest starting by testing an intensive form of the component. Once the intensive form is demonstrated effective, progressively less intensive forms of the intervention can be tested.

Joint evaluation across multiple states is a third strategy to increase sample size. Doing so has contractual and operational challenges. Clearly, the intervention being tested needs to be common—or at least similar. Slightly less obviously, the basic RESEA programs—before the component is added, deleted, or changed —should be similar. For example, it might be problematic to pool estimates of the impact of an

[58] FY 2020 is more recent but covers the (highly unusual) pandemic period, and state reporting is incomplete.

intervention across states that vary on some other significant feature (e.g., improved messaging both in states that "suspend until attend" and in states that do not implement "suspend until attend").

The implication of this analysis is that to imagine and argue for an evaluation for which it is infeasible to gain enough sample—even with these strategies—presages disappointment when the results come in.

Appendix C: Plans for Studies of RESEA

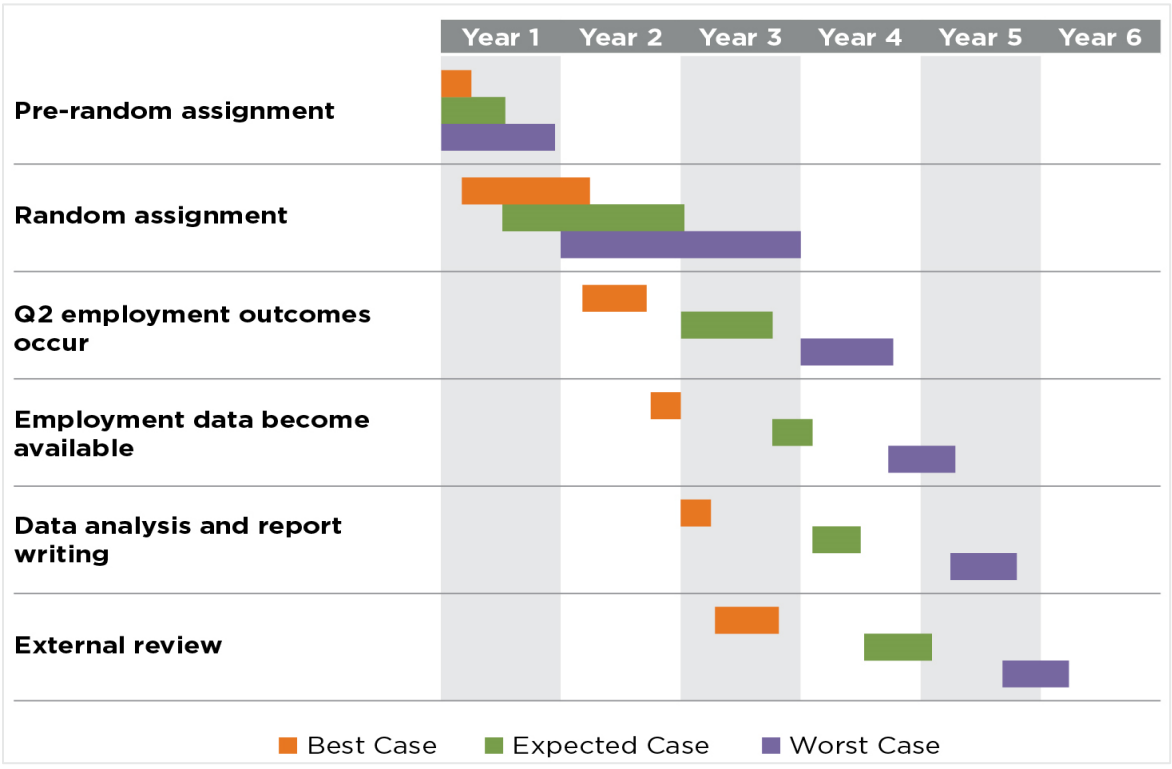
As noted in Chapter 4, the most direct way to meet the statutory evidence requirement would be for a generic RESEA program to be deemed as "demonstrated effective," with a High rating. That point will be reached when CLEAR identifies two state whole-program evaluations that both meet standards (i.e., receive a high or moderate rating) and detect an impact on both *UI weeks* and *Q2 employment.*

As of mid-2021, there were no completed studies of RESEA and no studies that had begun random assignment. This appendix considers the status of state RESEA evaluations and when there might be enough completed studies—of sufficient credibility, which detect impacts—for CLEAR to deem a generic RESEA program demonstrated effective. Specifically, Appendix Section C.1 considers the timeline for an experimental evaluation of RESEA. Appendix Section C.2 considers where state evaluation activities are relative to that timeline. Appendix Section C.3 puts the two pieces together to project when CLEAR is likely to deem a generic RESEA program demonstrated effective. Appendix Section C.4 notes some joint DOL and CLEAR steps that might allow states to use RESEA evidence a plan year earlier.

C.1. State Evaluation Timelines

Exhibit C-1 graphically depicts the nearly best-case timeline for an experimental evaluation of a whole-state RESEA program. The timeline *starts* from when an evaluator is engaged (i.e., an appropriate selection process has been completed and an appropriate contract executed) and the basic design has been selected (usually random assignment of the state's whole RESEA program). If the evaluator is to be selected by an open and competitive bidding process, these steps could require six months or more (to develop and issue the Request for Proposals, allow bidders to write compliant proposals, review the proposals, select a contractor, and execute the contract). Time for these contract activities is *not* included in the timelines below.

Exhibit C-1. Timeline for State Evaluations—Best, Expected, and Worst Cases



Source: Abt Associates
Note: Timelines start from when a contract has been executed and a basic design (i.e., random assignment, for which offices) has been settled.

There follows:

- **Pre-random assignment** (3+ months). Develop, review (by the state and likely by the DOL-funded evaluation TA provider), and approve an Evaluation Design Report. Then develop, insert, and test the random assignment processes.
- **Random assignment** (12+ months). Larger states will be able to achieve the target 30,000 RESEA-eligible claimants in a year. Smaller states will likely need two or more years. Even larger states might not detect impacts in a year. If not, they will probably want to continue random assignment until they detect an impact.
- **Q2 employment outcomes occur** (6 months). By definition, *Q2 employment* outcomes—that is, the second full calendar quarter after random assignment—will not occur until six to nine months after randomization.
- **Employment data become available** (3+ months). Employers report employment data to state UI programs for calendar quarters and with a lag of about three months. This three-month lag estimate, thus, implicitly assumes that random assignment begins on the first day of a calendar quarter. Otherwise, this interval could be as long as six months.
- **Data analysis and report writing** (3+ months). Presumably, writing the report will involve an initial draft, internal technical review by the evaluator team, and then professional editing and formatting. Furthermore, this three-month estimate assumes that data processing and analysis programming were generated on incomplete data, then when final data become available all that needs to happen is to re-run already tested programs. That is an expensive approach. A lower-cost approach would test the programming on final data. But doing that increases the time required to perhaps six months. Furthermore, this three-month estimate assumes a simple and short report.
- **External review** (6 months). State staff will want to review the report before publication. The contractor will need to revise. Several rounds of review and revision are not uncommon. Pre-publication, third-party review—perhaps as part of an as-yet-unfunded DOL evaluation TA effort—seems appropriate. There follows post-publication review by CLEAR.

Thus, if a state took the minimum time at each step, evaluation would run about 33 months. That is the best case. Sample size considerations imply that many states will need to conduct random assignment for two or more years. In addition, relative to the minimum time at each state, moderate schedule slippage is the norm. Thus, most experimental evaluations are likely to publish their final reports four to five years after they start—that is after signing a contract with an evaluator.

C.2. Details of State Evaluations as of Mid-2021

This section attempts to summarize the status of state evaluations as of mid-2021. Many states had planned to start an impact study in 2020; however, 2020 was not a normal year. COVID's spread started in February. UI caseloads spiked. In response, Congress created new UI programs. State management attention was focused on dealing with the twin challenges of spiking caseloads and new programs to be implemented. In net, as of early 2021, no state appeared to have executed a contract for an evaluation.

This situation changed rapidly in the spring of 2021. According to responses to Wave 3 of the RESEA Implementation Study Survey (fielded in March/April 2021), 17 states plan experimental (that is random assignment) designs and 10 other states plan a non-experimental impact study (see Exhibit C-2).

Exhibit C-2. Planned Evaluations as of Wave 3 Survey (March-May 2021)

Current RESEA evaluation plans		
Response option	**Number of states**	**% of states**
Planning a whole program evaluation using random assignment	7	14.0%
Planning a component evaluation using random assignment	6	12.0%
Planning a random assignment evaluation including both whole program and component sub-studies	4	8.0%
Planning an impact evaluation that does not involve random assignment (aka, a quasi-experimental impact evaluation)	10	20.0%
Planning a non-impact evaluation[a] only	11	22.0%
Not sure/nothing	12	24.0%
Total	50	100.0%

Source: RESEA Implementation Study Survey Wave 3, Q6.c.5 and Q6.c.6
[a] Non-impact evaluation types include: Outcomes studies; Process studies; Implementation studies; and Cost studies
Note: Total reflects the number of respondents. The number of selected responses may not sum to the total, because respondents could select more than one response.

Of the states planning experimental studies, seven plan only a whole-program study, six plan component studies,[59] and four plan both component and whole-program. Components to be tested are additional meetings (five states), remote services (four states), job search assistance (three states), individualized services (two states), and selection criteria (one state).[60]

Exhibit C-3. Components to be Evaluated as of Wave 3 Survey (March-May 2021)

Components planned for study by states using random assignment design	
Response option	**Number of states**
Planning a component evaluation[a] using random assignment (*n* = 10)	
Career and labor market information	0
Criteria used to select RESEA claimants (e.g., likelihood of exhaustion)	1
Ways to develop a reemployment plan	0
Job search assistance	3
Approaches to reduce failure to report	0
Penalties for non-compliance/failure to report	0
Providing more individualized career services	2
Adding or removing subsequent RESEA meetings	5
Remote vs. in-person services	4
Other	0

[59] It seems likely that many of these studies will also include a whole-program sub-study.

[60] Some states selected more than one option.

Source: RESEA Implementation Study Survey Wave 3, Q6.c.5 and Q6.c.6.
[a] States considered to be planning an evaluation of RESEA components if they indicated that they would be evaluating the following aspects: Career and labor market information; Criteria used to select RESEA claimants (e.g., likelihood of exhaustion); Ways to develop a reemployment plan; Job search assistance; Approaches to reduce failure to report; Penalties for non-compliance/failure to report; Providing more individualized career services; Adding or removing subsequent RESEA meetings; Remote vs in-person services; Other (please specify).
Note: Total reflects the number of respondents. The number of selected responses may not sum to the total, because respondents could select more than one response.

As of the Wave 3 survey, those planned experimental evaluations were at various stages (see Exhibit C-4). Most were still in the planning stage. Few had selected an evaluator and even fewer had begun data analysis. Additional discussions with states since the Wave 3 interview suggest steady, but slow, progress through July 2021. A few more states have completed statements of work and fewer have selected an evaluator.

Exhibit C-4. Status of Random Assignment Evaluations as of Wave 3 Survey (March-May 2021)

Current progress on RESEA evaluation plans	
Response option	**Number of states**
Planning a whole program evaluation using random assignment (*n* = 11)	
Still in the planning stages (no Statement of Work has been developed)	1
Statement of Work (or similar planning document) is complete but do not yet have an evaluator	5
Have an evaluator but have not begun data collection	4
Have an evaluator and have begun data collection	1
Planning a component evaluation[a] using random assignment (*n* = 10)	
Still in the planning stages (no Statement of Work has been developed)	6
Statement of Work (or similar planning document) is complete but do not yet have an evaluator	1
Have an evaluator, but have not begun data collection	2
Have an evaluator and have begun data collection	1

Source: RESEA Implementation Study Survey Wave 3, Q6.c.5, Q6.c.6, and Q6.c.2.
[a] States considered to be planning an evaluation of RESEA components if they indicated that they would be evaluating the following aspects: Career and labor market information; Criteria used to select RESEA claimants (e.g., likelihood of exhaustion); Ways to develop a reemployment plan; Job search assistance; Approaches to reduce failure to report; Penalties for non-compliance/failure to report; Providing more individualized career services; Adding or removing subsequent RESEA meetings; Remote vs in-person services; Other (please specify).

C.3. Timeline to CLEAR Determination about a Generic RESEA Program

States are required to provide evidence that their RESEA programs are evidence-based. This section considers when states are likely to be able to rely on a statement from CLEAR that a generic RESEA program is effective; that is, has clearly positive impacts on UI weeks and Q2 employment. As noted in Section 2.1, that determination is likely to be based on experimental studies. This discussion therefore focuses on those studies.

Review of Wave 3 survey results, state FY2021 RESEA Plans, and specific state Request for Proposals suggests the following.

- At least six states (and as many as 17) will start experimental evaluations (i.e., finalize contracts) for random assignment impact evaluations during 2021. A few of those evaluations will start in the first half of the year, most in the second half of the year.
- A few of those evaluations will include only one year of random assignment; most will include two or more years of random assignment. Thus, for this first round of evaluations, a few final reports might be publicly released as early as 2024, but that seems unlikely. More of those final reports will be publicly released in 2025, 2026, and 2027.
- Most, but not all, of the evaluators appear to have considerable experience. With moderate levels of evaluation TA, most of these experimental evaluations seem likely to yield results that will meet CLEAR standards for evidence quality.
- Concerns about sample size remain. It seems likely that most—but not all—whole program studies will detect favorable impacts on both outcomes. It is far less clear whether any studies of individual program components will detect such impacts. Most states that are planning such studies have samples that seem likely to be too small to detect impacts on the statutory outcomes.

What does this imply for when states can use generic evidence?

- Perhaps one or two final reports will appear in mid- to late-2024. More will appear in 2025, 2026, and 2027.
- Perhaps two-thirds of the completed experimental studies will meet quality standards and find impacts on both outcomes.
- CLEAR needs a few months to review the studies and generate a summary document.
- States need the CLEAR determination by the fall in order to write their RESEA Plans for the following year.

These considerations suggest that there is a small chance that states will be able to cite generic RESEA evidence in late 2024 for their FY2025 RESEA Plans.[61] It seems possible, but not likely, that states will be able to do so for their FY2026 RESEA Plans. FY2027 seem likely, but not certain. FY2028 seems almost certain.

C.4. Moving Up State Use of RESEA Evidence

Given the annual state RESEA Plan cycle, timing matters. In some cases, a delay of a month or two in CLEAR deeming a generic RESEA program effective will push back this timeline by a full RESEA Plan year. There appear to several steps that DOL could take that would move the date of determination that a generic RESEA program is effective up by a month or two.

---

[61] To see this, consider the nearly best case. Evaluation starts July 2021, with one year of random assignment, and the final report appears 36 months later in July 2024. In that case, CLEAR might be able to complete its review by the end of CY2024, but probably too late even for this state's FY2025 RESEA Plans.

For another state to use generic RESEA evidence, CLEAR would need to find two studies that meet standards and find favorable evidence for both outcomes. This seems unlikely in time for state FY2025 RESEA Plans.

Furthermore, most one year of random assignment studies will likely need more than 36 months and most studies will conduct more than one year of random assignment. Thus, the summary statements in the body of the document

1. DOL CLEAR could speed up its process. Specific steps might include: (1) Instructing DOL CLEAR to give priority to this task. (2) Asking states to share early—i.e., pre-final and pre-publicly released—drafts of reports. This might allow DOL CLEAR to have completed reviews as of when the reports are publicly released or shortly thereafter. (3) DOL CLEAR might pre-write its summary statement about evidence for RESEA, so that it could be publicly released shortly after the second study shows effectiveness.

2. DOL might rethink its strategy for state RESEA Plans. Specific steps might include: (1) delaying the due date for state RESEA Plans; and (2) allowing states to fill in evidence after initial submission of the plan.

Note, however, that these strategies with respect to CLEAR and state RESEA Plans only matter if timing is "close". They might make up a few months (perhaps three to five), but not a lot more.

Appendix D: Non-Experimental Designs

Although in some contexts, non-experimental designs are frequently used to estimate impact, four considerations lead this project to focus on experimental (i.e., random assignment) designs:

1. Experimental designs more plausibly estimate true impact. Random assignment guarantees that there are no systematic differences between those claimants selected and those not selected (among those randomly assigned). In the RESEA context in particular, there likely are systematic differences between those selected and non-selected in observational data (i.e., when who is selected is not determined randomly).

2. The pre-REA and REA evidence (see Appendix A) clearly shows that experimental designs are feasible. Experience implementing random assignment suggests that doing so it not more than minimally burdensome. The changes occur in a centralized computer system and are blind to local offices and those conducting the RESEA meetings.

3. Experimental designs have smaller appropriate sample sizes. Given that sample size is a major challenge for addressing many causal questions of interest to RESEA programs, smaller appropriate sample sizes are a key advantage.

4. Evaluations of reemployment programs have been much more likely to satisfy CLEAR's standards if they use experimental designs than if they use non-experimental designs.

Focusing on the last point, this section enumerates the non-experimental designs for which CLEAR or CLEAR-like reviews have standards and considers their feasibility in the RESEA context. This section draws heavily on the *CLEAR Causal Evidence Guidelines* (2015). The *RESEA Evaluation Toolkit* (Mills De La Rosa et al., 2021)[62] provides additional discussion of these issues.

D.1. Prospective vs. Retrospective Designs

Impact studies can be either *prospective* or *retrospective.* Those two types of impact studies are distinguished as follows:

- **Prospective designs** use outcome data that will be collected in the future as part of the evaluation. The evaluator designs the study and then conducts the study, including enrolling and providing services to participants. Finally, outcomes occur and data are collected on those outcomes.
- **Retrospective designs** use data that have already been collected on participants' program participation and outcomes.

Experimental studies are necessarily prospective. Non-experimental designs can be either prospective or retrospective.

The main ***advantage of retrospective designs*** is that—when the right conditions were in place in the past—data for a sufficiently large sample may already be available, potentially for multiple years of claimants. In turn, the major potential advantage of non-experimental designs is that the studies can, in theory, be retrospective, and thus more quickly completed. A retrospective non-experimental design can

[62] A link to the *RESEA Evaluation Toolkit* can be found on WorkforceGPS's "RESEA Evaluation and Evidence Resources" page. URL: https://rc.workforcegps.org/resources/2019/07/30/17/32/RESEA_Evaluation_Evidence_Resources.

yield a report in less than two years, perhaps within one year. In contrast, prospective designs studying Q2 employment will usually have timelines of four or more years from start to finish.[63]

But ***retrospective designs also have several disadvantages***. Because the data were not collected based on the needs of any particular study, available data are often not exactly what an evaluation would want. For example, the data might not contain important participant characteristics relevant for demonstrating similarity between the intervention group and comparison group. Retrospective designs also limit the study to testing interventions that have been (relatively broadly) applied in the past. If a state were interested in evaluating a new program component—or a whole program that includes an important new component—a retrospective design would be impossible. Finally, a retrospective design provides less control over how the program is implemented and for whom.

The discussion below suggests that, in most cases, states likely cannot answer impact questions of interest using retrospective methods. Further, if a prospective design is preferred, then for reasons discussed above, experimental designs will usually be preferable to non-experimental designs in the RESEA context.

We also discuss one design that is inherently prospective as CLEAR defines it—interrupted time series (ITS). We include it because it is the one non-experimental design that is capable of qualifying for a High study rating from CLEAR.

D.2. Designs That Can Achieve a High Causal Evidence Rating

CLEAR will consider giving a High causal evidence rating to an experimental design (i.e., random assignment; what CLEAR calls an "RCT/Randomized Controlled Trial") and to an ITS design. CLEAR has indicated that once it develops standards for regression discontinuity designs (RD), those designs may be eligible to receive a High causal evidence rating. Experimental designs were discussed in Appendix B; this section discusses ITS and RD.

Interrupted Time Series. ITS is a design in which impacts are estimated by comparing pre- and post-intervention outcomes for the same unit. The analysis can be done at the individual-level or at the group-level. For ITS, CLEAR requires that the decision of which units got the intervention when must not be related to any characteristic of the unit itself (e.g., levels or trends of some outcome). The decision of how to time the rollout must also be chosen by the researcher. Thus, for CLEAR, ITS is an inherently prospective design.[64]

Precision considerations imply that the number of claimants in the offices for which we have pre- and post-intervention outcomes should be in the tens of thousands. There do not appear to be any state that meets those criteria. The RESEA program operates nearly statewide in most states in which it operates. Furthermore, at the start of RESEA, it simply replaced REA. Over the last few years, some states have

---

[63] A prospective study estimating impacts on intermediate outcomes (e.g., RESEA meeting attendance) could be completed much more quickly. This is because the outcomes occur more quickly and the required sample sizes are smaller.

[64] Elsewhere in the research literature, authors might use the term *interrupted time series* to refer to a class of designs that use staggered timing of intervention implementation to estimate the intervention's impact, irrespective of whether or not the timing of the intervention was controlled by the researcher. In fact, most commonly, researchers *retrospectively* use timing that just happened to have been staggered (e.g., using differences among states in when they increased their minimum wage in order to estimate the impact of higher minimum wages on employment). As is discussed below, CLEAR refers to this kind of retrospective design as "difference-in-differences" or "fixed effects," not as ITS.

expanded REA/RESEA to additional offices, but those offices tend to be small. Any estimates based on that rollout would therefore likely be too imprecise to be useful.

A state might roll out some component office by office. If so, ITS might be feasible. But if that is the plan, *group random assignment*—that is, randomly choosing the timing with which an office gets the component—seems strongly preferable. As noted at the end of Section 2.1, group random assignment usually has sample size requirements several times as large as individual-level random assignment. These sample size considerations seem to make both ITS and group random assignment, in practice, infeasible for all but the largest states.

Regression Discontinuity. RD is feasible when who is selected for RESEA is determined by whether the claimant scored above or below the cutoff value on a measure. For RESEA evaluations, this is a promising design because many states select for RESEA based on whether a claimant scores above a particular value on a profiling score. **Option 2.1-2/Regression Discontinuity** considers this for **RQ1/Whole Programs**.

RD compares outcomes for those just on either side of the cutoff. For instance, in an RESEA evaluation, if a state set its profiling score selection threshold at 60, the evaluation would compare employment rates of claimants with scores just above 60 versus employment rates of claimants with scores just below 60 [65] Those claimants should be similar except that some were selected for RESEA and others were not. Said differently, RD looks for a sudden jump in outcomes (a "discontinuity") at the selection cutoff, as illustrated in Exhibit 2-2.

In practice, RD is likely to be feasible only for whole-program evaluations. RD would be feasible for components if a state used a profiling score or some other cutoff value to select some UI claimants a version of the program that included some component (e.g., intensive services) and others for a version of RESEA without that component. We are unaware of any state selecting for a component in that way. As with ITS, a state considering doing so prospectively would likely find random assignment a preferable option (**Option 3.2-1**).

In addition, in practice, sample size considerations imply that RD is unlikely to be feasible for any but the largest states. Formal analyses suggest that RD requires sample four or more times larger than does random assignment and all of the claimants should be "near" the cutoff (Deke & Dragoset 2012). Given that samples sizes for random assignment are challenging, sample sizes for RD are likely infeasible.

D.3. Designs That Can Achieve Only a Moderate Causal Evidence Rating

CLEAR will give a Moderate, but not High, causal evidence rating to an evaluation using several other non-experimental designs. This section considers some of those designs.

Matched Comparison Group and Other Regression Methods (including matching). In the RESEA context, these designs would compare outcomes for claimants who are selected for RESEA to outcomes

[65] Although the intuition easy to understand through a visualization (see Exhibit 2-2), the design's implementation is more complicated. If there were a very large number of claimants (tens of thousands) with scores immediately to either side of the cutoff, then the analyses would be relatively straightforward. In practice, RD designs become more complex because, for sample size reasons, analyses must use claimants with scores somewhat further from the cutoff (see the "band sample" in Exhibit 2-2). Using observations far from the cutoff introduces subtle decisions about which observes should be included (i.e., how far from the cutoff). In addition, using observations further from the cutoff requires the evaluator to be much more careful about how differences in profiling scores are controlled for in the analyses.

for claimants not selected for RESEA. To receive even a Moderate causal evidence rating, the study must show that the two groups are equivalent in their observable characteristics at baseline.

This is usually impossible for RESEA. For whole-program evaluations, if the state uses a profiling score to select among claimants who are RESEA-eligible, then those selected will have much higher profiling scores than those not selected. If claimants not selected are drawn from those who are not RESEA-eligible, then whatever made the claimants not RESEA-eligible (e.g., union hiring hall, out-of-state claim, remote office) will also imply that the two groups are not equivalent.

Similarly, for components, claimants who are assigned to a component in a non-experimental setting are likely to be very different from those who are not. Consider the likely reasons that someone would be assigned to a component: they might have major employment barriers or a past employment history that suggests greater risk to failing to find work. Or consider the likely reasons that an RESEA participant might decide to not use services offered to them: practical challenges (like transportation or health), less motivation, or they are confident that they already have promising job options identified. Because claimants who use a component differ in fundamental (and unobserved ways) from those who do not, it is not possible to create a valid comparison group.

Difference-in-Differences. These designs require multiple outcomes over time for the same claimant or group. Multiple observations for the same claimant would imply multiple UI spells over several years. Multiple observations for a group might imply observing outcomes for the same office over time. **Fixed effects** models explore how outcomes vary as an individual or group switches—from no RESEA to RESEA, from without the component to with the component, or from with the component to without the component. Difference-in-difference models also have observations—in the same time periods—for individuals or groups that did not experience a change. As with other non-experimental designs, CLEAR requires equivalence of the groups before the intervention (in levels and in trends).

For evaluating components, difference-in-differences at the state level would explore how outcomes change as a state adopts a component. This is **Option 3.2-4/Cross-State Interrupted Time Series of Observational Data**. There are two major limitations on the feasibility of this strategy. First, multiple states (probably a dozen or more) need to all adopt the same component. Second, there would need to be a record of when components were adopted by each state. **Option 3.3-4/Annual Survey of State Program Characteristics** would collect that information annually.

Instrumental Variables (IV). This design exploits some random factor affecting who was assigned to each group (e.g., pseudo-random assignment of cases to judges, or perhaps claimants to RESEA caseworkers).

Applying IV to RESEA faces two major challenges. First, there must be something random that has a strong effect on selection. We are unaware of a candidate for such a random factor. Second, appropriate sample sizes vary with the strength of the effect on selection. Even for a strong relation, appropriate sample sizes are several times those for experimental designs. Given that sample sizes are a challenge even for experimental designs, even when such a random factor is available, IV is likely to be infeasible.

References

Black, D. A., Smith, J. A., Berger, M. C., & Noel, B. J. (2003). Is the threat of reemployment services more effective than the services themselves? Evidence from random assignment in the UI system. American economic review, 93(4), 1313-1327.

Bloom, H. S. (2003). Using "short" interrupted time-series analysis to measure the impacts of whole-school reform: With applications to a study of accelerated schools. *Evaluation Review,* 27(1), 3-49. https://doi.org/10.1177/0193841X02239017

Bloom, H. S., Hill, C. J., & Riccio, J. A. (2003). Linking program implementation and effectiveness: Lessons from a pooled sample of welfare-to-work experiments. *Journal of Policy Analysis and Management, 22*(4), 551-575.

Boardman, A. E., Greenberg, D. H., Vining, A. R., & Weimer, D. L. (2017). *Cost-benefit analysis: concepts and practice.* Cambridge University Press.

CLEAR [Clearinghouse for Labor Evaluation and Research]. (2015). *CLEAR Causal Evidence Guidelines, Version 2.1.* U.S. Department of Labor. https://clear.dol.gov/sites/default/files/CLEAR_EvidenceGuidelines_V2.1.pdf

CLEAR [Clearinghouse for Labor Evaluation and Research]. (2018). *Research synthesis: What do we know about the effectiveness of reemployment initiatives?* U.S. Department of Labor. https://clear.dol.gov/sites/default/files/ResearchSynthesis_Reemployment_0.pdf

Darling, M., O'Leary, C. J., Perez-Johnson, I. L., Lefkowitz, J., Kline, K. J., Damerow, B., … & Chojnacki, G. (2017). Using Behavioral Insights to Improve Take-Up of a Reemployment Program: Trial Design and Findings. https://www.mathematica.org/download-media?MediaItemId={9A0114B8-21CC-4A26-BB1B-4D22CDF0963D}

Deke, J., and Dragoset., L. (2012). Statistical power for regression discontinuity designs in education: Empirical estimates of design effects relative to randomized controlled trials [Working paper]. Mathematica Policy Research, Inc. https://files.eric.ed.gov/fulltext/ED533141.pdf

Epstein, D., & Klerman, J. A. (2012). When is a program ready for rigorous impact evaluation? The role of a falsifiable logic model. *Evaluation Review*, 36(5), 375-401.

Epstein, Z., Klerman, J. A., Clarkwest, A., and D. Nightingale, D. (2022). State of the Evidence Report: *Evaluation to Advance RESEA Program Evidence.* Rockville, MD: Abt Associates.

Fortson, Kenneth, Rotz, Dana, Burkander, Paul, Mastri, Annalisa, Schochet, Peter, Rosenberg, Linda, McConnell, Sheena, & D'Amico, Ronald. (2017). *Providing public workforce services to job seekers: 30-month impact findings on the WIA Adult and Dislocated Worker programs.* Mathematica Policy Research.

Greenberg, D. H., & Appenzeller, U. (1998). *Cost analysis step by step: A how-to guide for planners and providers of welfare-to-work and other employment and training programs.* Connections to Work.

Greenberg, D. H., Michalopoulos, C., & Robins, P. K. (2003). A meta-analysis of government-sponsored training programs. *ILR Review, 57*(1), 31-53.

Klerman, J. A. (2017). Editor in chief's comment: External validity in systematic reviews. *Evaluation Review, 41*(5), 391-402. https://www.doi.org/10.1177/0193841X17746740.

Klerman, J. A., & Danielson, C. (2011). The transformation of the Supplemental Nutrition Assistance Program. *Journal of Policy Analysis and Management, 30*(4), 863-888. https://onlinelibrary.wiley.com/doi/abs/10.1002/pam.20601

Klerman, J. A., Saunders, C., Dastrup, E., Epstein, Z., Walton, D., and Adam, T., with Barnow, B. S. (2019). Evaluation of impacts of the Reemployment and Eligibility Assessment (REA) Program: Final report. Prepared for the U.S. Department of Labor. Cambridge, MA: Abt Associates.

Mastri, A., & McCutcheon, A. (2015). *Costs of services provided by the WIA Adult and Dislocated Worker programs*. U.S. Department of Labor.

Mayer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics, 13*(2), 151-161. https://www.jstor.org/stable/1392369?seq=1

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics, 142*(2), 698-714.

Michaelides, M., & Mueser, P. (2016). *The labor market effects of U.S. reemployment programs during the Great Recession* (Working Paper No. 08-2015). Nicosia, Cyprus: University of Cyprus, Department of Economics.

Mills De La Rosa, S., Souvanna, P., Clarkwest, A., Kappil, T., Epstein, Z., Rothschild, L., Kuehn, D., Wall, A., Klerman, J. A., & Nightingale, D. (2021). *Reemployment Services and Eligibility Assessment (RESEA) evaluation toolkit: Key elements for state RESEA programs.* U.S. Department of Labor. https://rc.workforcegps.org/resources/2019/07/30/17/32/-/media/0CA5268944284F3DB41E6DC92A313C2E.ashx

Minzer, A., Klerman, J., Epstein, Z., Savidge-Wilkins, G., Benson, V., Saunders, C., Cristobal, C., and Mills, S. (2017). *REA Impact Study: Implementation Report.* Abt Associates. https://www.abtassociates.com/sites/default/files/migrated_files/055f9b46-7166-460e-90c5-242e70fa8a6e.pdf

Molina, I., & Rao, J. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics, 38*(3), 369-385.

Poe-Yamagata, E., Benus, J., Bill, N., Carrington, H., Michaelides, M., & Shen, T. (2011, June). *Impact of the Reemployment and Eligibility Assessment (REA) initiative*. IMPAQ International. https://www.impaqint.com/sites/default/files/files/ETAOP_2012_08_Impact_of_the_REA_Initiative.pdf

Rao, J. N. K. (2014, November 27). Small area estimation. *Wiley Stats. Ref: Statistics Reference Online*. John Wiley & Sons. https://doi.org/10.1002/9781118445112.stat03310

Saunders, C., Dastrup, E., Epstein, Z., Walton, D., Adam, T., Klerman, J. A., & Barnow, B. S. (2019). *Evaluation of Impacts of the Reemployment and Eligibility Assessment (REA) Program: Final Report*. https://www.dol.gov/sites/dolgov/files/OASP/evaluation/pdf/REA%20Impact%20Study%20-%20Final%20Report.pdf

Schaberg, K., & Greenberg, D. H. (2020). *Long-term effects of a sectoral advancement strategy: Costs, benefits, and impacts from the WorkAdvance demonstration.* MDRC.

Toohey, D., (2017). *The effectiveness of work-search requirements over the business cycle: Evidence for job rationing*. https://sites.google.com/site/desmondtoohey/research.

Trutko, J., Trutko, A., Clarkwest, A., Souvanna, P., Klerman, J. A., Spaulding, S., Briggs, A., Scott, M., Hecker, I., Islam, A., & Katz, B. (2022). *RESEA Program Strategies: State and Local Implementation*. Report submitted to U.S. Department of Labor, Chief Evaluation Office. Rockville, MD: Abt Associates.

USDOL [U.S. Department of Labor]. (2019). *Expectations for states implementing the Reemployment Service and Eligibility Assessment (RESEA) program requirements for conducting evaluations and building program evidence.* Unemployment Insurance Program Letter No. 1-20. Employment and Training Administration, U.S. Department of Labor. https://wdr.doleta.gov/directives/attach/UIPL/UIPL_1-20.pdf

USDOL [U.S. Department of Labor]. (2021). *Fiscal Year (FY) 2021 Funding allotments and operating guidance for Unemployment Insurance (UI) Reemployment Services and Eligibility Assessments (RESEA) grants.* Unemployment Insurance Program Letter No. 13-21. Employment and Training Administration, U.S. Department of Labor. https://wdr.doleta.gov/directives/attach/UIPL/UIPL_13-21.pdf

Valentine, J. C., Wilson, S. J., Rindskopf, D., Lau, T. S., Tanner-Smith, E. E., Yeide, M., & Foster, L. (2017). Synthesizing evidence in public policy contexts: the challenge of synthesis when there are only a few studies. *Evaluation review*, 41(1), 3-26.

Vollmer, L., Mastri, A., Maccarone, A., & Sama-Miller, E. (2017). *The Right Tool for the Job: A Meta-Regression of Employment Strategies' Effects on Different Outcomes*. Mathematica Policy Research. https://ideas.repec.org/p/mpr/mprres/c721bdfdff0e4e78bf45fd939d0c6376.html

Weiss, Michael J., Mayer, A. K., Cullinan, D., Ratledge, A., Sommo, C., & Diamond, J. (2015). A random assignment evaluation of learning communities at Kingsborough Community College—Seven years later. *Journal of Research on Educational Effectiveness 8*(2), 189-217.