



Common Pitfalls in Designing RESEA Causal Evaluations

Andrew Clarkwest and Victoria Wheat (Abt Associates)

April 2023

Causal evaluation designs aim to estimate the impact *caused by* an intervention. Causal evaluations of the Reemployment Services and Eligibility Assessment (RESEA) program can produce evidence of causal impact to help states understand how effective their programs are overall or to determine the effectiveness of different program approaches. Evidence produced from evaluations using causal designs are also the basis for the U.S. Department of Labor’s (DOL’s) determinations of whether “interventions and service delivery strategies [qualify for] a ‘high’ or ‘moderate’ causal evidence rating.” Specifically, the evaluation’s design and conduct have produced evidence such that audiences can be confident the program is effective in improving participants’ employment outcomes and reducing their Unemployment Compensation (UC) duration.

This document addresses Frequently Asked Questions (FAQs) about designing causal evaluations. It aims to provide states with critical information to design RESEA impact evaluations that (1) produce credible evidence of the impact of their RESEA interventions and (2) demonstrate to DOL that interventions they are using are effective. The first is important for informing states’ decisions about program design. The second is important to qualify for federal RESEA program funds. Specifically, this FAQ addresses a set of design choices and issues that have arisen in proposed evaluation designs to date. States that do not avoid these potential “pitfalls” could come to the end of an evaluation and find that the evidence produced does not meet those two needs.

DOL funded this FAQ resource as part of a suite of evaluation technical assistance resources to states conducting evaluations of their RESEA programs. The resources, including the [RESEA Evaluation Toolkit](#), can be found online on the [WorkforceGPS RESEA Evaluation and Evidence Resources](#) page. [The RESEA Evaluation and Evidence FAQ](#) and [Evaluation Design Webinar](#) provide tips on conducting impact evaluations and other types of evaluations.

Question #1: How does DOL rate studies’ credibility?

DOL’s Clearinghouse for Labor Evaluation and Research (CLEAR) reviews studies on a wide range of labor-related interventions, including studies of reemployment interventions such as RESEA. CLEAR aims to bring labor-related research together in one place and help users make sense of that evidence. CLEAR’s system of **causal intervention ratings for studies** is one tool for doing that. Those study ratings indicate the credibility of a study’s findings. The three possible ratings are “high,” “moderate,” and “low,” indicated in CLEAR by gas gauge icons (see Box 1 below).

This document 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.

CLEAR reviews studies using the [CLEAR Causal Evidence Guidelines](#), which apply specific criteria to determine credibility and provide ratings. Studies that do not meet the guidelines specific to the type of evaluation design receive lower ratings. This FAQ is intended to help evaluations avoid pitfalls which could cause a study to not meet certain criteria outlined in the Causal Evidence Guidelines. Therefore, specific criteria from the causal evidence guidelines are referenced throughout the document.

When we say “**credible**,” what do we mean? In technical language, it means that the estimates are “internally valid” or “unbiased.” In less technical language, it means that effects estimated by a study truly reflect the intervention’s impacts on the participants who were studied, rather than reflecting something else.¹

To credibly estimate impact, an evaluation requires a comparison group that can provide a “counterfactual.” A counterfactual represents what the intervention participants’ outcomes would have been if they had not had access to the intervention. The difference in outcomes between the participants with access to the intervention versus the participants without access to the intervention is the *impact* of the intervention.

Box 1: Icons for CLEAR’s Causal Evidence Ratings for Studies



High Causal Evidence:
Strong evidence the effects are caused by the examined intervention.



Moderate Causal Evidence:
Evidence that the effects are caused to some degree by the examined intervention.



Low Causal Evidence:
Little evidence that the effects are caused by the examined intervention.

CLEAR’s study ratings assess whether the comparison group provides a valid counterfactual—that is, do the outcomes of the comparison group truly reflect what the outcomes of the intervention group would have been *if not for the intervention*? Or, in less technical terms, does the comparison group provide an apples-to-apples comparison with the intervention group? CLEAR assigns a “high” study rating if it determines that the comparison group provides a credible enough counterfactual that we can be confident the resulting findings reflect the impacts of the intervention and not some other difference between the groups.

Valid counterfactual comparisons can be difficult to create and maintain. With both experimental designs, known as randomized controlled trials (RCT) and quasi-experimental (QED) designs (e.g., a matched comparison group), much can go awry. Many of the questions below address issues that have arisen as states have been developing RESEA impact evaluation designs and that can lead to comparisons that are not credible. Problems with the comparison group that fail to provide valid counterfactuals can occur in QEDs and RCTs. Studies that rely on invalid comparisons are likely to receive “low” study ratings from CLEAR.

An important point to note is that CLEAR causal intervention ratings for *studies* do not indicate anything about *what* the studies found. For instance, a “high” study rating does not mean that the study found positive impacts, only that the analysis meets high standards for research design and execution and that its estimates of causal impacts are credible, whether those impacts were positive, negative, or null.

¹ Particularly, results that are statistically significant. It is important to note that findings that are not statistically significant (often referred to as “null” findings) from high-rated studies do not necessarily imply that the intervention has no impact. If standard errors (also known as “margins of error”) are large, then null findings could indicate that the study’s sample was too small to be able to detect an impact, even if an impact were present.

Question #2: How are CLEAR's study ratings related to RESEA intervention ratings?

In addition to study ratings (see Question #1), CLEAR also has **causal evidence ratings for RESEA interventions**. These too are referred to as “causal evidence ratings,” but unlike *study* ratings, *intervention* ratings **do** indicate what the studies found. These ratings reflect findings from all relevant studies of an intervention, while also considering the credibility of those studies.

Specifically, intervention ratings, indicated by thermometer icons (see Box 2) focus on whether studies of an intervention have found credible evidence that the intervention is effective. Interventions can be rated “high,” “moderate,” “potentially promising,” or “no rating.”

“High” and “moderate” intervention only consider findings from studies that are deemed sufficiently credible. Specifically, those rating categories only consider findings from studies that have received “high” or “moderate” *study* ratings (see Question #1). Findings from a low-rated study can be used to contribute to a “potentially promising” intervention rating. But findings from low-rated studies are not considered in determining whether an intervention qualifies for a “high” or “moderate” *intervention* rating.

Box 2: Icons for CLEAR's Causal Evidence Ratings for RESEA Interventions



High: At least two credible impact studies have found favorable effects on UI duration and employment with strong statistical significance ($p < .05$).



Moderate: At least one credible impact study has found favorable effects on UI duration and employment (can be different studies) with modest statistical significance ($p < .10$).



Potentially Promising: At least one impact study (irrespective of study rating) found favorable impacts on UI duration or employment with modest statistical significance ($p < .10$).



No Rating: No impact studies of the intervention exist yet or studies that do exist do not show favorable impacts; more research is needed.

Question #3: For a whole-program evaluation, can we estimate the program's impact using a quasi-experimental approach that compares outcomes of claimants selected for RESEA versus outcomes of claimants not selected for RESEA?

Likely not. Why not? Consider how claimants are selected for RESEA. Typically, claimants are selected based on profiling scores, which reflect an estimate of how likely they are to exhaust UC benefits. Specifically, RESEA programs typically select claimants with *higher* profiling scores. Because claimants selected for RESEA have higher profiling scores than those not selected, we can expect claimants selected for RESEA to have lower employment rates and longer UC durations than other claimants. Thus, when comparing the outcomes of UC claimants selected for RESEA versus those not selected, average outcomes for the two groups are likely to differ from each another for many reasons other than RESEA itself. These other systematic and consequential differences between the groups mean the comparison is not “apples-to-apples.”

Sometimes advanced statistical techniques—such as matching or regression discontinuity—can create a valid comparison group in situations even if program participants differ from non-participants.² However,

² *Matching* uses an algorithm to find people who did not receive the intervention who are similar to those who did receive it. For RESEA, matching requires that there are some claimants selected for RESEA who are like claimants not selected—in particular, there must be claimants in both groups with the same profiling scores. If selection occurs via a strict score cutoff, this would not be true. But in many states, selection for RESEA is based on staff capacity at local offices, so there can be some overlap in claimants' profiling scores who are or are not selected. But because the claimants with the same profiling scores will

those advanced statistical techniques create a valid comparison group only when stringent conditions are met. In addition, the advanced statistical techniques are complex and require specific expertise, likely require very large samples, and have greater risk of failing to qualify for even “moderate” study ratings from CLEAR.

For reference, see CLEAR’s Regression Model [Criterion #1](#) in the Causal Evidence Guidelines:
Before the intervention, were the intervention and comparison groups similar?

Question #4: For a whole-program evaluation, can we estimate the program’s impact using a quasi-experimental approach that compares outcomes of claimants who attended an RESEA meeting to the outcomes of claimants who were selected for RESEA but never attended a meeting?

No. Why not? Consider how claimants who attend the meeting are likely to be different from those who do not attend. Would the ways that they differ tend to make us expect that they would have different outcomes? Those who do not attend an RESEA meeting likely fall in one of two groups:

- Some claimants might not attend because they found a job or are confident that they will do so shortly, and so do not need assistance from the RESEA program. Thus, compared to typical claimants selected for RESEA, we would expect these non-attendees to have shorter UC durations and higher average employment rates.
- Other claimants might not attend the meeting because they have barriers that prevent them from doing so. For example, they might have challenges attending a meeting because of technology access (if remote), transportation (if in person), or competing obligations. Perhaps they did not understand the requirement due to limited English proficiency. These same barriers also can make it harder for them to become reemployed. Thus, compared to typical claimants selected for RESEA, we would expect these non-attendees to have lower-than-average employment rates.

Importantly, these kinds of differences between attendees and non-attendees probably cannot be captured in the evaluation’s data on claimants’ pre-intervention characteristics. Because those claimant traits are unmeasured, the differences cannot be addressed through statistical adjustments.

For reference, see CLEAR’s Regression Model [Criterion #1](#) in the Causal Evidence Guidelines:
Before the intervention, were the intervention and comparison groups similar?

often come from different offices or have been randomized at different points in time, it will likely prove difficult to create truly comparable RESEA and non-RESEA samples.

If selection is based on a strict profiling score cutoff, then a regression discontinuity approach might be feasible. *Regression discontinuity* designs compare the outcomes of claimants with scores just above the selection cutoff to the outcomes of claimants who have scores just below the cutoff. The two sets of claimants are thus similar other than that some were selected for RESEA and some were not. Regression discontinuity designs can provide highly credible impact estimates but are challenging to implement correctly and require very large samples.

See the [RESEA Evaluation Toolkit](#) for more details.

Question #5: For an evaluation of a component of a program, can we estimate impacts using a quasi-experimental approach that compares the outcomes of claimants selected for RESEA after the new component was implemented to outcomes of claimants selected for RESEA before the component was implemented?

No. Why not? Suppose the state had kept its RESEA program the same and had not implemented the new component. Are we confident that outcomes would have been the same for the claimants who attended RESEA early compared to those who attended it later? Almost certainly not. The outcomes for these two groups could differ for a range of reasons, including changes over time in the economy and UC claimant composition. We would expect UC claimants' employment and claim duration outcomes to differ correspondingly in earlier versus later periods, even without any changes to RESEA.³

For reference, see applicable CLEAR Criteria:

CLEAR's Regression Model [Criterion #1](#) in the Causal Evidence Guidelines: Before the intervention, were the intervention and comparison groups similar?

CLEAR's Regression Model [Criterion #2](#) in the Causal Evidence Guidelines: Were there confounding factors?

Question #6: Suppose that some of our RESEA participants are referred to a particular service (such as a job search workshop) and others are not referred. Can we estimate the impact of that service by using a quasi-experimental approach that compares the outcomes of those who were referred versus the outcomes of those who were not?

Likely not, given the typical reasons for providing referrals. Why not? If the basis of the referral is that some claimants need the service and others do not, then comparing the outcomes of the two groups will not give you a valid estimate of the impact of the service. For example, we might expect the referred group to have a lower reemployment rate than the claimants who were not referred if the reason for the referral is that the referred claimants were perceived to have difficulty finding a job without the service, whereas the claimants not referred were perceived as being likely to find a job without assistance. To create a valid comparison group, referrals would need to be made in such a way that they are unrelated to claimants' characteristics that are related to their outcomes.

For reference, see CLEAR's Regression Model [Criterion #1](#) in the Causal Evidence Guidelines: Before the intervention, were the intervention and comparison groups similar?

³ This design is likely to fail CLEAR's regression model Criterion #1 (because claimant populations are likely to differ over time in ways that may not be entirely captured in available data) but also Criterion #2, which states, "One somewhat common confound is known as the N=1 confound, when all cases from one study arm were in one state, county, local workforce investment area, school, or cohort. In such cases the effect of the intervention cannot be disentangled from the effect of the single unit in which the intervention was delivered." In this case, the intervention and comparison claimants were drawn from entirely different time periods. All intervention claimants were selected from a later period and all comparison claimants were selected from an earlier period. As such, there is no way to tell whether differences in outcomes are because of the new component or because the intervention claimants came from a different time and cohort than the comparison claimants.

Question #7: For an evaluation of a service that is a component of a program, can we estimate impacts using a quasi-experimental approach that compares the outcomes of claimants who were referred to the service and used the service versus those who were referred but did not use the service?

No. Why not? The reasons here are similar to the reasons in Question #2. Claimants who decide to use a service are likely to be fundamentally different from those who did not decide to use it. Further, they are likely to be different in ways that are related to their probability of reemployment.

For reference, see CLEAR’s Regression Model [Criterion #1](#) in the Causal Evidence Guidelines:
Before the intervention, were the intervention and comparison groups similar?

Question #8: Let’s say that our program develops a simple, plain language tool to help claimants more easily follow through on service referrals in their individual reemployment plan. We then randomly assign some claimants to be given that tool. Can we estimate the impact of the tool by comparing employment outcomes of claimants who received the new tool versus employment outcomes of those who received a “business-as-usual” reemployment plan (one without the new tool)?

Yes, but with caveats, depending on the outcomes of interest. Why? Because the evaluation design uses random assignment to determine who gets the tool, it is likely to meet CLEAR standards. However, it is important to think carefully about the *intermediate* outcomes that the new component intends to affect that, in turn, are meant to affect *final* outcomes, and to measure the program’s impact on those intermediate outcomes. This consideration is particularly important for many component evaluations that affect only one part of the RESEA program. In this instance, the key outcome would be for claimants to follow through on referrals. The evaluation results would tell you whether the intervention produced the shorter-term result (using the referred services) that is expected to lead to the longer-term result (employment).

An issue for component evaluations is that though most states are likely to have caseloads (and, in turn, samples sizes) that are large enough to be able to measure intermediate outcomes, detecting impacts on some outcomes—particularly employment outcomes—could require a sample larger than some states can obtain.⁴ To obtain larger sample sizes, a state could randomly assign claimants over a longer period of time or pool its results with results from other states evaluating that same component.

⁴ To illustrate, suppose the new tool increases follow-through on referrals by 10 percentage points. That would be a very good result for this type of relatively low-cost intervention, and you likely could detect that size of impact on uptake of referrals without needing a very large sample. However, referrals are only one part of RESEA and the tool would affect uptake for only 10 percent of claimants. Suppose that following through on referrals and receiving services cuts UC duration by 2 weeks. Even with an impact of that size, the average impact on the sample as a whole is only one-tenth that large (because only one-tenth of the sample was induced to take up services), or 0.2 weeks. That’s a much smaller impact than 2 weeks, and a much smaller impact will, in turn, require a much larger sample to detect.

Question #9: For random assignment in a whole-program evaluation, can we change the percentage of claimants assigned to receive the intervention (treatment group) versus those who do not receive it (control group) if we need to vary assignment rates across local areas or over time?

Yes, but you will need to take account of variation during analysis. Why? Program logistics could make it necessary to vary the assignment rates to the treatment group and control group across different times or places. For example, some American Job Centers (AJCs) might have a capacity to serve a larger proportion of new UC claimants than other AJCs. If so, it might make sense to have a higher proportion of claimants assigned to the treatment group in AJCs that have higher capacity. If you do vary the percentage of claimants assigned to each group, however, the two groups will no longer be similar. The treatment group will have a larger share of claimants from high-capacity AJCs than will the control group.

In this case, we *know* the groups differ, meaning the design does not suffer from the design pitfalls described above. As a result, during analysis your evaluator can make statistical adjustments to correct for these differences in the treatment and control groups. For example, the evaluator might use techniques such as *weighting* (by the inverse of the assignment probability) or “*fixed effects*” (by whatever unit you varied your assignment probabilities, such as by AJC or by time period). For your evaluator to make those adjustments, you will need to record what random assignment ratio you used at the time and place that each claimant was assigned. **An evaluator with strong experience in conducting impact studies will know how to make those statistical adjustments, which can maintain the valid treatment/control comparison and satisfy CLEAR’s standards for a “high” study rating.**

For reference, see CLEAR’s RCT [Criterion #3](#) in the Causal Evidence Guidelines: Did the probability of assignment into treatment or control groups vary over time without appropriate adjustment?

Additional RESEA Resources and Evaluation Technical Assistance

States play a central role in building evidence, so DOL provides evaluation technical assistance to states to support their efforts to conduct high-quality evaluations of RESEA programs. The RESEA Evaluation Technical Assistance Team is available to support states in all aspects of evaluation planning and execution. Let us know how we can help!

RESEA Helpline: RESEA@abtassoc.com

WorkforceGPS evaluation resources:

https://rc.workforcegps.org/resources/2019/07/30/17/32/RESEA_Evaluation_Evidence_Resources

Suggested citation: Clarkwest, A. & Wheat, V. (2023). *Common Pitfalls in Designing RESEA Causal Evaluations*. RESEA Evaluation Technical Assistance Brief. Prepared for the U.S. Department of Labor, Chief Evaluation Office. Abt Associates.