

Workforce Intelligence Lab

Evaluating Workforce Programs

Evaluation Toolkit

A practical framework for measuring delivery and outcomes, and determining when an effect claim is warranted.

PUBLICATION NUMBER

WIL-2026-003

VERSION

1.0

PUBLICATION DATE

24 September 2026

PUBLICATION TYPE

Evaluation Toolkit

ABSTRACT

A workforce program needs an evaluation design that matches the decision it is meant to inform. This toolkit separates delivery monitoring, participant outcomes, and causal impact. It supplies a practical sequence for defining cohorts, specifying measures, checking follow-up, choosing a defensible comparison, and reporting uncertainty. The package contains reusable planning, data-dictionary, outcome-reporting, and analysis-specification templates. It reports no participant outcome estimates or program effects because no program-level outcome analysis has been conducted for this publication.

PUBLICATION METADATA

Publication number	WIL-2026-003
Publication type	Evaluation Toolkit
Components	Planning, data-dictionary, outcome-reporting, and analysis-specification templates
Version	1.0
Publication date	24 September 2026

SUGGESTED CITATION

Workforce Intelligence Lab. (2026). *Evaluating Workforce Programs* (Publication No. WIL-2026-003, Version 1.0). *Evaluation Toolkit*, 24 September 2026.

CONTENTS

1	Start with the Decision and the Program Theory	4
2	Separate Three Forms of Evidence	4
3	Define the Cohort and Observation Unit	4
4	Specify Outcomes and Denominators Before Extraction	5
5	Audit Data Before Estimating Results	5
6	Select a Design That the Data Can Support	6
7	Write the Analysis Specification	6
8	Report Outcomes, Costs, and Decisions Together	7
9	What the Available Sources Can and Cannot Contribute	7
R	References	7
	About This Document	9

1 Start with the Decision and the Program Theory

Specify the decision before selecting indicators: whether to continue delivery, repair an implementation problem, change eligibility, expand a service, or commission an impact study. Name the decision owner, the implementation window, the eligible population, the service offer, and the alternative available to participants. A training provider's completion target and a funder's question about additional earnings are different evaluation questions.

Describe the pathway from resources to activities, outputs, intermediate outcomes, and longer-term outcomes. For each link, list the assumption that must hold. For example, completing training may improve credentials, but employment may also depend on transportation, employer demand, licensing, care responsibilities, and access to vacancies. These are hypotheses to test, not results of this toolkit.

Preserve any submitted grant protocol, funder reporting definitions, and registered analysis plan. Maintain a crosswalk between required reporting and the additional research questions. If a new design requires a change, record it through the applicable amendment process before analysis. This toolkit is an optional operational aid; it does not revise existing grant commitments.

2 Separate Three Forms of Evidence

Delivery monitoring asks whether the service reached its intended population and was delivered as planned. Measures may include eligibility processing, service receipt, attendance, completion, and referral follow-through. The appropriate denominator depends on the process: all applicants, eligible applicants, enrolled participants, or those scheduled to receive a service.

Outcome analysis asks what happened to a defined cohort over a specified period. Employment, earnings, credential attainment, job quality, and retention require operational definitions and data capable of observing the outcome. A favorable participant outcome is descriptive until a credible counterfactual addresses what would have happened without the program.

Impact evaluation asks whether an intervention changed outcomes relative to that counterfactual. Neither a before-and-after increase nor comparison with the regional unemployment rate establishes an effect. Labor-market changes, eligibility, participant selection, and incomplete follow-up can all alter observed outcomes. Label conclusions according to the design actually executed.

3 Define the Cohort and Observation Unit

The cohort specification should record eligibility, entry dates, assignment or offer dates where relevant, enrollment, participation, exit, and the follow-up schedule. Define repeat episodes and transfers before deduplicating. Preserve a secure person-to-episode crosswalk so that the analysis can distinguish a unique individual from multiple program spells.

Use an individual as the unit for a person-level outcome and an episode when the research question concerns service episodes. Do not count both without explaining the estimand. For cluster-level assignment or delivery, retain the cluster identifier and analyze uncertainty at the appropriate level. Assignment, enrollment, receipt, and completion are distinct states; collapsing them can change the question being answered.

Store direct identifiers separately from analytical records. Obtain the applicable authority and data-use agreements before linking administrative data. The templates specify fields and roles, but they do not authorize access or set a universal disclosure threshold. The data steward must approve retention, access, small-cell treatment, and release rules for the actual data.

4 Specify Outcomes and Denominators Before Extraction

For each outcome, record the concept, observation window, numerator, denominator, unit, source, linkage rules, exclusions, missingness codes, and reporting lag. Align windows across comparison groups. If a funder requires statutory or administrative performance measures, use the current governing definitions and keep a versioned reference [D1]; do not substitute a convenient local measure with a similar name.

An observed employment rate is the number observed employed divided by the number with ascertainable employment status in the defined cohort and window. Report that denominator alongside the full eligible cohort and the number whose status is unknown. An alternative all-cohort rate is valid only when its treatment of unknown status is explicit; unknown is not evidence of nonemployment.

Report earnings for the specified population and period. Earnings conditional on employment and earnings across the full cohort answer different questions. Use zero only when the source and coverage rules support a genuine observed zero; an absent match or missing record is not automatically zero. Specify whether amounts are nominal or inflation-adjusted and retain the deflator if used.

Credential and retention measures require equally explicit definitions. Identify qualifying credentials, the qualifying population, attainment dates, and observation windows. Define whether retention means continued work for the same employer, continued employment anywhere covered by the source, or repeated positive earnings. These measures must not be interchanged.

5 Audit Data Before Estimating Results

Freeze a dated extract and preserve its query, schema, row counts, and access-controlled lineage. Check unique keys, valid dates, temporal ordering, duplicates, missingness, linkage rates, and source coverage. Confirm that follow-up has matured for the entire reported cohort; a recent cohort may appear to perform worse simply because its outcomes are not yet observable.

Produce a cohort flow showing eligible, offered, enrolled, receiving service, completed, linked, and observed at follow-up. Record reasons for exclusion and loss to follow-up. Reconcile totals at each step without assuming that the states are interchangeable or strictly nested when the program permits re-entry.

Compare follow-up and linkage completeness across groups and relevant baseline characteristics. Administrative wage records may omit forms of work or jurisdictions outside the source's coverage. A no-match flag should remain distinct from observed nonemployment. Survey follow-up requires response-rate and nonresponse analysis. If coverage differs materially, narrow the question or report the limitation before comparing outcomes.

6 Select a Design That the Data Can Support

Random assignment can support an intention-to-treat estimate of the effect of an offer when assignment is implemented and recorded correctly. Preserve assignment regardless of participation, investigate attrition and contamination, and document any cluster design. An effect among participants requires additional assumptions; it is not obtained by simply discarding nonparticipants.

A comparison design requires a defensible account of how comparison units approximate the untreated outcomes of the treated group. Matching or regression adjustment addresses measured baseline differences under assumptions; it does not establish that unmeasured selection is absent. Check overlap and balance and avoid extrapolating to groups without support.

Difference-in-differences compares changes across groups and relies on a credible parallel-trends assumption, alongside appropriate treatment timing and no unaddressed concurrent shocks. Examine pre-treatment patterns where available, document anticipation and spillover risks, and use an estimator appropriate to the timing structure. A regression discontinuity design requires a real assignment threshold, credible continuity assumptions, and checks for manipulation. Do not choose a design simply because a software routine is available.

Where no credible comparison is available, publish an implementation or descriptive outcome study. That is useful evidence when its scope is clear. A recommendation to obtain stronger evidence may be more defensible than an unsupported effect estimate [D2].

7 Write the Analysis Specification

The reusable specification covers the research question, estimand, primary outcome, population, intervention, comparison, time window, adjustment set, inference method, missing-data handling, and sensitivity checks. Specify the primary analysis before reviewing treatment differences. Keep exploratory analyses clearly labeled and retain dated amendments.

Plan sample-size and precision calculations from the actual assignment unit, plausible baseline outcome variation, expected participation, clustering, and follow-up loss. This publication supplies no invented minimum sample size or power claim. If the available data cannot support decision-relevant precision, narrow the question or redesign the evaluation.

For a descriptive rate, show numerator and denominator. For a between-group difference, show the component group estimates and their uncertainty. For an effect estimate, disclose the model, inference method, assumptions, and sensitivity to reasonable alternatives. Multiple outcomes and subgroup analyses require a stated approach to multiplicity and interpretation. Avoid presenting a selected favorable result as the program's overall effect.

8 Report Outcomes, Costs, and Decisions Together

The report template separates implementation findings, descriptive outcomes, impact estimates if justified, and limitations. Each finding should link to a reproducible table and an explicit decision implication. Missing outcomes, incomplete follow-up, and unsupported counterfactuals belong in the main interpretation, not only in a technical appendix.

Cost per participant and cost per completer are accounting ratios with defined cost scope and denominators. Incremental cost-effectiveness requires an incremental outcome estimate and a defensible comparator. Do not divide program cost by a descriptive outcome difference and label the result a causal return on investment.

Use the decision worksheet to record the evidence, alternative explanations, practical response, owner, and next review trigger. An implementation failure may call for service repair; an imprecise estimate may call for more observation; a credible effect may support a bounded expansion decision. None of those conclusions follows from this toolkit alone.

9 What the Available Sources Can and Cannot Contribute

QCEW and OEWS describe labor-market conditions, and ACS and LODES can help characterize population or commuting context within their source definitions [B2, B3, C1, C2]. They do not identify program participation or establish participant outcomes by themselves. Posting data measures advertised opportunities rather than realized hires. Sponsor or occupation registers do not measure apprentice starts, completions, or causal training effects.

No linked participant cohort, matured outcome file, comparison-group analysis, or impact estimate is included here. The toolkit therefore makes no claim that any program improved employment, earnings, completion, or retention. Source availability is a feasibility input, not evidence of effectiveness.

R References

[D1] U.S. Department of Labor. WIOA performance accountability. <https://www.dol.gov/agencies/eta/performance/performance-indicators>

[D2] U.S. Department of Labor, CLEAR. Causal evidence guidelines. <https://beta.dol.gov/research-data/clear/about/review-rating-guidelines/causal-evidence-guidelines-version-22>

[B2] U.S. Bureau of Labor Statistics. QCEW overview. <https://www.bls.gov/cew/overview.htm>

[B3] U.S. Bureau of Labor Statistics. OEWS frequently asked questions. https://www.bls.gov/oes/oes_ques.htm

[C1] U.S. Census Bureau. ACS data profiles. <https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/>

[C2] U.S. Census Bureau. LEHD data. <https://lehd.ces.census.gov/data/>

Reference list checked 24 September 2026. Applicable program reporting requirements must be versioned for the evaluation being undertaken. No source listed here supplies an outcome estimate for this publication.

ABOUT THIS DOCUMENT

Published by

Workforce Intelligence Lab
37 N. Orange Avenue, Suite 222, Orlando, Florida 32801
www.WILInstitute.org

The Workforce Intelligence Lab is an independent nonprofit research organization, recognized as tax-exempt under Section 501(c)(3) of the Internal Revenue Code.

© 2026 Workforce Intelligence Lab. This document may be quoted or reproduced with attribution to the Workforce Intelligence Lab.

DISCLOSURES

WIL is organizationally and operationally distinct from AlphaHire, a commercial workforce intelligence company. The two organizations are related — WIL may use appropriately governed research access made available by AlphaHire, and relevant relationships are disclosed in research outputs — but WIL retains sole authority over its research questions, design, analysis, interpretation, and publication, and no affiliated or sponsoring organization determines its findings. Organizations cited as sources are references only and imply no affiliation or endorsement.

DOCUMENT REVISION HISTORY

VERSION	DATE	SUMMARY OF CHANGES
1.0	September 2026	Initial publication.