Title IV-E Prevention Services CLEARINGHOUSE

OPRE

Abt



Designing Rigorous Quasi-Experimental Studies for Review by the Title IV-E Prevention Services Clearinghouse

Abt Global: Daniel Gubits, Rebecca McGill-Wilkinson, and Sandra Jo Wilson.

BRIEF | December 2024 | OPRE Report 2024-341

1. Introduction

The Title IV-E Prevention Services Clearinghouse[1] (hereafter the Clearinghouse) conducts independent, transparent reviews of research on mental health and substance use prevention and treatment programs and services, in-home parent skill-based programs and services, and kinship navigator programs intended to provide enhanced support to children and families and prevent foster care placements. This brief provides guidance for evaluators on designing and reporting high quality QEDs aligned with the Clearinghouse's *Handbook of Standards and Procedures, Version 2.0*. Using this brief, along with the *Reporting Guide for Study Authors*, can help evaluators design and conduct rigorous QEDs and avoid common reasons that QEDs receive low ratings on the Clearinghouse's design and execution standards.

The Clearinghouse currently[2] reviews Randomized Controlled Trials (RCTs) or Quasi-experimental Designs (QEDs) with at least one intervention condition and at least one comparison condition.

- RCTs are studies in which units are assigned to intervention and control conditions via a random process (e.g., a lottery).
- QEDs are studies in which units are assigned to intervention and comparison conditions via a non-random process.

The goal of both types of study designs is to produce credible estimates of the impact of interventions by creating intervention and comparison conditions that are as similar as possible; in this way, the only substantial difference between the two conditions is the presence of the intervention, and any differences in outcomes observed after the intervention can be attributed to the intervention and not some other factor.

What types of QEDs[3] are eligible for review?

A wide range of QEDs can be eligible for review including, but not limited to, matched and non-matched designs, natural experiments, before-after-control-impact designs, controlled before-and-after designs, and difference-in-difference designs. Conditions may be formed through natural processes or formed by researchers. Researchers may also use statistical techniques to match individuals (or groups of individuals) who participated in an intervention to individuals who did not participate based on their measurable characteristics (e.g., propensity score matching). Intervention and comparison conditions may be formed before or after outcome data are collected.

In the following sections, we describe the design and execution standards all QEDs reviewed by the Clearinghouse must meet to receive a rating of *Moderate* Support for Causal Evidence – the highest rating QEDs can receive. Specifically, QEDs must: (1) have no confounding factors; (2) use reliable and valid outcome measures that are consistently measured across the intervention and comparison conditions; (3) demonstrate baseline equivalence; (4) use an acceptable missing data approach; and (5) rely on appropriate statistical models to estimate intervention impacts. Comprehensive information about these and other Clearinghouse standards can be found in the *Handbook of Standards and Procedures, Version 2.0*.

[1] The Prevention Services Clearinghouse was established by the Administration for Children and Families within the U.S. Department of Health and Human Services. The standards and procedures used by the Clearinghouse were developed in accordance with the Family First Prevention Services Act of 2018 (FFPSA) as codified in Title IV-E of the Social Security Act.

[2] The Prevention Services Clearinghouse is planning a pilot to develop and implement standards and procedures for reviewing studies that use single case designs and determine how such designs may contribute to promising ratings.

[3] Although regression discontinuity designs are types of quasi-experimental designs (Shadish, Cook, & Campbell, 2002), separate eligibility and review criteria for such designs would need to be applied. The Prevention Services Clearinghouse does not currently have standards for rating regression discontinuity designs.

2. Avoiding Design Confounds

Clearinghouse Design Confound Standards (Handbook Section 5.9.3)

QEDs **must be free of** substantially different characteristics confounds and n=1 confounds in order to receive a moderate causal evidence rating by the Clearinghouse. Studies that exhibit either or both confounds receive a low causal evidence rating.

Evaluators designing QEDs need to plan ahead to avoid design confounds before beginning a study. A design confound is any factor other than the intervention that is both plausibly related to the outcome measures and completely or largely aligned with either the intervention condition or the comparison condition. In such cases, the confounding factor may have a separate effect on the outcome that cannot be eliminated by the study design or isolated from the intervention effect. When a confound is present, it is impossible to separate how much of any observed difference on the outcomes is related to the intervention and how much to the confounding factor. There are different kinds of design confounds, but all diminish the ability to observe the true impact of an intervention. The Clearinghouse recognizes the following types of confounds: (i) the substantially different characteristics confound, and (ii) the n=1 person-provider confound or administrative unit confound. See **Exhibit 1** for examples of the different types of confounds and how evaluators can avoid them when designing studies.

Exhibit 1. Types of Confounds and How to Avoid Them

Confound type	Substantially different characteristics confound	N=1 person-provider confound	N=1 administrative unit confound
This confound is present if...	A characteristic of the intervention or comparison condition, or a characteristic of the service provider for the intervention or comparison condition, is systematically different from that of the other condition.	All participants in the intervention condition or all participants in the comparison condition receive their respective services from a single provider that does not provide services to the other condition.	All participants in the intervention condition or all participants in the comparison condition receive intervention or comparison services in a single administrative unit (e.g., clinic, community).
An example of this confound is...	An intervention is offered to families who have had at least one substantiated claim of abuse or neglect in the past year and the evaluator identifies comparison condition families from the same community who are socioeconomically and demographically similar but who have not had any claims of abuse or neglect.	Intervention services are provided by a single therapist and the pre-post gains of her patients on a mental health assessment are compared with the gains of patients of another therapist.	Intervention services are provided in one state. Comparison services are provided in another state.
This confound results in a low causal evidence rating because...	The prior claim of abuse or neglect may be related to the outcome, thus confounding measurement of the intervention effect.	It is impossible to disentangle the effect of the intervention from the skills of the therapists.	The intervention effect may be confounded with the capacity of the administrative unit to produce better outcomes.
Evaluators can avoid this confound by...	Not using a specific participant or provider characteristic to determine who will receive the intervention.	Ensuring intervention and comparison conditions each receive services from more than one service provider *or* that participants in both conditions receive services from the same service provider.	Ensuring intervention and comparison conditions each receive services in more than one administrative unit *or* that participants in both conditions receive services in the same administrative unit.

It is especially important to check for confounds at the design stage of a QED. Because study participants are not randomly assigned to intervention and comparison conditions there is potential for one or more unobserved factors to be aligned with only one of the conditions, creating a design confound. Establishing baseline equivalence or including baseline covariates in a model used to assess the impact of the intervention cannot eliminate a confound if one is present (see the Baseline Equivalence section below for more information).

3. Reliable, Valid, and Consistent Measures of the Outcomes

Clearinghouse Measurement Standards (Handbook Section 5.9.2)

Outcome measures in QEDs **must** be reliable, face valid, and measured consistently across intervention and comparison groups in order to receive a moderate causal evidence rating by the Clearinghouse.

Evaluators designing strong QEDs should consider their outcomes carefully in the context of study objectives and research questions and select measures that are **reliable, face valid, and measured consistently** across intervention and comparison conditions.

Reliability

Using a measure that has low reliability means that there is more measurement error, which affects the precision of impact estimates and, by extension, affects the likelihood of finding statistical significance. To meet Clearinghouse design and execution standards, outcome measures and measures used to demonstrate baseline equivalence in a QED must meet internal consistency, test-retest reliability, or interrater reliability requirements or be assumed to be reliable (see box at right).

Measures assumed by the Clearinghouse to be reliable include:

- **Administrative records** obtained from schools, child welfare or other service organizations, hospitals, or clinics.
- **Demographic characteristics**, such as age, race/ethnicity, education level, socioeconomic status, etc.
- **Medical or physical tests**, such as urinalysis, blood pressure, etc.

Face Validity

The Clearinghouse also requires that measures be face valid. A measure that is face valid is clearly defined, has a direct interpretation, and appears to measure the underlying construct it was intended to measure. A measure that claims to be a measure of emotional adjustment, for example, but asks questions about favorite foods, exercise regimens, and smoking behavior does not have face validity. Other ways that measures can lack face validity occur when they lack a clear direction, meaning it is not clear whether low or high scores are preferable. For example, consider a measure of parenting behaviors in which lower scores reflect parenting that is lax or inconsistent and higher scores reflect behaviors that are harsh and negative. A measure like this does not have a clear favorable direction because both high and low scores can be interpreted as a poor outcome.

Measurement Consistency

Finally, to meet Clearinghouse design and execution standards, outcomes must be consistently measured across the intervention and comparison conditions. To demonstrate consistent measurement in a study, measures should be constructed in the same way and collected at the same time using the same collection methods for the two conditions. When the timing of measures is not the same for intervention and comparison conditions, it is impossible to be sure if differences at follow-up are due to the intervention or because the outcomes were measured differently in the two conditions.

4. Baseline Equivalence

Clearinghouse Baseline Equivalence Standards (Handbook Section 5.7)

QEDs **must** meet baseline equivalence standards by demonstrating baseline equivalence and by statistically controlling for baseline differences when required. QEDs that do not meet baseline equivalence standards cannot receive a moderate causal evidence rating by the Clearinghouse.

Baseline equivalence **must** be established using measures that are reliable, face valid, and consistently measured.

Baseline equivalence **must** be established using the analytic sample.

Baseline equivalence – or, how similar the intervention and comparison conditions are at *baseline* – plays an especially important role for QEDs. When conditions differ at baseline, any differences observed at *posttest* may be due to differences that were already present, rather than to the intervention itself. In such cases, it is impossible to know if the intervention truly affected the outcome or not. Randomized studies are expected to create conditions that are identical on both measured and unmeasured characteristics because randomly assigning participants to conditions means that membership in the intervention or comparison condition does not depend on any underlying characteristics of the participants.[4] QEDs, on the other hand, may be more vulnerable to baseline differences because assignment to intervention or comparison conditions is non-random and therefore can be related to underlying characteristics of the study participants. For example, study participants with greater needs may be more likely to be referred to a new program or service than to receive services as usual. As a result, designing and conducting a high-quality QED study requires paying special attention to baseline equivalence.

Baseline is a point in time prior to the start of the intervention for the intervention condition and the equivalent point in time for the comparison condition.

Posttest is a point in time after the start of the intervention at which outcomes are measured for intervention and comparison conditions. Follow-up may occur during an intervention, immediately after the intervention is completed, or at a later time point.

There are several strategies evaluators can use to construct intervention and comparison conditions with similar characteristics. For example, an evaluator could identify intervention and comparison conditions from among families in the same community or from individuals seeking treatment at the same clinic. Evaluators can also use statistical matching techniques, such as propensity score matching, to ensure intervention and comparison conditions are similar on baseline variables (see the Using Matching Strategies section below). Regardless of how intervention and comparison conditions are formed, evaluators should collect information about the participants in their intervention and comparison conditions prior to the start of the intervention so that they can assess baseline equivalence.

Using data collected prior to the start of the intervention, evaluators should demonstrate baseline equivalence between the conditions on *baseline* measures that are related to the outcomes. This could include direct pretests of the outcomes, other measures that are similar to or correlated with the outcome, and sociodemographic characteristics. For example, for a study comparing a court-mandated therapy program versus typical community mental health services on parental depression, an evaluator should ensure, at a minimum, that the parents in the court-mandated therapy program and the parents in typical services have similar levels of depression symptoms prior to beginning therapy.

The study participants for whom evaluators collect baseline data are referred to as the *baseline sample*. The sample used to estimate the effect of the intervention is the *analytic sample*. To ensure differences in outcomes between intervention and comparison conditions reflect only the effect of the program or service and not prior differences, baseline equivalence must be established on the analytic sample to meet design and

[4] In some cases, randomized designs do have issues with baseline equivalence, such as when many participants or groups of participants leave a study before it is over.

execution standards (see example in box). The Clearinghouse requires that each outcome in a QED study meet baseline equivalence standards on the respective analytic sample available for that outcome.[5]

Acceptable Variables for Establishing Baseline Equivalence

In most cases, the Clearinghouse prefers establishing baseline equivalence using a *direct pretest* (i.e., a pretest that uses the same or nearly the same measurement instrument as is used for the outcome) or a *correlated pretest* (i.e., a measure in any eligible outcome domain that has a correlation of 0.60 or higher with the outcome in the analytic sample). This is because pretest measures are assumed to be highly correlated with the study outcomes and correlated pretests demonstrate a high correlation with study outcomes.

In certain cases, assessing baseline equivalence on a direct pretest or correlated pretest may not be possible. When this occurs, the Clearinghouse recommends demonstrating baseline equivalence using a *pretest alternative* (i.e., a measure that is in the same or similar domain as the outcome). For example, if collecting a pretest measure of prior abuse or neglect reports is not possible because no children had prior reports at baseline, a measure of children maltreatment risk could serve as a pretest alternative. If no direct or correlated pretests are available and an appropriate pretest alternative is not possible, the Clearinghouse requires baseline equivalence to be demonstrated on *multiple sociodemographic characteristics*[6].

Assessing Baseline Equivalence

The Clearinghouse uses a standardized mean difference effect size to determine whether there is baseline equivalence between the intervention and comparison conditions. The Clearinghouse calculates the effect size and then determines whether the groups are similar by comparing the effect size for each baseline variable against the standard for baseline equivalence (see **Exhibit 2)**.[7] If intervention and comparison conditions are equivalent at baseline on the selected baseline measure, no statistical adjustments are required. If the baseline effect size is within the adjustment range[8], the analytic models must adjust for the baseline measure, such as by including the baseline measure as a covariate. When the baseline effect size is in the adjustment range, statistical adjustments are expected to effectively counteract any bias that would result from pre-existing differences between conditions. If baseline differences are outside the adjustment range, the groups are far from being equivalent and statistical adjustments are not sufficient to meet Clearinghouse design and execution standards. There are a variety of ways to adjust for baseline measures. Section 5.8 of the *Prevention Services Clearinghouse Handbook of Standards and Procedures, Version 2.0* describes the methods that are accepted by the Clearinghouse.

Why is baseline equivalence on the analytic sample required by the Clearinghouse?

Example: In a study of a court-mandated therapy program, participants in the baseline sample for the intervention and comparison conditions have similar levels of depression symptoms at the beginning of the study. However, participants in the intervention condition with the most serious symptoms at baseline drop out of the study. Although the intervention and comparison conditions appear equivalent when looking at the *baseline* sample, the intervention and comparison conditions do not have similar levels of depression symptoms at baseline when looking at the *analytic* sample because the most serious cases left the study. At the end of the study, the remaining intervention condition participants appear to have fewer depression symptoms than the comparison condition participants. This could be due to the fact that participants with more serious symptoms dropped out of the intervention condition rather than that the intervention was effective at reducing depression symptoms. Without knowing whether the participants who are in the sample used to estimate the effect of the intervention (the analytic sample) are equivalent at baseline, we cannot know whether the impact we observe is due to the intervention.

---

[5] The *Handbook of Standards and Procedures, Version 2.0* provides more information on how studies are evaluated for baseline equivalence (Section 5.7), including formulae for calculating effect sizes (Section 6.4).

[6] The *Handbook of Standards and Procedures, Version 2.0* provides information on when and how sociodemographic variables can be used to establish baseline equivalence (Section 5.7)

[7] If baseline effect sizes or information that can be used to calculate baseline effect sizes are not reported in the study, the Prevention Services Clearinghouse will contact the author to request the needed information.

[8] Effect sizes between 0.05 and 0.25 standard deviation units are considered to be in the adjustment range.

Exhibit 2. Prevention Service Clearinghouse Baseline Equivalence Standards for QEDs

Effect Size (ES) Difference in Standard Deviation (SD) Units	Baseline Equivalence Characterization	Requirement for Statistical Control for Baseline Difference	Highest Possible Causal Evidence Rating
ES < 0.05 SD	Equivalent	No statistical control required	Moderate
0.05 SD ≤ ES < 0.25 SD	Within the adjustment range	Must statistically control for baseline measure	Moderate
ES ≥ 0.25 SD	Outside of the adjustment range	Not applicable	Low

Using Matching Strategies

Evaluators may decide to use one or more of a variety of matching strategies to help form intervention and comparison conditions that are equivalent (or as close to equivalent as possible) on variables that can be measured at baseline. These techniques vary in complexity and range from one-to-one matching of intervention and comparison participants on pretests or other characteristics to statistical matching using propensity scores.

A comprehensive discussion of how to employ matching strategies in the design of a QED is not covered in this brief. Evaluators who wish to use matching are encouraged to consult other available resources on matching techniques.[9] In addition, evaluators whose matching studies may be considered for review by the Clearinghouse should pay particular attention to the general guidance below. Failure to follow these practices is the most common reason matching studies do not meet Clearinghouse design and execution standards.

General Guidance for Matching Strategies

- **Adjust for matching with replacement.** Matching with replacement — that is, when a single comparison participant is matched to multiple intervention condition participants — can improve baseline equivalence between intervention and comparison conditions because it provides more opportunity to find close matches for intervention condition participants. However, it introduces statistical dependencies among comparison condition members. These dependencies must be accounted for in impact analyses, or the results of the study will be biased. When writing up their studies, evaluators should: (i) indicate whether matching procedures were conducted with or without replacement; and (ii) if matching procedures were conducted with replacement, adjust their standard error estimates to account for a comparison unit being matched multiple times (for example, by applying a clustering correction). Evaluators should also adjust any baseline statistics reported in the same way and report the analytic sample size using the number of unique participants in the analysis.

- **Do not include endogenous variables in the matching model.** An endogenous variable is one that is measured or obtained after baseline and that could have been influenced by the intervention. Including endogenous variables in a matching model can lead to over- or under-estimating the effect of the intervention. A matching model that includes an endogenous variable would not meet Clearinghouse statistical model standards (see the Appropriate Statistical Models section for more information).

- **Report baseline equivalence on pretests and other baseline variables.** Although it is best practice in studies that use statistical matching to demonstrate baseline equivalence on propensity scores or other distance metrics, evaluators should also report equivalence on the individual variables included in the matching model after completing the matching. The Clearinghouse cannot assess baseline equivalence by comparing propensity scores.

- **Including only the propensity scores in an impact model would not be sufficient if the Clearinghouse requires adjustment for baseline differences in the analyses.** For example, if the pretest effect size is in the adjustment range after matching, the pretest score must be included in the impact analyses in order to meet design and execution standards even if the pretest is included in the propensity score model.

[9] This publication provides guidance on these approaches in the context of evaluating home visiting programs.

5. Addressing Missing Data

Clearinghouse Missing Data Standards (Handbook Section 5.9.4)

QEDs **must** use an acceptable approach to address missing data, when present, in order to receive a moderate causal evidence rating. For some missing data approaches, the Clearinghouse will assess the study for bias due to imputation of missing data.

It is common for studies to have **missing data** for some members of the study sample. Missingness can occur on baseline measures (such as if a participant skipped some questions on a survey or was not able to attend a baseline interview) or outcome measures (such as if a participant left the study or moved to a different community and could not be located). Missing data can affect the credibility of a QED because it can decrease the sample size and introduce bias into the study findings. For example, if a study is missing baseline data for a portion of the sample, baseline equivalence cannot be established on the analytic sample.

It is important to try to minimize opportunities for missing data from the very start of a study. Strategies include rigorously tracking participants, ensuring that data collection is accessible and within the capabilities of the participants, and developing data collection protocols that are only as complex or as extensive as needed to meet study objectives. There may still be missing data even in the most carefully monitored study. In such cases, evaluators should carefully consider the options for addressing missing data in their analyses because how evaluators address missing data in their studies can protect against or introduce bias and affect the credibility of the findings.

When missing data are present in studies reviewed by the Clearinghouse, the missing data standards apply. The Clearinghouse missing data standards in Section 5.9.4 of the *Handbook of Standards and Procedures, Version 2.0* describe several acceptable missing data approaches that are acceptable for QEDs. The Clearinghouse first assesses whether the missing data strategy is acceptable and then, in some cases, will assess the potential for bias that could result from the strategy. The Clearinghouse ***Reporting Guide for Study Authors*** describes the information the Clearinghouse uses to assess bias when there are missing data and provides table shells that evaluators can use to summarize the data elements that are required for the bias assessment.

6. Appropriate Statistical Models

Clearinghouse Statistical Model Standards (Handbook Section 5.9.1)

QEDs **must not** include endogenous covariates in the statistical models in order to receive a moderate causal evidence rating from the Clearinghouse.

Evaluators have many options for analyzing the impacts of the interventions in a QED and should select statistical models that are appropriate for their research questions, study design, and data structure. When assessing the statistical models employed in a study under review, the Clearinghouse reviews the models for issues that could lead to biased impact estimates (that is, estimates that over- or underestimate the effzect of the intervention).

One common issue is when evaluators include *endogenous covariates* in their statistical models. As mentioned earlier, endogenous covariates are variables that

Example of an Endogenous Covariate

Evaluators notice that participants assigned to the intervention condition spent different amounts of time in the therapy sessions and they decide to include a covariate in the impact model that accounts for time spent in therapy. This dosage covariate is endogenous because the intervention might have created systematic differences in dosage between the intervention and comparison conditions that would bias the estimate of the effects of the intervention. The findings from a model like this would receive a low causal evidence rating from the Clearinghouse. Questions about the influence of dosage may still be an important element of an evaluation. In such cases, evaluators may decide to report their analyses with and without the endogenous covariate to facilitate Clearinghouse review.

are measured or obtained after baseline and that could have been influenced by the intervention. Models with endogenous covariates receive a low causal evidence rating.

Evaluators should avoid endogenous covariates (or report models with and without the covariates). However, including non-endogenous covariates in analytic models can be beneficial. For example, including the pretest in a statistical model may increase statistical precision because the pretest is assumed to explain some of the variance in the outcome. Including other non-endogenous covariates that explain some variance in the outcome may also increase statistical precision.

7. Conclusion

QEDs can contribute valuable evidence to understanding the effectiveness of programs and services intended to provide enhanced support to children and families and prevent foster care placement. In this brief, we provide information about designing and analyzing high-quality QEDs and relevant Prevention Services Clearinghouse design and execution standards. Applying the guidance described here and using the ***Reporting Guide for Study Authors*** when writing up study results can result in high-quality QEDs and can also facilitate the Clearinghouse's review of QEDs.

For additional information, we recommend:

- ***Title IV-E Prevention Services Clearinghouse Handbook of Standards and Procedures, Version 2.0***
- ***Reporting Guide for Study Authors***
- Responses to **frequently asked questions (FAQs)**
- Sign up for the Clearinghouse's **email list**, which we use to announce additional resources as they become available.

Designing Rigorous Quasi-Experimental Studies for Review by the Title IV-E Prevention Services Clearinghouse

OPRE Report 2024-341
December 2024
Daniel Gubits, Rebecca McGill-Wilkinson, and Sandra Jo Wilson

Submitted to:
Christine Fortunato, Jenessa Malin, and Laura Nerenberg, Project Officers
Office of Planning, Research, and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services

Contract Numbers: HHS P233201500069I | GS00F252CA

Project Director: Sandra Jo Wilson
Abt Global
6130 Executive Blvd.
Rockville, MD, 20852

This report is in the public domain. Permission to reproduce is not necessary. Suggested citation: Gubits, D., McGill-Wilkinson, R., and Wilson, S. J. (2024). *Designing Rigorous Quasi-Experimental Studies for Review by the Title IV-E Prevention Services Clearinghouse*, OPRE Report # 2024-341, Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Disclaimer
The views expressed in this publication do not necessarily reflect the views or policies of the Office of Planning, Research, and Evaluation, the Administration for Children and Families, or the U.S. Department of Health and Human Services.

This report and other reports sponsored by the Office of Planning, Research and Evaluation are available at **www.acf.hhs.gov/opre**.

Subscribe to OPRE News and Follow OPRE on Social Media