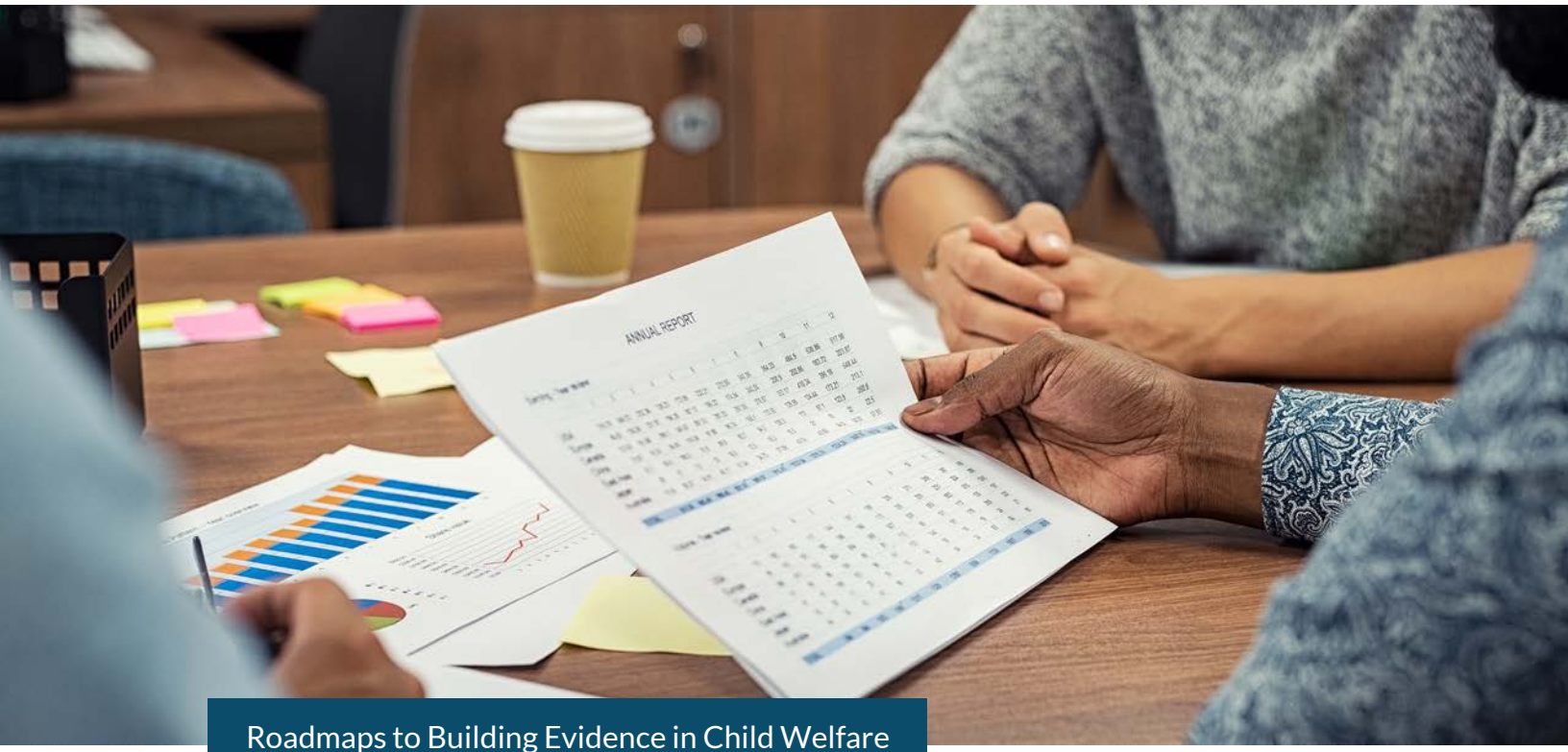


Roadmaps to Building Evidence in Child Welfare

RESEARCH REPORT

Quasi-Experimental Designs in Child Welfare Evaluations

Opportunities for Generating Rigorous Evidence

Laura Packard Tucker

OPRE Report #2022-42

August 2022

CWEST | CHILD WELFARE EVIDENCE STRENGTHENING TEAM

Quasi-Experimental Designs in Child Welfare Evaluations

Laura Packard Tucker

OPRE REPORT #2022-42
AUGUST 2022

SUBMITTED TO

Kathleen Dwyer and Alysia Blandon, project officers
Office of Planning, Research, and Evaluation
Administration for Children and Families
US Department of Health and Human Services

Contract Number: HHS P233-2015-00064I

SUBMITTED BY

Michael Pergamit, Principal Investigator
Urban Institute
500 L'Enfant Plaza SW
Washington, DC 20024

DISCLAIMER

Sign-up for the OPRE newsletter

Follow OPRE on Twitter @OPRE_ACF

facebook.com/OPRE.ACF

Connect on LinkedIn company/opreacf

Follow OPRE on Instagram @opre_acf

Contents

Acknowledgments

This report is part of activities to support evidence building in child welfare through a contract to the Urban Institute funded by the Department of Health and Human Services, Administration for Children and Families. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission. We would also like to thank our project officers Kathleen Dwyer and Alysia Blandon for their guidance and input as well as Cara Kelly and other reviewers at OPRE and the Children's Bureau for their valuable comments. The Supporting Evidence Building in Child Welfare project (https://www.acf.hhs.gov/opre/project/supporting-evidence-building-child-welfare-2016-2025) includes conducting rigorous evaluations of child welfare programs, practices, and policies as well as building evaluation capacity in the child welfare field. All the Roadmaps to Building Evidence in Child Welfare products can be found at https://www.urban.org/projects/roadmaps-building-evidence-child-welfare.

The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at urban.org/fundingprinciples.

Quasi-Experimental Designs in Child Welfare Evaluations

Impact Evaluations and Evidence Building

The child welfare field requires more and better evidence about what works to support children and families. Conducting rigorous impact evaluations is one way to generate that evidence. As an alternative to randomized controlled trials (experimental designs), quasi-experimental designs can also yield reliable evidence about program effectiveness and may be more appropriate in certain contexts. This report defines quasi-experimental designs (QEDs) and summarizes their benefits and challenges. It provides an overview of four common QEDs and uses child welfare examples to show where child welfare administrators might use these evaluation designs in their work.

Making better decisions in the child welfare field requires evidence. Evidence is information used to support an observation, claim, hypothesis, or decision. In other words, evidence answers the question, "How do you know?" (Wulczyn et al. 2014). Child welfare agencies need evidence that their programs are effective. This evidence helps you understand if programs are supporting children and families well.

Research evidence is evidence you build using scientific principles. It can come from various sources (such as administrative data, focus groups, or surveys) and may be quantitative or qualitative (Wulczyn, Alpert, and Monahan-Price 2016). Evaluation produces one type of research evidence and answers basic questions about a program's implementation and effectiveness (OPRE 2018). Several evaluation types exist. Your choice of evaluation type should depend on your research questions, the maturity of your intervention, and the local context. For example, the way a state rolls out a program's implementation may make certain evaluation types impractical. Or the kinds of child-level data a child welfare agency collects may make certain types more workable.

One type of evaluation is an impact evaluation. Impact evaluations test the effectiveness of a well-defined intervention. A well-defined intervention has a manual or other clear, written guidance dictating its operation (Walsh et al. 2015).[1] To be ready for an impact evaluation, the intervention should operate in a stable context and have clearly defined, expected outcomes. Expected outcomes would come from the intervention's theory of change or from past experience. A formative evaluation

[1] "Implementing Evidence-Based Practice," Child Welfare Information Gateway, accessed July 20, 2021, https://www.childwelfare.gov/topics/management/practice-improvement/evidence/implementing/.

could help you observe and refine your expected outcomes before undertaking an impact evaluation. Impact evaluations can answer the question, "Are families better off after participating in my intervention than if they had not?" You often answer this question by comparing families who participated with those who did not. Rigorous impact evaluations can identify interventions that positively impact children and families. They can also identify interventions that have no impact on children and families or even negative impacts. This knowledge helps child welfare agencies fund efforts to better meet their missions and help children and families.

Randomized controlled trials (RCTs) are considered the "gold standard" of rigorous impact evaluations (White and Sabarwhal 2014). But they may not always be workable or the most appropriate design, given the research questions or local context. You should conduct RCTs when possible and appropriate. If you cannot conduct an RCT, QEDs are a good alternative.

An Overview of Quasi-Experimental Designs

What Is a QED?

To understand QEDs, it is helpful to first understand why RCTs are considered the "gold standard." RCTs are very good at assessing the effectiveness of interventions. This is because RCTs randomly assign people to either the intervention (the treatment group) or services as usual (the comparison group). This creates two similar groups, except for the services they receive. By making the groups similar, RCT evaluations can measure an intervention's impact on outcomes while ruling out other explanations (Hanson and Pergamit 2022; JBA 2013).

Sometimes RCTs are impossible, impractical, or not appropriate to carry out. This may be the case when there is not excess demand for services. Excess demand means that a program does not have enough resources to serve all the people who could benefit from it. Without excess demand, creating a control group for an RCT would mean denying an intervention to those who might benefit because you could have provided the intervention to some or all of the people in the control group. It is also possible that the type of program may not be a good fit for an RCT. For example, a kinship navigator program may provide navigation and mentoring services. These services help caregivers learn how to navigate the child welfare system or other systems that help families meet general needs. Kinship caregivers who participate in the program could share this knowledge with other kinship caregivers who did not participate in the program. In this case, even if you randomly assigned services, the treatment may still affect caregivers not in the treatment group, making an RCT impractical.

In cases where an RCT is not suitable, QEDs can offer more accessible evaluation methods. QEDs do not randomly assign people to treatment and comparison groups. Instead, QEDs use other methods to identify a comparison group. These methods try to make the comparison group as similar as possible to the treatment group based on preintervention (baseline) characteristics such as age, gender, or race. In this report, we review four common types of QEDs: matched-group, interrupted time series (ITS), difference-in-difference (DiD), and regression discontinuity (RDD).

The Challenge of Causality

When providing a service to children and families, you want to know whether it is working. Are your services having a positive effect? Impact evaluations using QEDs can help answer that question. You may observe through routine program monitoring that families who got an intervention are more likely to experience a positive outcome than families who did not get the intervention. But you should not assume this observation proves that the intervention *caused* the outcome. In other words, you do not know whether the link between the intervention and the outcome is causal.

The key challenge to crafting a QED is establishing this causal link. To claim an intervention caused an outcome, you must rule out other possible explanations. The most common other explanation is preintervention, or baseline, differences between your comparison and treatment groups. These group differences could be the participants' age, race, socioeconomic status, motivation, or something else. Differences may also come from the program context, such as where, when, or by whom the intervention was implemented. These differing characteristics are confounding variables. A confounding variable is an "extra" variable that you did not account for when looking at the association between an intervention on an outcome. A confounding variable can make it look like an intervention caused an outcome when that is not true.

Participant Characteristic Example

Imagine a parent training course meant to increase positive parent-child interactions. An evaluation may find positive outcomes for families involved compared with families who did not participate. But these outcomes could occur because parents who are more motivated to improve their parenting were more likely to choose to attend the course. The parents' motivation to improve is a confounding

variable.[2] If the comparison group is not similarly motivated, this difference could make it look like the training course caused the outcome.

Program Context Example

Imagine a state implemented an intervention meant to prevent reentry into foster care. They targeted recently reunified families and focused on service coordination. The state implemented the intervention in a couple of urban counties and compared their outcomes with those observed in a selection of rural counties. The geographic area where they implement the intervention may impact outcomes. For example, the types and levels of services available may differ between urban and rural areas. Differences we find in outcomes may have more to do with geography than treatment.

QEDs and Confounders

No matter the type of QED, you need to consider confounding variables as a potential issue. Each type of QED approaches the problem of confounding variables differently. QED methods try to make the comparison and treatment groups as similar as possible. QEDs do not randomly assign participants to treatment, so the possibility always exists that confounding variables are impacting outcomes. Some confounding variables may not exist in the data you have or are difficult to measure. These could include motivation to change, mental health status, or any characteristic related to outcomes. For more information on confounding variables, check out the *What Researchers Mean By* guide, created by the Institute for Work and Health (Moser and Vu 2017).

Clearinghouses and Confounders

Some evidence-rating clearinghouses provide guidance on confounding variables (Brewsaugh and Prendergast 2022). For example, the Prevention Services Clearinghouse[3] requires that evaluators compare population differences between the treatment and comparison groups in socioeconomic status, race/ethnicity, and age. The *Title IV-E Prevention Services Clearinghouse Handbook of Standards and Procedures* (Wilson et al. 2019) also warns of situations where service acceptance is used to create the

[2] This kind of difference between treatment and comparison groups is referred to as self-selection or joiner bias. Self-selection bias happens when people choose whether or not to participate in the intervention and the group that chooses to participate is different than the group that opts out.

[3] "Welcome," Title IV-E Prevention Services Clearinghouse," accessed August 1, 2022, preventionservices.acf.hhs.gov.

treatment group. For instance, this could happen when the treatment group includes individuals who accepted treatment and most or all of the comparison group may have refused treatment. Or the program (treatment) may only have been offered in some parts of the state. These communities could differ significantly in social context (such as in housing costs, labor market, or availability of social services), and these differences could impact outcomes (Wilson et al. 2019).

The Pros and Cons of QEDs

Pros. Although they may not apply in every situation, there are several possible advantages to evaluations using QED methods:

- **Greater stakeholder support.** You may be able to gain stakeholder support more easily than with RCTs. QEDs do not use random assignment to allocate services, and random assignment can make some stakeholders uncomfortable because of the appearance of denying some individuals services, even if the evaluation does not cause fewer people to be served. QEDs may provide another design option while still generating rigorous evidence.
- **Higher external validity.** QEDs can often test interventions and policies in real-world settings without many, if any, program or policy implementation adjustments. This means they can sometimes have higher external validity compared with RCTs. External validity is how well a study's findings would apply in a different context. Without external validity, you cannot apply results from your study to other settings. RCTs can involve altering the service context such as by changing the way individuals are referred to a program. In contrast, QEDs are often implemented in conditions that differ little, if at all, from the usual service context. This can make it more likely that QED findings will apply to real-world conditions.
- **Many opportunities to use administrative data.** Because QEDs often rely on available administrative data, they may not need to collect new data. Specifically, QEDs often use administrative data to identify and create a comparison group. Child welfare agencies collect administrative data as part of daily operations, typically in their Statewide Automated Child Welfare Information System (SACWIS) or Comprehensive Child Welfare Information System (CCWIS) systems (Packard Tucker and Zhou 2022). So using administrative data can reduce the time, capacity, and funding needed for your evaluation.
- Easier evaluation implementation. As we already said, QEDs can test interventions and policies in real-world settings. You can often start a QED evaluation without making major changes to

program implementation. It can also be easier to implement certain QEDs, rather than an RCT, after program implementation is under way. This means that QEDs can be simpler in some ways to implement than RCTs.

Cons. QEDs have some possible drawbacks:

- **Lower internal validity.** Compared with RCTs, QEDs may offer weaker evidence that an intervention caused an outcome (internal validity). This means that they have lower internal validity relative to RCTs. For your study to have internal validity, you need to be able to rule out alternate explanations for your findings.
  - » Threats to internal validity can occur when anything other than the program can provide an explanation for the outcome. This can happen when the treatment and comparison groups are different on a key characteristic. The characteristic may be unknown or unmeasured. These group differences limit your confidence that outcomes were caused by the program.
  - » Selection bias threatens the internal validity of QEDs. Selection bias happens when the individuals who end up in the treatment group differ from those who end up on the comparison group because of the way they were selected for or chose to enter the program. Think back to the parent training course example above. There, outcomes changed because parents who were more motivated to improve their parenting were more likely to choose to attend the course.
  - » QEDs can be more challenging to implement than RCTs because of the need for careful consideration of how the comparison group is formed.
- **Underlying QED model assumptions.** Each QED type contains a set of underlying assumptions. A QED's success relies on how well the chosen study type (such as matched-group, ITS, DiD, or RDD) adheres to these assumptions. But it is often difficult or impossible to test the validity of those assumptions. We outline the assumptions underpinning four common types of QEDs in the following section.
- **Fewer opportunities for primary data collection.** Although collecting primary data is a resource-heavy activity, it can provide important information for an evaluation. In QEDs, you may have fewer opportunities to collect primary data than in RCTs. For many QEDs, you will not have contact with the comparison group. For example, you may find them in historical data in an ITS design or in a neighboring county in a DiD design. Without contact with the comparison group, you lose the ability to collect new data from them via primary data collection.

- **Conflicting Clearinghouse requirements.** Evidence-rating clearinghouses accept QEDs, but most only accept some types. Most clearinghouses accept matched-group QEDs. A few accept nonmatching QEDs, such as RDD, DiD, or ITS. See table 2 in the Guide to Designing Rigorous Impact Evaluations in Child Welfare (Brewsaugh and Prendergast 2022) to determine which QEDs are accepted by clearinghouses relevant to child welfare programs.

Successfully conducting a QED requires experience. Because of the considerations and challenges listed above, you should partner with an experienced evaluator to design and implement your QED. We recommend that you partner with an evaluator experienced in using your desired QED type.

Common QEDs

The following sections review four common types of QEDs. We describe their methods, assumptions, and requirements. We will also provide insight into the opportunities and challenges posed by these QEDs through real-world examples. You can find a summary of the common QEDs' characteristics in appendix A.

Matched Group

DEFINITION

QEDs try to make the comparison and treatment groups as similar as possible. In a matched-group design, you do this by matching the characteristics of the treatment group members with those of the comparison group members (Hanita, Ansel, and Shakman 2017).

The evaluator matches the treatment and comparison groups on matching variables. A matching variable is a characteristic of a person (such as race, age, family size, socioeconomic status, or child welfare service history) or place. You use matching variables to find someone in the comparison group who looks like a treatment group member (CEBP 2014). You should make matches based on characteristics that could affect the outcome of interest. If possible, we recommend also matching the two groups on the outcome of interest, measured at baseline (Brewsaugh and Prendergast 2022).

Matched-group QEDs have two matching strategies (figure 1):

1. **Exact matching.** In exact matching, each person in the treatment group is exactly matched to a person in the comparison group on every matching variable. Exact matching can be done manually, but the process quickly becomes complicated with many matching variables. When

many characteristics are used, it may be impossible to find exact matches for everyone, and statistical matching may be a better option.

2. **Statistical matching.** In statistical matching, you match people in the treatment and comparison groups using a statistical method. A common method used is propensity score matching (PSM). Under PSM, you match the treatment and comparison groups using a propensity score, which is the likelihood of a person being in the treatment group, based on their characteristics (Austin 2011). The higher the score, the better the match. In other words, is the person in the comparison group similar enough to people in the treatment group that they could just as likely have received treatment?

FIGURE 1

Two Methods to Create Matched Groups

Source: The author developed this graphic to visually demonstrate exact matching and statistical matching.
Notes: The colored shapes to the right stand for unique combinations of the characteristics you may use to create matched groups. For example, an orange circle can be a 19-year-old Latina female with a high school diploma while a pink circle is a 19-year-old Black female with a GED. The dotted lines are matched pairs of people in the treatment and comparison groups. Exact matching selects comparison group members who perfectly match on all chosen matching variables (same color and shape). Unmatched participants are dropped out of the analysis (shown with an X). Statistical matching selects comparison group members that match each treatment group member as closely as possible (e.g., same color but different shape, same shape but different color, or same shape and color).

EVALUATION OPPORTUNITY

A matched-group design works when a group of individuals like the treatment group is not receiving treatment. You also need to have access to the same type of data for both groups. At a minimum, you will need demographic data to match the groups and outcome data to look at intervention effectiveness.

It would also help to have data on other characteristics (such as data on their child welfare history). These data may exist when you stagger the rollout of an intervention either by provider or geographically. A staggered rollout could provide a similar comparison population in the areas where the intervention has not yet been implemented. By comparing the outcomes in the matched comparison group with the treatment group, you can evaluate an intervention's effect.

ASSUMPTIONS AND CONSIDERATIONS

Administrative data (e.g., SACWIS or CCWIS data) are a valuable resource for matched-group QEDs. You can link child welfare administrative data with data from other sources (such as program data or data from other public programs such as Temporary Assistance for Needy Families or Medicaid). You can use these data to look at outcomes and establish matching variables.

A matched-group QED depends on the treatment and comparison groups' similarity at baseline, so you must confirm the baseline equivalence of the two groups. Baseline equivalence is the extent to which the treatment and comparison groups are similar to one another at baseline. Be sure to check baseline equivalence on all matching variables and on any variables that could provide alternative explanations for group differences after treatment.

ADVANTAGES AND DISADVANTAGES

A primary challenge in QEDs is creating groups that mimic random assignment as closely as possible. An advantage of a matched-group design is that the matching minimizes selection bias when random assignment is not possible. As we have said before, selection bias is a threat to your evaluation having equivalent treatment and comparison groups, which weakens your internal validity (the ability to say that the intervention caused the outcomes observed). There is no way to remove all internal validity threats in a QED. But matching on observed characteristics that will likely impact the outcome is a good way to limit the presence of internal validity threats. This increases confidence in your findings.

BOX 1

Matched-Group Evaluation Example—Youth Villages's Intercept Program in Tennessee

Intercept™ is an intensive in-home services program that aims to strengthen families to prevent or limit the need for foster care. The program targets families with children at risk of placement and families with children who are already in foster care. An evaluation in Tennessee looked at Intercept provided by Youth Villages. The researchers used a matched-group design with exact matching. Matching variables included gender, race/ethnicity, age, perpetrator type, family income, family safety, investigation details, and the young person's education, developmental challenges, mental health, and substance use.

The study showed a positive program impact on preventing placement and permanency. The Prevention Services Clearinghouse gave Intercept a Supported rating based on this study. A Supported rating reflects a program with evidence that included at least one comparison group, met the clearinghouse's design standards, and showed an effect at least 6 months beyond the end of treatment.

Sources: Scott Huhr and Fred Wulczyn, Do Intensive In-Home Services Prevent Placement? A Case Study of Youth Villages' Intercept® Program (Chicago: Chapin Hall at the University of Chicago; Center for State Child Welfare Data, 2020), https://fcda.chapinhall.org/wp-content/uploads/2019/10/YV-Intercept-Results-1-8-2020-final.pdf; Scott Huhr and Fred Wulczyn, *Do Intensive In-Home Services Promote Permanency?: A Case Study of Youth Villages' Intercept® Program* (Chicago: Chapin Hall at the University of Chicago; Center for State Child Welfare Data, 2020), https://fcda.chapinhall.org/wp-content/uploads/2020/09/Permanency-YVIntercept-final-982020.pdf; "Program and Service Ratings," Title IV-E Prevention Services Clearinghouse, accessed July 20, 2021, https://preventionservices.abtsites.com/review-process/psr; "Intercept®," Title IV-E Prevention Services Clearinghouse, accessed July 20, 2021, https://preventionservices.abtsites.com/programs/238/show.

Matching poses both challenges and opportunities. When matching, you have to judge how to match and on what characteristics. Additionally, the data necessary for a good matching strategy may not be available. Data may be missing because the necessary information is not collected, not collected well, or not observable. Even in rigorous matched-group designs, you cannot remove the possibility that treatment and comparison groups differ on some characteristics.

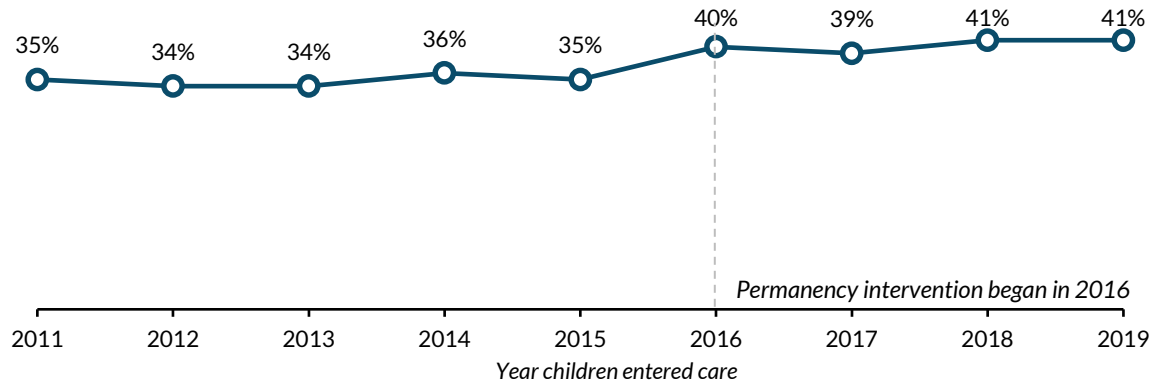
Interrupted Time Series (ITS)

DEFINITION

A time series design involves repeated measurements taken over time for a group of people. In an ITS design, the evaluator tracks an outcome over time to show an underlying trend in that outcome. Then, the evaluator assesses whether the implementation of an intervention "interrupted" that trend. You use an ITS design to attribute change over time in a population-level outcome to an intervention (an outcome like the rate of placement in the general population or the likelihood of permanency for children entering out-of-home care). The population before the intervention acts as the comparison group, and the same population after the intervention acts as the treatment group.

FIGURE 2

ITS Example—Permanency within 12 Months



Source: The author developed this figure to display a hypothetical example.
Notes: This figure looks at a hypothetical state's share of children entering out-of-home care who then exit to permanency within 12 months. In 2016, this state implemented an intervention that aimed to improve timely permanency for all children entering care.

In figure 2, our time series tracks a state's rate of permanency within 12 months for 9 years. From 2011 to 2015, the rate varies slightly but hovers around 35 percent. In our hypothetical example, we imagine this state implemented an intervention in 2016 meant to improve permanency. After this implementation, the average share of children exiting to permanency within 12 months increases to 40 percent. In this case, the new intervention "interrupted" the permanency trend and improved outcomes.

EVALUATION OPPORTUNITY

You may want to consider an ITS design when there is a large-scale intervention implementation or a new policy change. ITS is well suited to look at the impact of these larger implementations under two conditions. First, you need to have introduced the intervention for a population over a clearly defined period. And the intervention is meant to target the whole population (Lopez Bernal, Cummins, and Gasparrini 2017).

ASSUMPTIONS AND CONSIDERATIONS

An ITS design requires you to meet two assumptions:

- Any trends in the outcome before the intervention need to be along a straight or nearly straight line. The trend can occur with a steady increase or decrease of an outcome over a period or, as in figure 2, the trend line can be flat, showing no real change over time. You can check whether trends are along a straight line by looking at the visual trend and performing statistical checks.

- Any trends in the outcome before the intervention would have continued unchanged. In other words, without the intervention, you would not have expected the outcome trends to change. In figure 2, under this assumption, we would expect the rate of permanency to continue at about 35 percent through 2019 without the intervention. This assumption includes any changes because of external factors affecting the outcome trend, such as other interventions or policy changes.

Also, your data need to meet a few requirements:

- There must be a clear cutoff date in the data between the preintervention period and the postintervention period.
- Outcome data need to be available at regular time points both before and after the intervention. Measuring your outcome at, say, monthly, 6-month, or annual intervals will create a time series.
- Regular demographic and other characteristic data must be available pre- and postintervention as well. This is because you must examine the composition of the population both before and after the intervention. A basic ITS model cannot handle changes in the population composition over time. But there are statistical methods to handle that situation if it arises.

ADVANTAGES AND DISADVANTAGES

An advantage of ITS designs is their ability to use administrative data well. Because they focus on population-level outcomes, ITS designs can capitalize on longitudinal administrative data. Using administrative data can reduce the data collection burden in evaluations while still providing the information needed to measure key child welfare outcomes (such as child safety, placement type, and permanency) across time (Packard Tucker and Zhou 2022).

BOX 2

ITS Evaluation Example—Screening and Child Maltreatment in the Netherlands

In the Netherlands, seven hospital emergency departments implemented a new checklist for screening for child abuse along with additional training for nurses. Researchers used an ITS design to compare the screening and detection rates for child abuse before and after implementation in those seven hospitals. By comparing pre- and postintervention outcomes, they found that child abuse screening and detection rates increased significantly after implementation of the new checklist and training.

Source: Eveline C. F. M. Louwers, Ida J. Korfage, Marjo J. Affourtit, Dop J. H. Scheewe, Marjolijn H. van de Merwe, Anne-Françoise S. R. Vooijs-Moulaert…Harry J. de Koning, "Effects of Systematic Screening and Detection of Child Abuse in Emergency Departments," *PEDIATRICS* 130, no. 3 (2012): 457–64, https://doi.org/10.1542/peds.2011-3527.

When you're looking at outcomes over time, you may see trends in the preintervention period. Another advantage to the ITS model is that it accounts for these preintervention trends (Kontopantelis et al. 2015). The ITS model assumes that this preintervention trend would have continued as is if the intervention had not been implemented. ITS differs from other designs in that it uses a before-after comparison within a single population instead of a comparison group during the same period. This limits selection bias and confounding because of differences between the treatment and comparison groups (Lopez Bernal, Cummins, and Gasparrini 2018). Some cofounding variables stay fairly consistent over time in populations. These include characteristics such as the distribution of age or socioeconomic status in a population. ITS designs are generally unaffected by slow population shifts of this nature. But ITS is sensitive to any rapid population, context, or system changes (such as a recession, natural disaster, or major changes in the broader service system) (Lopez Bernal, Cummins, and Gasparrini 2017).

A disadvantage of ITS designs is that they require a very specific implementation context. Many interventions do not have a clear date that distinguishes the preintervention period from the postintervention period. Implementation may take place over several months or years. And it may take time for the intervention to reach the full target population or be implemented fully with fidelity.

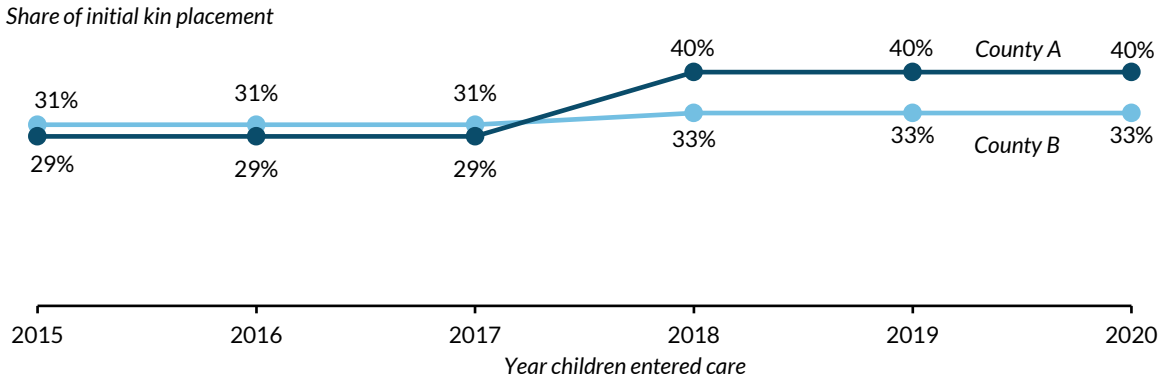
Difference-in-Difference (DiD)

DEFINITION

DiD works by comparing change over time in outcomes for two groups—a treatment group and a comparison group. People who received an intervention make up the treatment group. People who did not receive the intervention are in the comparison group. Like an ITS design, you look at outcomes both before and after you implement an intervention. But, unlike ITS, you have a group receiving services as usual during the posttreatment period.

FIGURE 3

DiD Example—Kinship Placement Rate



Source: The author developed this figure to display a hypothetical example.
Notes: This figure looks at the share of children entering out-of-home care into a kinship placement in counties A and B. In 2018, county B implemented an intervention that increased support and capacity for efforts to find kin for children coming into care.

In figure 3, we track a population-level outcome over six years in two counties (the initial kinship placement rate). County A implemented no initiatives to increase kinship placement. Children coming into care in county A experienced services as usual throughout the six-year period. In 2018, county B implemented an intervention to increase the level of kinship placements in their county. Around the same time in 2018, county A saw a 2 percent increase in kin placements. County B saw an 11 percent increase. A DiD design would look at the difference of those two differences. County B's change (11 percent) minus county A's change (2 percent) equals a 9 percent change attributable to the intervention in county B.

BOX 3

DiD Statewide Evaluation Example—Triple P in North Carolina

In 2011 and 2012, North Carolina implemented the Positive Parenting Program (Triple P) in a third of their 100 counties. Researchers used a DiD design to compare outcomes between the treatment and comparison counties, before and after implementation. They examined child maltreatment and placement rates in all 100 counties from 2008 to 2015. The evaluation found that implementation of Triple P in treatment counties was associated with lower rates of child maltreatment and placement.

Source: Samantha Schilling, Paul Lanier, Roderick A. Rose, Meghan Shanahan, and Adam J. Zolotor, "A Quasi-Experimental Effectiveness Study of Triple P on Child Maltreatment," *Journal of Family Violence* 35, no. 4 (2020): 373–83, https://doi.org/10.1007/s10896-019-00043-5.

EVALUATION OPPORTUNITY

A DiD design may work in situations where the intervention was implemented at different times across sites (in states, counties, offices, or providers). For example, a state implements a new intervention in some counties before implementing it in others. Or a county implements a new program with some providers but not all.

ASSUMPTIONS AND CONSIDERATIONS

The primary assumption in a DID design is the "parallel trends" assumption. Under this assumption, the trend in outcomes in both treatment and comparison sites is similar before the intervention (going up, going down, or staying steady). And the assumption is that any differences in outcomes between the treatment and comparison groups would have been the same over time without the intervention. For example, in figure 3, we assume that without county B's intervention, county B's outcome would have changed by the same 2 percent as (and parallel to) county A's outcome.

DiD designs have several other requirements:

- The choice of sites to receive the intervention should be unrelated to the outcome. For example, treatment sites should not be chosen based on the perceived strengths or needs of the people or providers there.
- There should be no major differences between the treatment and comparison sites that may impact outcomes. For example, problematic site differences could arise because of differences in geography. If all the treatment sites were urban counties while all the comparison sites were rural counties, the geographic differences between the counties might impact outcomes.
- DiD requires outcome data measured from treatment and comparison sites at two or more different periods. You need to measure outcomes at least once before the intervention begins and at least once after treatment is implemented.
- Regular demographic and other characteristic data must be available pre- and postintervention as well. This is because you must examine the composition of the population both before and after intervention. The population characteristics in the treatment and comparison sites should remain constant over time. Like an ITS model, a basic DiD model cannot handle changes in population composition over time, but some statistical methods can handle that situation if it arises.

ADVANTAGES AND DISADVANTAGES

An advantage of DiD is that it can make a strong case for a causal effect of an intervention when it meets the assumptions. Like ITS, DiDs can take advantage of administrative data because they focus on population-level outcomes. Another advantage of a DiD is that the treatment and comparison groups can start at different levels of the outcome. Under the parallel trends assumption, it is important that the two groups have similar outcome trends in the pre-period (such as the outcome going up, going down, or staying stable). But the two groups do not have to have outcomes at the same level. For example, the treatment group could have different permanency rates than the comparison group in the preperiod.

DiD's requirement of a comparison site can be both an advantage and a disadvantage. A well-chosen comparison site can improve the strength of your findings. But there are restrictions around the choice of comparison site(s). You cannot use DiD if the treatment assignment (the intervention implementation rollout) is based on which site is most "ready," "able," or "disadvantaged." Also, you will need baseline population and outcome data for both treatment and comparison sites.

Regression Discontinuity (RDD)

DEFINITION

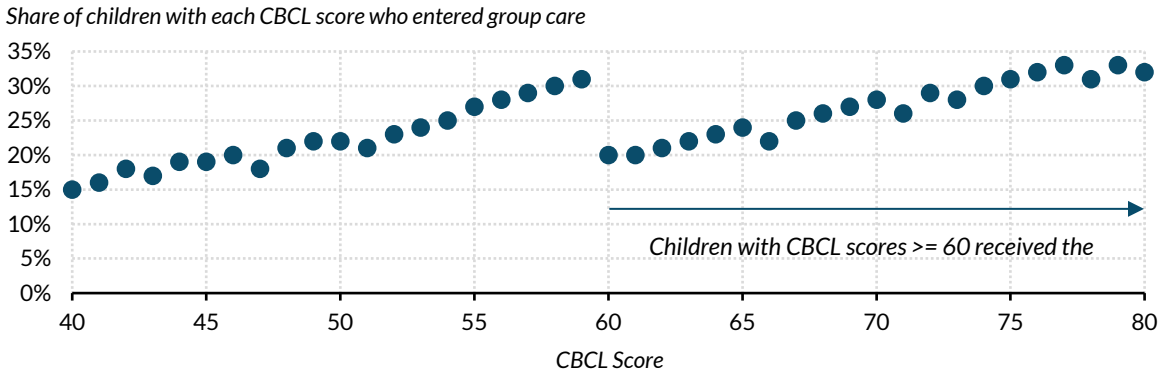
You can use RDD when an intervention (treatment) is assigned based on a cutoff score. RDD compares outcomes for the individuals whose scores are on either side of the cutoff to show the intervention's effect. This cutoff score may come from a measure such as a risk or functional assessment. This measure is called an assignment variable, because you use it to assign an individual to an intervention. In an RDD, you compare the outcomes of the group whose scores are immediately above the cutoff (assigned to the treatment) with the outcomes of the individuals whose scores are immediately below the cutoff (the comparison group). The difference between the average outcomes of those two groups is the impact of the intervention on the outcome.

In figure 4, we present a hypothetical example where treatment (wraparound services) is assigned based on a cutoff score. In this scenario, you would assign any child in care who scores 60 or higher on the Child Behavior Checklist (CBCL) to a wraparound services intervention (Leslie et al. 2000). An evaluation using an RDD compares the outcomes for children right above and below this cutoff point. Here, we see that children right above the cutoff score of 60 show a decreased likelihood of placement in group care compared with the children who scored below 60 and were not offered the intervention. For example, in the comparison group, 31 percent of the children who scored 59 on the CBCL entered group care. In the treatment group, only 20 percent of the children who scored 60

entered group care. Presumably, these children's needs are very similar, and the difference in outcomes is because of the intervention's impact.

FIGURE 4

RDD Example—Group Care Placement Rate by CBCL Score



Source: The author developed this figure to display a hypothetical example.
Notes: This figure looks at the share of children entering group care because of a disruption in a family placement. Each dot stands for a group of children who received a certain Total Problem Score on the CBCL (Leslie et al. 2000). A CBCL score of 60 is used as a cutoff score, above which children are assigned to a wraparound services intervention meant to reduce placement disruptions and group care use.

EVALUATION OPPORTUNITY

RDDs have not been widely used in child welfare, but the opportunity exists. RDDs can take advantage of an intervention that uses a measure to determine treatment assignment. You may be able to use an RDD if you use standardized measures for assessment purposes to decide whether a client receives a particular intervention. For example, caseworkers may use a risk assessment at investigation. They then might use that score to assign families at or above a cutoff value to a specific service.

ASSUMPTIONS AND CONSIDERATIONS

An RDD design requires some key assumptions:

- The measure used to assign people to the intervention should be continuous around a cutoff point. A continuous measure captures information on a continuum or scale. This means that the scoring should be without jumps or interruptions.
  - » The cutoff point, or threshold, should be clearly defined.
  - » The assignment variable is not caused or influenced by the treatment. In other words, whatever variable you use to assign treatment needs to be measured before the treatment starts (Jacob, Somers, and Bloom 2012).

    - » The assignment variable should include enough values below and above the cutoff point. Some clearinghouses recommend at least four values below and four values above the cutoff (Sama-Miller et al. 2020; Schochet et al. 2010).
    - » There must also be a large enough sample size around the cutoff. It is that part of the sample—people right around the cutoff point—from which you will calculate the intervention's effect.
- The characteristics of individuals close to the cutoff point should be similar (such as their demographics and level of need). The only difference between them should be whether they are receiving treatment or not.
- Besides the intervention, there are no other relevant ways you treat the treatment and comparison groups differently (Jacob, Somers, and Bloom 2012).
- The scoring of the assessment should be "blind" to treatment (Lee and Lemieux 2010)—meaning that whoever is scoring the assessment should do so without considering treatment assignment. Thinking back to our example in figure 4, caseworkers should not adjust CBCL scores over the 60-point cutoff to make sure certain children receive wraparound services (treatment).

ADVANTAGES AND DISADVANTAGES

Under the right implementation conditions, an RDD can take advantage of existing program and administrative data. Being able to conduct an RDD in an existing implementation context can enhance your study's external validity. Higher external validity means you can better apply results from your study to other settings. But RDDs have some disadvantages as well. As we've said, an RDD estimates the impact of an intervention based on differences in outcomes for the treatment and comparison group members right around the cutoff point. So RDD findings do not always apply to individuals with scores further away from the cutoff point. This limits how much you can generalize the findings of even a well-conceived RDD. Also, it may be difficult to have an adequate sample size around the cutoff value.

Conclusion

Establishing evidence of which programs work in child welfare and which do not is important. This information helps us provide effective services so children can grow up in safe and nurturing environments. You can use impact evaluations to test the effectiveness of a well-defined intervention. Impact evaluations can be RCTs or QEDs. In this report, we focused on QEDs and the opportunities they

present to generate rigorous evidence. We explained their structure, discussed how they can establish causality, and addressed their advantages and disadvantages. We also presented the basics of four common types of QEDs: matched-group, ITS, DiD, and RDD.

For child welfare administrators, we hope this report helps you think about how you can use a QED to evaluate programs in your child welfare system. Not every impact evaluation needs to be an RCT. But no matter the evaluation type, following best practices in evaluation design will increase the credibility of your findings. This report cannot cover all components that are important to consider when designing a QED evaluation in child welfare. We encourage you to use this information as a starting point. Good evaluation design goes beyond the basics and is always rooted in a deep understanding of the intervention's theory of change and service context. We recommend that those looking to evaluate child welfare programs consult with methodological experts about their specific context. Please also see our guides on best practices for designing evaluations and reporting evaluation findings for more detailed guidance (Brewsaugh and Prendergast 2022; Prendergast and Brewsaugh 2022).

Appendix A. Common QEDs Summary

The table below summarizes the basic elements, assumptions, and opportunities for each of the four common QED types reviewed in this report.

	Matched-group	Interrupted time series (ITS)	Difference-in-difference (DiD)	Regression discontinuity (RDD)
Definition	A matched-group comparison QED matches members of the treatment group with members of the comparison group. This is done either via exact matching or statistical matching.	An ITS tracks a population-level outcome over time to establish an underlying trend in that outcome. Then, ITS assesses whether the implementation of an intervention "interrupted" that trend.	DiD works by comparing change over time in outcomes for two groups: (1) people in areas who received an intervention and (2) people in other areas who did not receive the intervention.	When an intervention treatment is being assigned based on a cutoff score, RDD compares outcomes for the individuals whose scores are on either side of the cutoff to show the effect of the intervention.
Comparison group	A group of similar individuals who are not receiving treatment. Matching variables are used to match someone in the treatment group to someone who looks like them in the comparison group.	The population before the intervention acts as the comparison group.	People in other areas who did not receive the intervention make up the comparison group.	People whose scores on the assignment variable are right below the cutoff make up the comparison group.
Primary assumptions	A matched-group QED depends on the treatment and comparison groups' similarity at baseline, so you must confirm the baseline equivalence of the two groups.	▪ Any trends in the outcome before the intervention need to be linear. ▪ Any trends in the outcome before the intervention would have continued unchanged without the intervention.	▪ The trend in outcomes in both the treatment and comparison sites is similar before the intervention. ▪ The choice of sites to receive the intervention should be unrelated to the outcome. ▪ There should be no major differences between the treatment and comparison sites that may impact outcomes.	▪ The measure used to assign people to the intervention should be continuous around a cutoff point. ▪ The average characteristics of individuals close to the cutoff point should be similar. ▪ Besides the intervention, there are no other relevant ways you treat the treatment and comparison groups differently. ▪ The scoring of the assessment should be "blind" to treatment.

Evaluation opportunity	A matched-group design works when there is a group of individuals *similar* to the treatment group but who are not receiving treatment.	An ITS design may be appropriate when there is a large-scale intervention or policy implementation, with a clearly defined start date, targeting population-level outcomes.	A DiD design may work in situations where the intervention was implemented at different times across sites (such as in states, counties, offices, or providers).	An RDD can take advantage of an intervention that uses a measure to determine treatment assignment.

Appendix B. Glossary

This glossary lists common terms used in this report.

Baseline equivalence	The extent to which the treatment and control/comparison groups were similar to one another when the evaluation began. Equivalence is achieved when there are no statistically significant differences between the treatment and comparison groups on key measures (such as demographics, child welfare history, pretests, or other characteristics) at the start of the evaluation.
Causality	The logical process used to draw conclusions from evidence about what has been produced or "caused" by a program. To say a program produced or caused a certain result means that if the program had not been there (or if it had been there in a different form or to a different degree), then the result found (or level of result) would not have happened.
Comparison group	A group not exposed to a program or treatment—referred to as a control group in the case of random assignment.
Confounding variable	A confounding variable is an "extra" variable that you did not account for when looking at the effect of an intervention on an outcome. A confounding variable can make it look like an intervention caused an outcome when that is not true.
Difference-in-difference (DiD) design	Difference in difference (DID) is a statistical approach that tries to mimic experimental design using observational (available) data. It estimates the effect of an intervention on an outcome by comparing the average change over time in the outcome variable for the treatment group with the average change over time for the control group.
Exact matching	A method of creating groups in which every person in the treatment group is exactly matched to a person in the comparison group on every variable used to conduct the match.
External validity	The ability to generalize findings about the program from a specific evaluation to how the program would perform in other locations, providers, or populations. For example, findings about a program evaluation in a rural setting or only with English speakers may not be generalized to urban settings or non-English speakers.
Fidelity	A term used to describe if all program activities are delivered consistently with the intended quantity and quality according to the model.
Impact evaluation	A type of evaluation used to assess whether a program causes a desired outcome—for example, whether a parenting program causes a reduction in child maltreatment.
Internal validity	The ability to argue a program has caused the outcomes observed rather than alternative explanations.
Internal validity threats	Anything other than the program that is associated most strongly with either the treatment or control/comparison group and can provide a plausible alternative explanation for the outcome.
Interrupted time series (ITS)	A time series is a set of repeated measurements taken over time for a group of people. In an interrupted time series (ITS) evaluation, the time series of an outcome of interest is used to find an underlying trend, which is interrupted by an intervention at a specific point in time. ITS is used to attribute change over time in a population-level outcome to an intervention.
Matched-group QED	A method of forming groups by matching people in the treatment group to people in the comparison group based on how similar they are on baseline characteristics.
Primary data	Data collected directly from a data source (such as from a participant) for the evaluation.
Quasi-experimental design (QED)	An evaluation design where group assignment is decided through some nonrandom process, such as matching, cutoff scores, or time.
Randomized controlled trial (RCT)	A study design where group assignment is decided through a random process, such as a lottery, dice roll, or computer algorithm.

Regression discontinuity design (RDD)	A regression discontinuity design (RDD) is a QED where group assignment is based on a cutoff score on a measure (such as an assessment of behavior) above or below which participants are assigned to treatment or comparison groups.
Sample size	The number of people from the program's target population who will be study participants.
Selection bias	Selection bias occurs when individuals are assigned to groups based on some characteristic that can predispose them to a particular outcome, such as by manipulating random assignment. Selection bias threatens the internal validity of program evaluations whenever selection of treatment and control groups is done nonrandomly.
Statistical matching	A method of creating groups where people in the treatment and comparison groups are matched using some type of statistical method, such as propensity score matching.
Target population	The group of people whom the program is meant to serve.
Theory of change	A statement about why and how a particular program is expected to bring about change.
Treatment group	The group of people that receives the program—also referred to as the experimental or program group.
Variable	An attribute or characteristic that describes an individual, group, or system that can change and have different values from one individual, group, or system to another—for example, gender, educational level, income.
Well-defined intervention	A well-defined intervention has a clear theory of change and a manual or other comprehensive, written guidance describing how it should operate.

References

Austin, Peter C. 2011. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." *Multivariate Behavioral Research* 46 (3): 399–424. https://doi.org/10.1080/00273171.2011.568786.

Brewsaugh, Katrina, and Sarah Prendergast. 2022. *Ten Key Design Elements for Rigorous Impact Evaluations in Child Welfare: A Desk Reference for Evaluators*. OPRE Report #2022-171. Washington, DC: Urban Institute.

CEBP (Coalition for Evidence-Based Policy. 2014. "Which Comparison-Group ('Quasi-Experimental') Study Designs Are Most Likely to Produce Valid Estimates of a Program's Impact?: A Brief Overview and Sample Review Form." Washington, DC: CEBP.

Hanita, Makoto, Dana Ansel, and Karen Shakman. 2017. "Matched-Comparison Group Design: An Evaluation Brief for Educational Stakeholders." Waltham, MA: Education Development Center.

Hanson, Devlin, and Michael Pergamit. 2022. *Conducting a Randomized Controlled Trial (RCT) in Child Welfare: A Guide to What, Why, and How for Child Welfare Agency Staff*. Washington, DC: Urban Institute.

Jacob, Robin, Pei Zhu, Marie-Andrée Somers, and Howard Bloom. 2012. *A Practical Guide to Regression Discontinuity*. New York: MDRC.

JBA (James Bell Associates). 2013. *Conducting Randomized Controlled Trials in Child Welfare Practice Settings: Challenges and Solutions.* Arlington, VA: JBA.

Kontopantelis, Evangelos, Tim Doran, David A. Springate, Iain Buchan, and David Reeves. 2015. "Regression Based Quasi-Experimental Approach When Randomisation Is Not an Option: Interrupted Time Series Analysis." *BMJ* 350:h2750. https://doi.org/10.1136/bmj.h2750.

Lee, David S, and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48 (2): 281–355. https://doi.org/10.1257/jel.48.2.281.

Leslie, Laurel K., John Landsverk, Roxanne Ezzet-Lofstrom, Jeanne M. Tschann, Donald J. Slymen, and Ann F. Garland. 2000. "Children in Foster Care: Factors Influencing Outpatient Mental Health Service Use." *Child Abuse and Neglect* 24 (4): 465–76. https://doi.org/10.1016/S0145-2134(00)00116-2.

Lopez Bernal, James, Steven Cummins, and Antonio Gasparrini. 2017. "Interrupted Time Series Regression for the Evaluation of Public Health Interventions: A Tutorial." *International Journal of Epidemiology* 46(1): 348–55. https://doi.org/10.1093/ije/dyw098.

———. 2018. "The Use of Controls in Interrupted Time Series Studies of Public Health Interventions." *International Journal of Epidemiology* 47 (6): 2082–93. https://doi.org/10.1093/ije/dyy135.

Moser, Cindy, and Uyen Vu. 2017. *What Researchers Mean by...Easy-to-Understand Definitions of Common Research Terms in the Health and Social Sciences.* Toronto, Ontario: Institute for Work and Health.

OPRE. 2018. *The Program Manager's Guide to Evaluation, Second Edition*. Washington, DC: HHS, ACF, OPRE.

Packard Tucker, Laura, and Xiaomeng Zhou. 2022. *Administrative Data in Child Welfare Evaluations: Using Administrative Data to Understand Populations and Measure Outcomes*. OPRE Report #2022-26. Washington, DC: Urban Institute.

Prendergast, Sarah, and Katrina Brewsaugh. 2022. *A Guide to Writing High-Quality Evaluation Reports in Child Welfare*. OPRE Report #2022-43. Washington, DC: Urban Institute.

Sama-Miller, Emily, Julieta Lugo-Gil, Jessica Harding, Lauren Akers and Rebecca Coughlin. 2020. *Home Visiting Evidence of Effectiveness (HomVEE) Systematic Review: Handbook of Procedures and Evidence Standards, Version 2.*

OPRE Report # 2020-151. Washington, DC: US Department of Health and Human Services (HHS), Administration for Children and Families (ACF), Office of Planning, Research, and Evaluation (OPRE).

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J.R., Porter, J., and Smith, J. 2010. "Standards for Regression Discontinuity Designs." Washington, DC: What Works Clearinghouse.

Walsh, Cambria, Jennifer Rolls Reutz, and Rhonda Williams. 2015. *Selecting and Implementing Evidence-Based Practices: A Guide for Child and Family Serving Systems*, 2nd ed. San Diego: California Evidence-Based Clearinghouse for Child Welfare.

Welti, Kate, Esther Gross, Alexandria Wilkins, and Karin Malm. 2018. "Evaluation of the Upfront Family Finding Pilot." Bethesda, MD: Child Trends.

Wilson, Sandra J., Cristofer S. Price, Suzanne E. U. Kerns, Samuel D. Dastrup, and Scott R. Brown. 2019. *Title IV-E Prevention Services Clearinghouse Handbook of Standards and Procedures, version 1.0.* OPRE Report # 2019-56. Washington, DC: HHS, ACF, OPRE.

White, Howard, and Shaun Sabarwhal. 2014. "Quasi-Experimental Design and Methods." Florence: UNICEF Office of Research.

Wulczyn, Fred, Lily Alpert, and Kerry Monahan-Price. 2016. "Research Evidence Use by Child Welfare Agencies." Minneapolis: CW360 Child Welfare Reform.

Wulczyn, Fred, Lily Alpert, Britany Orlebeke, and Jennifer Miller Haight. 2014. "Principles, Language, and Shared Meaning: Toward a Common Understanding of CQI in Child Welfare." Chicago: Center for State Child Welfare Data, Chapin Hall at the University of Chicago.

About the Authors

Laura Packard Tucker is a senior research associate at the Urban Institute in the Center on Labor, Human Services, and Population. Her research focuses on issues related to child welfare, with an emphasis on fiscal analysis, program evaluations, and transition-age young people. Through her work at Urban, she is committed to improving the ways communities and government come together to support children, young people, and families. Packard Tucker holds an MS in financial analysis from Portland State University.

Statement of Independence

The Urban Institute strives to meet the highest standards of integrity and quality in its research and analyses and in the evidence-based policy recommendations offered by its researchers and experts. We believe that operating consistent with the values of independence, rigor, and transparency is essential to maintaining those standards. As an organization, the Urban Institute does not take positions on issues, but it does empower and support its experts in sharing their own evidence-based views and policy recommendations that have been shaped by scholarship. Funders do not determine our research findings or the insights and recommendations of our experts. Urban scholars and experts are expected to be objective and follow the evidence wherever it may lead.

500 L'Enfant Plaza SW
Washington, DC 20024

www.urban.org