



Evaluation Technical Assistance Brief

DECEMBER 2017 • NUMBER 3

Sarah Avellar, Robert Santillano, and Debra Strong

Tips for Planning an Impact Evaluation

The question of how to best protect vulnerable children, stabilize families, and support healthy parenting is one without easy answers. Even if stakeholders strongly believe in a program and families like it, an impact evaluation is the only way to know if a program is effective in achieving its ultimate goals.¹ An impact evaluation must have a program group—which is offered program services—and a comparison group—which is not—to test whether outcomes change, and by how much, solely because of the program. Rigorous impact evaluations are increasingly preferred or even required by funders because they are key to learning what works for families.

Impact evaluations, though important, can be challenging. This brief provides tips—drawing on the experiences of previous cohorts of RPG grantees—about how to plan an impact evaluation. The tips focus on impact evaluation designs in which groups were formed either through a randomized controlled trial (RCT) or a quasi-experimental design (QED). The discussion of these two strategies highlights pitfalls and proposed solutions.

Why consider impact evaluation designs?

The primary reason to design a meaningful evaluation is to capture changes that occur. Children grow up, circumstances evolve, and attitudes shift. The goal of an impact evaluation is to isolate the changes that a program caused in people apart from changes that took place for other reasons and would have occurred even without the program. Funders and practitioners want to know what changes happened because of a program.

Who should read this brief?

The Children's Bureau funded this brief for groups that receive a Regional Partnership Grant (RPG) or other grants and want to conduct an evaluation that can identify the causal impacts of a program designed for families involved with—or at risk of involvement with—child welfare. It draws on the experience of previous RPG grantees to identify challenges of rigorously evaluating a program and presents ways to address them. Grantees may want to use these tips when working with their evaluator to design and conduct an impact study.

Isolating changes caused by a program from other influences requires a credible comparison group—one that is initially as similar to the program group as possible. If the groups have similar characteristics and situations at the beginning of the study but only one group has access to the program being tested, later differences in outcomes between the two groups were likely caused by the program.

Evaluation designs vary in how well they can estimate a program's effects. With an RCT, people are first enrolled in the study and then randomly assigned (for example, with a coin flip) to the program or comparison groups. Because the process is random, the groups will be similar, on average, at the beginning of the study. This similarity means well-executed RCTs can provide the strongest evidence of effectiveness.

QEDs include a wide range of designs, but in all of them, the people in the study are in

MATHEMATICA
Policy Research



RPG

Regional Partnership Grants
and Cross-Site Evaluation

¹ <https://www.cdc.gov/std/Program/pupestd/Types%20of%20Evaluation.pdf>.

What is RPG?

The RPG program supports partnerships between child welfare agencies, substance use disorder (SUD) treatment providers, and other systems to address the needs of children who are in, or at risk of, out-of-home placement due to a parent's or caretaker's SUD. The grant funder is the Children's Bureau within the Administration on Children, Youth, and Families; Administration for Children and Families; U.S. Department of Health and Human Services.

The legislation that authorizes the partnerships requires the agencies to collect and report on a set of performance measures. The Children's Bureau also requires partners to evaluate their programs and participate in a national cross-site evaluation (Administration for Children and Families 2012, 2014).

the program or comparison groups through a process that is not random. For example, a program group may be made up of families in the grantee's service area and the comparison group might be drawn from a different location that is not implementing the program. Previous grantees that conducted a QED have typically used a matched comparison design in which families for the comparison group are selected to have similar characteristics to those in the program group. In all QEDs, there is always some uncertainty about the initial similarity of the groups because researchers use a non-random process. It is not possible to measure and include in the analysis all characteristics of the groups, so the potential for unmeasured differences remains.



Design tip: Balance the strength of evidence an evaluation design can produce with practical factors that can affect whether the design can be implemented successfully

Each design has pros and cons (Table 1). RCTs can estimate program effects better than QEDs, but stakeholders might be resistant. QEDs may seem more palatable because the study typically does not affect who receives services. However, implementing a QED requires finding a credible comparison group where the groups can be shown to be similar at the start, which can also be difficult. Knowing the trade-offs to each design can be helpful when considering what to use for an evaluation. The remainder of this brief provides additional tips for each design.

Tips for planning randomized controlled trials (RCTs)

Although RCTs can provide the strongest evidence for assessing program impacts, if a study is to be successful, stakeholders must be willing to participate. Stakeholders who

Table 1. Strengths and challenges for comparison group strategies

Consideration	RCT	QED
Strength of evidence	Strong evidence: If executed well, can answer whether a program is effective in achieving its goals	Moderate evidence: Can never prove that program caused results
Acceptability to stakeholders	Moderate to difficult: Need for ongoing efforts for continued stakeholder buy-in	Easy to moderate: Likely less resistance from stakeholders, but may be difficult to recruit agency to provide comparison group
Ability to form similar program and groups	Easy to moderate: Built into strategy, but attrition must be low to maintain benefits of design	Moderate to difficult: Similar families often hard to find, and similarity difficult to demonstrate
Recruiting the desired number of study sample members	Moderate to difficult: Most programs need to increase recruitment to have enough people to form a program group (that fills all available spaces in a program) and a comparison group	Easy to difficult: Depends on number of people in target populations who program serves and comparison settings and other circumstances

are not fully informed about the importance and value of the RCT may not support it. For example, in a previous RPG evaluation, a state official required a grantee to drop random assignment because of concerns about negative publicity resulting from perceptions that families were being denied services. In another example, at one grantee, staff believed they received fewer referrals for services because the referring agencies feared that needy families might not receive any help. The sample size issue was so severe the grantee ended random assignment.²

Some stakeholders have ethical concerns about random assignment. These kinds of concerns must be regarded as important and valid, and planning should proactively raise and address them.

In both cases, not all key stakeholders were on board, even though the programs represented good candidates for a random assignment evaluation. Based on grantees' experiences, three tips and related strategies could improve the chances of successfully implementing an RCT.



RCT Tip 1: Anticipate concerns and proactively address them in the design

Some stakeholders have ethical concerns about random assignment. These kinds of concerns must be regarded as important and valid, and planning should proactively raise and address them. Consider the following common concerns and possible design solutions:

RCT Concern 1: Families in the control group are being denied services. Several design options can address this concern. The first is to offer standard best practice care to the comparison group. Many grantees are trying something new and innovative that may or may not be more effective than standard or typical services. In these cases, traditional services could be offered—through the grantee or another agency—to the comparison group. A second option is to offer them a program with an alternative focus. For example, if the services being studied provide parenting skills, the alternative could be treatment for a SUD (see box). A third option is to vary the intensity of the program. Intensive services are usually costly, and it is hard to know the best dosage without testing. So the comparison group can

still receive services, but with a lighter touch in terms of length or frequency. A fourth option is to offer a waiting list. Families randomly assigned to the comparison condition would be offered the service after both groups have provided follow-up data. This would be a viable option, but with two considerations. First, longer-term outcomes usually cannot be captured in a waiting-list design. Typically, the wait-listed group is offered services after the first group completes them and both groups have provided data. Thus, expected changes must occur about when services end. Second, because the wait-listed group cannot receive services right away, the program must still be valuable for families even if they do not immediately receive them.

A caution when considering what services to offer to the comparison group is that the more similar the services that the program and comparison groups receive, the smaller the expected effect. The smaller the expected effect, the larger the sample size necessary to detect it statistically. The biggest expected effect is between intensive services and no services.

Using service order to form program and comparison groups

Parents often have multiple needs: for example, SUDs and underdeveloped parenting skills. Families could be randomly assigned to receive either substance abuse treatment or parenting skills services first. At the end of those services, the effects of each type could be examined. For example, are the parents who received services to address their SUD using substances less frequently than those in the other group? Conversely, have the parents in the parenting group improved their skills more than those in the SUD group? Each group could then switch treatment type. Results after the switch could be examined as well. For example, did the results differ depending on the order in which parents received treatment?

² Obtaining a sufficient sample size is often a challenge for evaluations. In RCTs, the challenge is greater because not only is it important to enroll a sufficient number of people in the study, the quality of the evaluation can depend on keeping them in the study (that is, collecting data on all sample members). A companion brief describes ideas for obtaining and maintaining the desired sample size. For more information, see Enrolling and Retaining Evaluation Participants: RPG Evaluation Technical Assistance Brief.

RCT Concern 2: If the program denies some families, available program slots will be wasted. A program might not recruit enough study participants to assign half to the comparison group and still fill all available program spaces. The first approach, as discussed in the companion brief, is to increase recruiting efforts, but another solution is to alter random assignment probabilities so programs operate at capacity. For example, the study could assign 60 percent of participants to the program and 40 percent to the comparison group. Evaluators can adjust these probabilities over time to reflect changes in service demand.

 **RCT Tip 2: Present a strong rationale to stakeholders for an RCT**

It might be necessary to show stakeholders, including program staff, referral sources, agency leaders and others, that an RCT is appropriate and worth doing. Consider the following reasons that directly address common concerns:

- **Fairness.** If the number of families who need services exceeds available slots, random assignment is an equitable way to allocate openings because everyone has the same chance of getting services. For example, if random assignment is 50/50, the first family identified for services has a 50/50 chance of receiving them, as does the 1,000th family.
- **A long-term focus on the mission.** Stakeholders often assume a new program model will be effective. However, many programs do not work better than usual services or, in rare cases, may even harm participants. This could be true even of programs that have already been evaluated in other settings, different target populations, or with meaningfully different implementation. Without an impact evaluation, providers and funders do not know if families are actually being helped by the new service or whether the comparison service is just as good or better. That understanding is essential to investing resources in what works best and improving future outcomes for many more families.

- **Identifying the best alternative.** In cases in which the comparison group participates in an alternative program (see previous sections for ideas), everyone is being served. The goal is to find the most effective alternative for families and the program.
- **A willingness to implement and respond to lessons from RCTs can attract future funding.** Willingness and capacity to conduct rigorous evaluations and make changes in programs based on evidence may help attract future funding to an organization. Funders increasingly want evaluations of services and evidence of effectiveness and appreciate evidence that organizations seeking funds have the capacity to conduct meaningful evaluations. The long-term sustainability of services may even depend on evidence of effectiveness from an RCT.

 **RCT Tip 3: Continually work with stakeholders to assess their ongoing support and address new or emerging concerns**

Change is constant not only for families but also agencies. New challenges can emerge over time. Staff turnover or new leadership can mean a referral source becomes less supportive of an evaluation. Regular monitoring and reminding of the benefits of the study design, or reminders that alternative services will still be available will help to maintain continuity of the evaluation.

Tips for planning quasi-experimental designs (QEDs)

A strong QED provides moderate evidence of a program's impacts (that is, the evidence of program effects is not as strong as an RCT, but still highly valuable). When an RCT is not possible, a well done QED is an excellent alternative. However, QEDs also have potential pitfalls. Comparison families must be similar to those you are serving, but receiving different services. Past grantees have sought referrals from other agencies to identify possible comparison group participants, which can be very time-consuming. To reduce that burden, a few grantees tried to partner with a single agency that served a similar population but provided substantially different services

A strong QED provides moderate evidence of a program's impacts (that is, the evidence of program effects is not as strong as an RCT, but still highly valuable). When an RCT is not possible, a well done QED is an excellent alternative.

than the grantee. However, most of the grantees could not find such a match. Further, even among those that did, the comparison group agency had little incentive to participate and some withdrew from the study. Based on these experiences, we have identified four tips to consider.

 **QED Tip 1: Allow enough time to find a suitable comparison group**

Setting up a comparison group requires a substantial commitment of time and resources. In RPG, some grantees wanted to focus first on their services before turning to the task of selecting a comparison group and initiating the evaluation. They eventually found that they could not get a comparison group together in time to meet the evaluation requirements.

 **QED Tip 2: Provide incentives to agencies that provide comparison group services and data**

Working with a partner agency to form a comparison group rather than trying to gather referrals from many sources can have many advantages for the evaluation, but benefits for that agency must be built into the design. As previously described, some partner agencies in RPG withdrew from the evaluation over time. Incentives might be monetary to offset participation costs, an exchange of services for clients (as long as this does not interfere with the evaluation), providing data about the clients, or some other arrangement. Whatever it is, minimizing the financial burden and demands on staff time on a comparison site may increase their cooperation (see QED Tip 3).

 **QED Tip 3: Direct contact by the grantee and evaluator with comparison group members is best**

Others might not be as invested in an evaluation as the grantee and its evaluator. It is often best to have grantee or evaluator staff interact directly with potential comparison group members. One RPG grantee had to wait for comparison group members to make contact after another party gave them information about the study. Unfortunately, very few people followed up. Other grantees planned to have

the agency staff serving the comparison group collect data from them, but the staff did not prioritize data collection. Consider collecting the data directly to reduce burden on your partner and increase response rates.

 **QED Tip 4: Consider less commonly used QED approaches**

As previously described, matched comparison group designs, in which program and comparison groups are selected based on their initial characteristics, are very common. But alternatives to matching QEDs are also an option grantees can discuss with their evaluators. Here are two examples:

Difference-in-differences (DID). This design compares two types of changes or differences to estimate effects. The first difference comes from comparing outcomes for families before and after a program is introduced. For this first comparison, the “before” families would have been eligible for the program but did not receive it because it was not yet offered. The “after” families were offered the program. The second difference compares “before” and “after” families who had never been eligible for focal program services—perhaps because they live in geographic area where the program was not offered.

Unlike a matched QED, in which the same people are in the “before” and “after” phases, difference-in-differences designs create groups based on time and other factors, such as geographic location or age. Because of this, different people may be in the “before” and “after” phases.

The design can capitalize on events outside of the evaluation, such as introduction of a program in one location but not others (see a hypothetical example on the next page). But there are two primary trade-offs.

DID Challenge 1. Data requirements.

First, this requires data on outcomes and characteristics for an extended period of time before and after program introduction. The evaluation should have at least two—but preferably more—serial observations in the time periods before the program is introduced (Gertler et al. 2011). Data for an

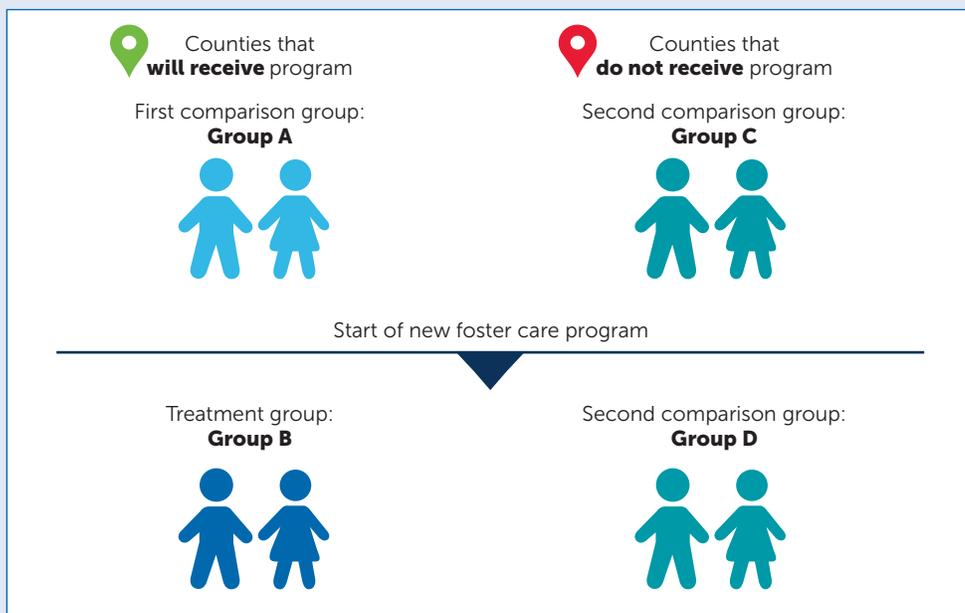
Others might not be as invested in an evaluation as the grantee and its evaluator. Whenever possible, try to interact directly with potential comparison group members.

Hypothetical example of difference-in-differences design

Suppose a new foster care program was introduced to a subset of counties on January 1 of the current year. All of the children in foster care in those counties before the change could be part of the first comparison group (A). All of the children in foster care in those counties after the change would represent the program group (B). The data should cover several years before and after the program's launch.

A second comparison group could be made up of children in foster care during the same time periods, but in the counties that did not have access to the program (group C before the program and group D after program). These children should not be affected by the introduction of the program. This would further strengthen the evidence, because if some other change during the evaluation affected the entire state—such as a new policy or economic recession—including this second comparison should help factor that out of the estimate.

The analysis uses a difference-in-differences calculation. For this example, that calculation could be changes in child welfare administrative data on reunification rates for the counties that got the program (B–A), and changes for counties that did not get the program (D–C). The difference-in-differences estimate then subtracts the non-program county changes from the program county changes to estimate the effects of the program. That is: $(B-A) - (D-C)$.



extended period of time after the program introduction is also important.

Data requirements are greater for a difference-in-difference design than a QED. For a difference-in-difference design, we would not be able to assume that the results from a single time point before or after are sufficient.

DID Challenge 2. Finding a “parallel” comparison group. The second

challenge is that a difference-in-differences model requires a comparison group that meets two conditions:

- If the program had not been introduced, outcomes would have increased or decreased at the same rate in both groups (sometimes known as equal or parallel trends assumption [Gertler et al. 2011]). There is no way to definitively prove this (similar to a matched comparison group QED for which it can

never be shown that there are no initial differences between the groups). But having multiple “before” program data points can increase confidence that this assumption has been met.

- The comparison group should not have been affected by the newly introduced program. For example, if a program were to be introduced in one county or area but not another, the comparison county can help eliminate non-program factors from the estimate that may lead to changes, such as statewide resources or an economic recession.

Regression discontinuity. In a regression discontinuity design (RDD), families are divided into either a program group or a control group based on a cutoff score from an assessment (such as a substance use or parenting stress assessment) or other characteristic that can be used to numerically rank families. An advantage of this design is that families above (or below) the cutoff point—those with the greatest needs—receive the program services. The key to a strong RDD is abiding by the cutoff score to place or not place participants into the program. Although this design has several advantages, such as potentially building on existing program operations, a substantial drawback is sample size requirements. To have the same ability to detect statistically significant effects as an RCT, an RDD requires a sample size about four times as large. For that reason, this design is likely appropriate only for a region-wide or statewide program.

Value of administrative data for impact evaluations

No matter which design stakeholders select, nothing can be done without data. Because collecting data can be difficult and expensive, using only administrative child welfare data can reduce costs for evaluations, so this has strong appeal (Permanency Innovations Initiative Evaluation Team 2016). As with the other designs, it also brings some challenges. Below are tips about using administrative data for evaluations as they apply to either a QED or an RCT.



Administrative Data Tip 1: Administrative-data-only QEDs require a lot of data

Administrative data can allow for comparison groups to be formed without ever engaging comparison group families. The key limitation for a QED is that the administrative data are also the only information available to determine whether families are similar. Because of this, the more data that are obtained, the better—especially information that would suggest that two families would be likely to experience similar outcomes. Because of this, whenever possible, (1) obtain multiple years of data for participants before and after the program, (2) pay careful attention that the time frame covered in the data is similar for program and comparison groups, and (3) obtain data on as many characteristics as possible for matching and establishing baseline equivalence.



Administrative Data Tip 2: Administrative data can facilitate an RCT

As discussed, RCTs are stronger designs than QEDs for detecting program effects. It is worth strongly considering whether and how administrative data can be used in an RCT. The “spotlight” on the next page describes how one grantee carefully developed an RCT within the child welfare system. Administrative data fuels the entire evaluation from random assignment through data collection. The result: a rigorous study with less burden for participants.

The bottom line

Conducting a successful impact evaluation is challenging but has important short- and long-term benefits for funders, providers, and, most importantly, families. Regardless of the evaluation design, it requires the grantee, provider, and evaluation staff to jointly support the evaluation to overcome any challenges. This brief primarily focuses on planning, because careful, proactive work during this phase can help prevent or avoid common problems. But evaluations require care and attention at all stages.

Despite the challenges, the benefits are many. The most important benefit is learning how to better serve families in your community. But in the end, successfully conducting an impact evaluation allows us to say whether a program caused a change for families. Knowing what works for families can help policymakers, practitioners, and communities across the country.

REFERENCES

Administration for Children and Families. "Regional Partnership Grants To Increase the Well-Being of, and To Improve the Permanency Outcomes for, Children Affected by Substance Abuse." Washington, DC: U.S. Department of Health and Human Services, 2012. Available at <http://www.acf.hhs.gov/grants/open/foa/view/HHS-2012-ACF-ACYF-CU-0321/pdf>. Accessed August 8, 2012. (Copies of announcements from the Children's Bureau for closed discretionary grant opportunities are available upon request. Contact info@childwelfare.gov.)

Administration for Children and Families. "Regional Partnership Grants To Increase the Well-Being of, and To Improve the Permanency Outcomes for, Children Affected by Substance Abuse." Funding Opportunity Announcement HHS-2014-ACFACYF-CU-0809, 2014. Available at https://ami.grantsolutions.gov/files/HHS-2014-ACF-ACYF-CU-0809_0.pdf. Accessed September 21, 2017.

Gertler, Paul J., Sebastian Martinez, Patrick Premand, Laura B. Rawlings, Christel M. J. Vermeersch. "Impact Evaluation in Practice." 2011. Washington, DC: The World Bank. Available at http://siteresources.worldbank.org/EXTHDOFFICE/Resources/5485726-1295455628620/Impact_Evaluation_in_Practice.pdf. Accessed September 13, 2017.

Permanency Innovations Initiative Evaluation Team. (2016). *Using Child Welfare Administrative Data in the Permanency Innovations Initiative Evaluation*. OPRE Report 2016-47. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Children's Bureau, and Office of Planning, Research and Evaluation. Available at https://www.acf.hhs.gov/sites/default/files/opre/using_child_welfare_administrative_data_in_pii_compliant.pdf. Accessed September 12, 2017.

RPG evaluation spotlight



The University of Kansas Center for Research Strengthening Families Program: Birth to Three (SFP B-3) Evaluation

The University of Kansas Center for Research is conducting an RCT using administrative data to examine the effectiveness of the Strengthening Families Program: Birth to Three (SFP B-3), a 14-week parenting skills training program. The evaluation includes families with SUDs and children up to 47 months old in foster care.

The team generated a list of eligible families, using two sources: (1) child welfare agency data on families with children placed in out of home care in the last 12 months and (2) referrals from agency front-line staff of additional families who meet the eligibility criteria. Families were then randomly assigned to the program or comparison group. The grantee team is receiving administrative data for both groups on permanency and safety.

Families assigned to the program group were contacted by the site coordinator for recruitment into the program. Program group members consent to receive the alternative services and provide additional data. Anyone who does not consent is counted as attrition, which is being monitored by the local evaluator.

A distinctive feature of this evaluation is that comparison group members need not be directly contacted about the evaluation. They are eligible to receive standard, best practice care and need not provide any additional data for the evaluation. In other words, they receive the same care they normally would, have no evaluation burden, but still are contributing to learning about what works for families.

The university's institutional review board (IRB) reviewed and approved the design. For any evaluation, an IRB is responsible for assessing and minimizing risks to participants.