From Evidence to Impact: Strengthening Evaluation in Child Welfare Services

Design Challenges

Michael Pergamit, Mark Courtney, Bridgette Lery
URBAN INSTITUTE, UCHICAGO, URBAN INSTITUTE

January 2026

Many challenges have limited the development of strategies for evaluating the effectiveness of the child welfare services system,[1] including choosing the best evaluation design for the child welfare context. Some issues create challenges regardless of design choice, while others are specific to randomized controlled trials (RCTs) or quasi-experimental designs (QEDs). Many of these issues could be overcome with investment in evaluation capacity within state and county child welfare agencies.

This brief is the last in a series of three briefs describing the need for rigorous evaluation of policies, programs, and practices affecting the experiences and outcomes of those served by the child welfare system in the United States. In the first brief, Context and Landscape, we identify the current evidence on the effectiveness of child welfare programs and services. In the second Hurdles for Agencies brief, we discuss the challenges child welfare agencies face, limiting the establishment of systematic evidence. In this brief, we discuss challenges to using various evaluation designs. Through these briefs, we hope to begin a conversation in the field about how to overcome the challenges to establishing evidence-based knowledge of the effectiveness of child welfare services based on participants' perspectives. Our aim is to equip families, child welfare agency staff, program developers, legal representatives, and advocates with insights that can guide systemwide improvements to these services.

Common Design Issues

Defining and Identifying the Target Population

Failure to clearly define a program's target population, the specific group of people it is meant to serve, before implementing a program can obscure the program's effectiveness. If people outside of the target population are admitted, it is difficult if not impossible to understand if the program helped its intended beneficiaries. As a result, it is also difficult to describe the population more likely to benefit from the program if implemented elsewhere (i.e., its external validity).

Even when a program's target population is clearly defined, reliably identifying eligible children and families can be difficult for staff connecting them to services, particularly when eligibility is based on the likelihood of experiencing an undesirable outcome. This challenge frequently arises in programs designed to serve families with children at imminent risk of foster care placement or to prevent youth already in care from entering congregate care. Historically, child welfare agencies have struggled to accurately assess a child or parent's likelihood of such outcomes. A consequence of this has been the low prevalence of the unwanted outcome among control group members found in many evaluations of child welfare programs. Yet, to determine that a program is effective, evaluations must focus on populations with a sufficiently high probability of experiencing that outcome in the absence of effective intervention.

To identify these target populations, agencies often rely on a combination of caseworker judgement and structured assessment tools. Structured assessments can help ensure consistency and objectivity, but only if they are used for all members of the population being screened, accurately assess risk levels, and are applied uniformly across staff. Overreliance on professional judgement alone can introduce bias and inconsistency. For example, caseworkers may label a family as "at imminent risk" to secure what they perceive as beneficial services, even when that risk is not truly imminent. In other cases, professionals may deem a family eligible simply because they believe no other appropriate service option is available.

Such patterns can inadvertently undermine evaluation validity. When the families referred to a program meaningfully differ from those the program is designed to serve, evaluators may only discover the mismatch after an evaluation is underway. For example, a recent random assignment evaluation of a reading and math tutoring program for teens in foster care found that nearly 14 percent of all youths referred to the program had both reading and math skills that rendered the program inappropriate for meeting their needs (Zinn and Courtney 2014). These misalignments highlight the importance of clear eligibility criteria, consistent assessment practices, and early evaluator involvement in referral and enrollment planning so that studies accurately measure a program's intended impact.

Estimating Target Population Size and Sample Flow

Conducting a successful program evaluation requires that the evaluator and evaluation partners have a reasonably accurate estimate of the size of the target population available over the course of the evaluation. Detecting program impacts requires a minimum sample size that varies depending on the program and intended outcomes. The evaluation plan will specify the minimum needed sample size. The sample must be large enough to generate treatment and comparison group populations by the end of the evaluation consistent with the evaluation plan's assumptions regarding the statistical power of impact analyses, with larger samples improving the ability to detect an impact that really exists. Accurate estimates of the pace of referrals to a treatment group are also needed to ensure that slots in a program are filled in a timely manner; overestimates can lead to slots being left empty, whereas underestimates can lead to members of the target population needing to wait for undesirably long periods before entering a program.

Public agencies trying to assess the evaluability of a program are often unable to estimate the number of clients they would refer to a program were they able to refer all members of a target population. In many cases, they do not systematically measure the specific needs of their families that are relevant to assessing appropriateness for a given program. This lack of an estimate not only impacts evaluations but also limits the program's ability to seek the amount of funding required to serve all families who could benefit from the program.

Another aspect of the target population size is the rate over time at which children or families become eligible, (i.e., the rate of sample flow expected for the evaluation). This could be determined by looking at past program referrals. However, some public agency administrative data systems do not track referrals to specific programs, so the agency cannot determine how many referrals have been made in the past, or consequently, how many to expect in the future. In some jurisdictions, referrals are captured by data systems, but idiosyncratic referral processes relying on caseworker discretion mean that past referrals are likely a poor proxy for the size of the target population. Ultimately, historical referral data cannot be relied on to estimate future referrals in the absence of data on how many clients exhibit the characteristics that lead workers to refer them to a particular program.

In some cases, estimates of the target population size are known and suggest that an evaluation is feasible, but no data exist on the rate of program take-up. Low take-up not only reduces the likelihood of finding an impact but also can change how the evaluation must unfold. Child welfare services agencies rightfully do not want to have open program slots for more than a short interval. If take-up is lower than expected, changes in the evaluation design may

be required, such as changing the randomization ratio of treatment to control. Such a change will then affect how long the evaluation must continue, to compensate for the reduced statistical power to detect impacts that a different ratio will create. If agencies find the evaluation too difficult to manage, they might not be open to extending the evaluation period, perhaps even terminating it before an adequate sample size can be achieved.

Barriers to Random Control Trials

RCTs are considered the gold standard in evaluation research because they create two equivalent groups, one offered the intervention and one provided an alternative, usually the standard services available in the community. Because the random nature of assignment to intervention and control groups results in groups that do not vary on individual characteristics that might be associated with evaluation outcomes, the groups are equivalent, and we can infer that any difference in their outcomes are attributed to the intervention and not to other factors. This is known as *internal validity*. RCTs are limited when they lack *external validity*, meaning they cannot accurately describe the impact of an intervention outside a specific place with a specific population where there are specific alternative services. However, most alternative designs do not provide strong external validity. Importantly, it is generally believed that the outcomes from an RCT of a well-developed program administered to a common child welfare services population, such as families with substance use issues, generally applies more broadly. The cumulation of multiple rigorous studies reinforces the strengths of those beliefs and is the reason many clearinghouses require more than one study in different locations to achieve the highest rating.

Sometimes Random Control Trials Are Not Feasible

As discussed in our companion brief on agency hurdles (Pergamit, Courtney, and Lery 2025), a significant barrier to rigorous evaluation is resistance in child welfare agencies, but RCTs are not always feasible when it is administratively difficult to faithfully randomize participants and keep the control group from obtaining the intervention.

If genuine saturation occurs, that is, if all eligible individuals or families can be served with existing resources, then one might consider that an RCT is unethical, which can happen with smaller programs or in resource-rich states or counties. However, we note that when agencies do not monitor the size of their target population, they may believe erroneously that they are serving all eligible individuals or families.

Relying on Outcomes from Primary Data Sources Has Trade-Offs

When measures of outcomes are not available in administrative data, outcomes can often be measured using primary data sources, such as surveys or assessments, but these sources carry heavy challenges and trade-offs related to validity, reliability, feasibility, and cost.

Validity refers to how well the data measure the intended purpose. Validity weakens when survey participation declines between baseline and follow-up measurements, especially for the control group members who usually have little reason to participate because they are not part of the intervention program. Measuring outcomes within participants over multiple time points is often desirable to see change over time, but attrition is likely to worsen as participants develop response fatigue or fall out of contact. Strategies to minimize attrition exist, like making periodic contact with control group participants or giving them small regular incentives to stay engaged, but too often evaluators do not allocate resources for these activities.

Surveys and assessments are useful for collecting information that individuals can *reliably* provide. But social desirability can yield unreliable responses on some topics, such as asking respondents about harsh parenting behavior. Unless the evaluation creates its own survey or assessment, existing instruments with good reliability and validity are available, but they can be costly and come with restrictions and requirements about use.

Both *reliability* and *validity* suffer when it becomes infeasible to collect the data. This is true for administrative data, too, but the challenges are different. Primary data require engagement of evaluation participants as well as those delivering the survey or assessment, often a service provider who is already engaged with the participants. However, service providers often do not have contact with the control group, making it challenging to ensure the instrument will be consistently administered across groups.

A final consideration when collecting primary data in an RCT is when to obtain participant consent for participation in evaluation of a program. The intent-to-treat principle for an RCT requires that everyone randomized to treatment and control groups must be included in the measurement of the program's effects, regardless of whether they went on to start or finish the program. If consent is obtained before random assignment, the sample is narrowed to those willing to participate in the evaluation, which leaves a self-selected group with reasons for participating that cannot be measured. In this case, the sample may not be generalizable to broader populations. If consent is obtained after randomization, then the sample may be representative of the target population but the ability to see effects may be weakened because some will not consent to data collection. The ability to detect differences could be further impacted by the fact that those in the treatment group may be more likely to consent to participate in data-collection activities.

Issues with Quasi-experimental Designs

Random assignment with a large enough sample is the most straightforward way to create two comparable groups whose outcomes can be compared. Some QEDs can establish this as well, but a viable QED requires identifying a good comparison group, one that is the same as the treatment group in every way that could influence the outcomes, except that they do not obtain the treatment.

However, finding such a comparison group is often more difficult than agencies realize and, depending on the available data, may not be feasible. Other individuals or families in the same location who do not receive the service are not a valid comparison, as some unmeasured reason may account for differences, such as a family's motivation to pursue the program or uneven knowledge among caseworkers about the services available. Within a state, a rigorous QED relies on the program being offered only in some areas and that the people and context of other areas are comparable with the areas where the program is offered. Furthermore, comparison groups drawn from regions of a state or county that are not offering the program may nevertheless receive services that are very similar to that program, resulting in a poor contrast between the treatment and services-as-usual condition.

Lack of Outcome Data Eliminates Many Quasi-experimental Designs Possibilities

When the outcomes of a program are captured in child welfare services data, such as removals, reunifications, or substantiated reports of abuse and neglect, a comparison group can be considered because the outcomes are known for those not in the treatment group. But many situations exist where outcome data are unavailable for any comparison group. Outcomes of many types of programs, such as substance use treatment or mental health services, will not be captured for people not participating in those programs. Similarly, outcomes collected by programs through surveys or assessments will have no counterpart for people not participating in the program. And adequate outcome data may not even be available for those served by a program. Most child welfare services interventions are delivered by contracted service providers. If any outcome data are collected, it is only for the treatment group, and among them, only those who completed the program or were successful.

Randomized Rollout Designs Can Be Useful but Often Agencies Do Not Meet the Necessary Conditions

When child welfare services agencies introduce new programs or services, they often roll them out across their state over time to allow for training workers and monitoring the process. In such situations, randomized rollout evaluation

designs offer an opportunity for a rigorous evaluation without using an RCT. These designs make use of variation in the timing of implementation start times across sites (e.g., counties, agency offices, organizations) to assess program impacts. Randomized rollout designs identify the impact of an intervention where it is implemented by comparing change over time in outcomes for the people in sites that received the intervention to those in sites that did not receive the intervention. Advantages of randomized rollout designs, when the assumptions needed for their use are met, include their ability to make a strong case for the causal effect of an intervention, relatively easy interpretation of their findings, and their capacity to measure the impact of interventions at a system level.

Among other conditions needed for a strong randomized rollout evaluation are the availability of outcome data on an intervention group and a comparison group at two or more periods, at least once before a program or policy began and at least once after, and that the reasons that sites are selected are unrelated to the outcome.

Unfortunately, we have found that using randomized rollout designs for evaluating programs is often met with challenges that eliminate this design from consideration. This is typically due to the lack of available data on outcomes for the target population prior to the implementation of the program in both treatment sites and comparison sites. Sometimes, selection bias prevents randomized rollouts from being implemented, such as when states decide to roll out an intervention first to areas where they believe the need is greatest, or in their largest counties, best offices, or politically prioritized areas. Thus, the early intervention sites are likely to differ from the comparison sites in ways that will confound the relationship between the intervention and outcomes.

Conclusion

Our experience designing evaluations with child welfare agencies has revealed a wide range of challenges to conducting rigorous studies, and none have been completely challenge free. Some barriers are universal regardless of the evaluation design. Rigorous evaluations require reliable, valid data on the size and needs of target populations, the services delivered, and the outcomes achieved. But designs also differ in key ways that should influence when and how a particular design is used, including their relative level of internal and external validity and the types of data they require.

The most important takeaway is the need for agencies to build evaluation planning into every stage of program development and implementation. Failure to do so can render a strong impact evaluation of a promising program infeasible, because operational procedures are already established that are incompatible with the assumptions of any strong evaluation design. Embedding evaluation design in everyday practice ensures that policies and programs are implemented with both learning and accountability in mind, putting the field on a clearer path toward understanding what works to improve outcomes for children and families.

Notes

[1] The child welfare system includes child protective services, family preservation and reunification services, out-of-home care, and adoption and guardianship services supported by Title IV-E and Title IV-B of the Social Security Act.

References

Pergamit, Michael, Mark E. Courtney, and Bridgette Lery. 2026. "From Evidence to Impact: Strengthening Evaluation in Child Welfare Services, Hurdles for Agencies." Washington, DC: Urban Institute.

Zinn, Andrew, and Mark E. Courtney. 2014. "Context Matters: Experimental Evaluation of Home-Based Tutoring for Youth in Foster Care." *Children and Youth Services Review* 47 (3): 198–204. https://doi.org/10.1016/j.childyouth.2014.08.017.

About the Authors

Michael Pergamit and Bridgette Lery are senior fellows in the Family and Financial Wellbeing Division at the Urban Institute. Mark Courtney is Samuel Deutsch Professor Emeritus at the Crown Family School of Social Work, Policy, and Practice, University of Chicago.

Acknowledgments

This brief was funded by the Annie E Casey Foundation. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission. 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.