

Creating equivalent research groups

Daniel Friend, Angela Valdovinos D'Angelo, Avery Hennigar, Armando Yanez, and Rebecca Piatt

The goal in causal evaluation is to ensure that the differences between two or more study groups are due to the intervention and not to initial (or baseline) variations between groups. To accomplish this, evaluators need to make sure their research groups demonstrate baseline equivalence—meaning the groups are “equivalent” on measured baseline characteristics (for example, age and sex). Evaluators need to monitor and ensure baseline equivalence during the ongoing enrollment into the evaluation and at the end of the study during the analysis stage. Equivalent research groups are essential to accurately estimate the impact of a program. If groups are not equivalent, then they initially differed on characteristics that could be related to outcomes of interest, and the evaluator cannot draw causal conclusions due to the high risk of bias. This brief contains information on: (1) how study design and implementation affect group equivalence, (2) why regular monitoring of equivalence is needed throughout the course of a study, and (3) when demonstrating baseline equivalence for the analytic sample is necessary before conducting analysis. The content of this brief is primarily based on guidance from the [What Works Clearinghouse](#) (WWC).¹



Select and implement an appropriate evaluation design to help develop equivalent groups

Using a randomized controlled trial (RCT) or a quasi-experimental design (QED) enables evaluators to determine whether a program caused a particular outcome—as long as the groups are equivalent at baseline. Below, we briefly describe how each design promotes the creation of equivalent groups:

- RCT.** RCTs use random assignment to create groups that are equal on observed and unobserved characteristics at the time that participants are randomly assigned—which typically occurs at enrollment into the evaluation (Shadish et al. 2002). For example, a Healthy Marriage and Relationship Education (HMRE) evaluation could randomly assign people to receive the HMRE program or not after they complete a baseline survey at program intake. Randomly assigning participants balances out any initial, potentially systematic differences between the groups, since assigning people by chance alone helps to ensure baseline characteristics of both groups will be inherently balanced, on average. If done correctly, randomization enables evaluators to assume that any observed differences between the two groups are due to the program itself.

¹ The WWC is operated by the U.S. Department of Education, Institute of Education Sciences. The WWC guides program evaluators using causal designs to provide rigorous evidence of program efficacy. The WWC contains information on demonstrating baseline equivalence. The Administration for Children and Families adapts clearinghouse standards, like those from the WWC, to provide evaluators with guidance on analysis and reporting of HMRE local evaluations.

The Administration for Children and Families (ACF) provides grants to fund healthy marriage and relationship education (HMRE) programs to strengthen and improve the quality of relationships. The programs offer a range of services from relationship education for high school students to marriage and relationship skills building for adult couples. Grant recipients may be funded to also conduct descriptive or impact evaluations of their funded programs. Independent local evaluators support grant recipients in conducting their local evaluations. This brief is part of a larger evaluation technical assistance (TA) toolkit developed by Mathematica to help HMRE local evaluators understand key program evaluation concepts, common evaluation challenges, and strategies to prevent or overcome challenges. The briefs are standalone documents that can be read in any order. The TA toolkit was developed with HMRE program staff, their local evaluators, and other partners in mind, but it is also relevant to other program areas and organizations.

- **QED.** QEDs do not use random assignment and instead evaluators select a nonrandom comparison group. When conducting a QED, evaluators should strive to select a comparison group that is as similar to the treatment group as possible, apart from the offer of services (Box 7.1). For example, an HMRE evaluator who is studying a program offered in a school district may select a comparison group from another school district within the same county that is not offering HMRE services. Although selecting an appropriate comparison group can help create equivalent research groups, it is not the same as an RCT. As participants are not randomly assigned to the treatment condition, the HMRE program, the treatment and comparison groups in the QED might not be equal on the observed *and* unobserved characteristics at baseline. As a result, evaluators must show that the groups are equivalent in their final analytic sample (WWC 2022).

Box 7.1. Options for selecting an appropriate comparison group for a QED

- Select a comparison group from an area (such as a nearby county) where HMRE services are not offered or the services are substantively different from what is being offered to the treatment group.
- Regardless of the area selected, characteristics of the area and its inhabitants should be similar to those where HMRE services are being offered. Before beginning data collection, use existing data sources (for example, Census data) to help determine the comparison group's characteristics before collecting baseline data from them. Consider using proxy measures if it is not possible to directly assess an outcome from the secondary source. For example, federally sponsored, publicly available data sources, like the Census, likely do not include measures of relationship satisfaction and coparenting. Marriage and divorce rates, child support claims, child living arrangements, and so on could potentially serve as proxy measures for these outcomes.
- Construct a baseline or pre-test survey to identify demographics and outcomes of interest that includes key covariates related to the outcomes of interest. For example, covariates of mental health might influence outcomes related to relationship satisfaction. The same could be true for covariates related to poverty and socioeconomic status. Note that it is necessary to apply any eligibility criteria for the treatment group to the comparison group.

Sources: Handley et al. (2018); Shadish et al. (2002).

Regardless of design choice, evaluators will need appropriate procedures to implement the design as intended to achieve equivalence. In an RCT, group equivalence can be compromised if the assignment procedure was not actually random (Office of Adolescent Health 2014). For example, an HMRE local evaluator or other staff member responsible for randomization might move randomly assigned participants to a different group. Or a staff member conducting random assignment might switch a couple originally assigned to the control group into the treatment group because the staff member perceives that the couple needs services—which would compromise the design by creating an imbalance between the two groups. To properly maintain the study's design, evaluators should clearly communicate key information about the study design to all staff, including how it could be compromised, and regularly monitor that random assignment is being implemented as expected. For example, if half of the randomly assigned group received HMRE services and the other half did not, then a periodic check of random assignment status should show a roughly 50/50 split in terms of sample size.

In a QED, nonequivalent groups can occur when the comparison group is selected from a different population than the treatment group (Office of Adolescent Health 2014). For example, groups would likely be nonequivalent if an HMRE local evaluator using a QED to evaluate a program for single Hispanic mothers selects comparison group members from another area (for example, county or community) that consists of married, non-Hispanic women. To help prevent this, evaluators should use the information in Box 7.1 to help select an appropriate group.



Regularly monitor equivalence to determine whether adjustments are needed

Regular monitoring throughout the evaluation can help evaluators assess when groups may be at risk of nonequivalence. As an evaluator sees groups becoming nonequivalent, they can modify their approach. For example, take a scenario where an evaluator is monitoring attrition and finds that, if the current rate of attrition persists, they will be at risk of having a high-attrition RCT. The evaluator then examines several key baseline characteristics and finds they are different between the experimental and control groups—again noting that, if the pattern continues, the groups will not be equivalent. Based on these conditions, the evaluator can work with staff to improve data collection efforts (for example, for participants with the nonequivalent characteristics) to improve response rates and make the treatment and comparison groups more similar at enrollment.

Sometimes, the evaluation procedures (or lack thereof), rather than attrition, contribute to nonequivalence. Early monitoring of equivalence could help an evaluator identify when staff are not adhering to evaluation procedures. For example, staff could incorrectly be randomly assigning participants. In this instance, early monitoring could help the evaluator revise the procedures or retrain staff to address the issue and correct any imbalances between the groups.

Equivalence between the groups is typically demonstrated by calculating effect size estimates (that is, standardized mean differences; Box 7.2) between the treatment and comparison groups on key baseline characteristics that are expected to influence the outcomes of interest (WWC 2022). To construct a monitoring tool for baseline equivalence, the evaluator will need to compute effect sizes to demonstrate baseline equivalence between the research groups. To do this, they will need each research group's means, standard deviations, and sample sizes for each baseline variable they wish to examine (Table 7.1). Using this information, evaluators should select the appropriate effect size estimate calculation and continue to monitor how these estimates fluctuate throughout data collection.

Box 7.2. Using effect sizes to demonstrate baseline equivalence

The suggested calculation of effect sizes depends on the type of variable.

- **Hedges' g.** An effect size index used with continuous variables such as number of children or ratings of relationship satisfaction.
- **Cox's Index.** An effect size index used with dichotomous variables like whether a participant was married at baseline (yes/no).

The [WWC](#) provides more information on how to calculate these effect sizes. Additionally, several effect size calculators are available online or come standard with statistical software.

Source: WWC 2022.

Table 7.1. Illustrative baseline equivalence monitoring tool

Baseline measure	Intervention group			Comparison group			Effect size of difference
	Mean (or %)	Standard deviation	Sample size	Mean (or %)	Standard deviation	Sample size	
<i>Knowledge of healthy relationships</i>	5.2	4.7	100	8.4	6.9	150	0.53
<i>Relationship satisfaction</i>	4	1	102	4.2	1	150	0.20
<i>Coparenting skills</i>	4	1.75	102	3.75	1.25	150	0.17



Demonstrate baseline equivalence before conducting final analyses

Demonstrating baseline equivalence means that an evaluator shows that the participants in the research groups composing their final analytic sample *remain* similar in terms of their baseline characteristics. Design choices and procedures undoubtedly influence how equivalent research groups are formed, but the need to demonstrate baseline equivalence will differ depending on whether the study is an RCT or a QED. Evaluators using a QED *must always* demonstrate baseline equivalence—as assignment to the treatment or comparison groups is nonrandom (WWC 2022). As a result, an evaluator employing this design needs to demonstrate baseline equivalence to ensure the research groups are similar, to attribute differences in observed outcomes at the end of the study to the program.

An RCT must demonstrate baseline equivalence if it has high rates of attrition. If an RCT loses too many participants over time, the groups may no longer be equivalent in terms of their initial characteristics, particularly when those differences are related to outcomes of interest (Shadish et al. 2002). For example, attrition-related bias would be introduced if participants who did not participate in follow-up data collection were, on average at baseline, less satisfied with their relationship than those who did respond to follow-up surveys. RCTs with high rates of attrition must demonstrate that the study groups are still equivalent at baseline based on who completed each round of data collection and for each outcome the study seeks to examine (WWC 2022). See the fourth brief in this toolkit on understanding and mitigating attrition for more information on calculating and determining when attrition is “high.”

Demonstrating baseline equivalence should be done *before* conducting any analyses. This is because, if the groups are not equivalent, the evaluator may be able to adjust their analysis to account for this. Evaluators can determine whether the two groups are similar by comparing whether the effect size estimates for each characteristic fall within an acceptable range.

Below is guidance for interpreting differences in effect sizes to determine whether groups are equivalent (WWC 2022):

- **Effect size is less than 0.05.** This indicates the groups are equivalent because baseline differences are relatively small and thus not a concern. In this case, the evaluator can proceed with their planned analysis. In Table 7.1, the coparenting variable falls within this range.
- **Effect size is greater than 0.05, but less than or equal to 0.25.** This indicates the groups are not equivalent. As a result, a statistical adjustment is required if the effect size is within this range. For example, the evaluator should use regression covariate adjustments or an analysis of covariance—where the covariate is any baseline variable that falls within this effect size range. In Table 7.1, the relationship satisfaction variable falls within this range.
- **Effect size is greater than 0.25.** This indicates the groups are not equivalent, and the differences are too large for statistical adjustment. In this instance, evaluators should consult their evaluation technical assistance providers and federal project specialists for more guidance on this issue. There may be additional analytic options for the evaluator to try that are too nuanced to detail in this brief. Typically, though, this means that the study cannot demonstrate baseline equivalence and can no longer be considered a causal evaluation, as the threat of bias is too great to make causal claims. In Table 7.1, the knowledge of healthy relationships variable falls within this range.

References

Handley, M.A., C.R. Lyles, C. McCulloch, and A. Cattamanchi. "Selecting and Improving Quasi-Experimental Designs in Effectiveness and Implementation Research." *Annual Review of Public Health*, vol. 39, 2018, pp. 5–25. <https://doi.org/10.1146/annurev-publhealth-040617-014128>.

Office of Adolescent Health. "Baseline Inequivalence and Matching." U.S. Department of Health and Human Services, 2014. <https://opa.hhs.gov/sites/default/files/2020-07/baselineinequivalence-tabrief.pdf>. Accessed June 15, 2023.

Shadish, W.R., T.D. Cook, and D.T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, 2002.

What Works Clearinghouse. *What Works Clearinghouse Procedures and Standards Handbook, Version 5.0*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE), 2022. <https://ies.ed.gov/ncee/wwc/Handbooks>.

Suggested citation: Friend, Daniel, Angela Valdovinos D'Angelo, Avery Hennigar, Armando Yanez, and Rebecca Piatt (2024). "Creating Equivalent Research Groups." OPRE Report #2024-145. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Acknowledgements: Many people contributed to this toolkit. First, we acknowledge staff at the OPRE in the Administration for Children and Families at the U.S. Department of Health and Human Services. We are particularly grateful for the direction and feedback from Samantha Illangasekare, Rebecca Hjelm, and Kathleen McCoy. We want to extend our gratitude to Sarah Avellar and Angela D'Angelo for reviewing drafts of the briefs. We also extend our appreciation to Effie Metropoulos and Bridget Gutierrez for editing and Yvonne Marki-Korosec and Gwyneth Olson for designing the graphics in this toolkit.