CNCS Research Guidance

Baseline Equivalence: What it is and Why it is Needed

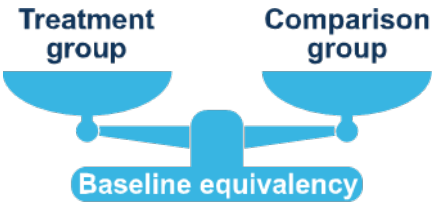
Mary Anne Anderson and Nan Maxwell

Learn how to determine if an impact study is likely to produce meaningful results.

If individuals who participated in a program have better outcomes than those who did not, can program managers say their program improves outcomes?

– Only if the study has baseline equivalence

When a study has baseline equivalence, members of the treatment group (those who participated in the program) are, on average, the same as members of the comparison group (those who did not participate) before the study began. The only observed difference between the two groups is that the treatment group participated in the program. All other observed characteristics—those that can be measured (such as age, race/ethnicity, and education)—are the same.



Researchers want these two groups to be the same so they can say that the program—and not some other factor—caused differences in outcomes between the groups. If the groups were different before the study began, those differences, and not the program, may have impacted outcomes.

Importantly, baseline equivalence must be established for the groups for whom outcomes are compared. Some individuals that were part of the treatment or comparison groups when the study began might not be included in the analysis because, for example, they dropped out of the study. If these individuals are not included in the analysis that compares outcomes between the treatment and comparison groups, they should not be included in analysis that shows baseline equivalence.

This guide is designed to help practitioners and researchers work together to design an impact study with baseline equivalence—or as close to it as possible. When funders and other stakeholders are deciding what programs to fund and scale, it's important that they see evidence of baseline equivalence to have confidence in the program's effectiveness.

How to establish baseline equivalence

Baseline equivalence is important for impact studies because those studies are designed to say whether a program actually *caused* outcomes to occur. Two types of impact studies—randomized controlled trials (RCTs) and quasi-experimental designs (QEDs)—can assess whether a program caused outcomes, ***if*** the study has baseline equivalence between treatment and comparison groups. Baseline equivalence is obtained in different ways, depending on whether the study is an RCT or QED.

- **RCT.** For a RCT, study participants are randomly assigned to either a treatment group that receives the program services or to a comparison group that does not. When random assignment is done correctly, these two groups likely have no differences on either observed or unobserved (that is, unmeasured) characteristics, which allows the study to examine the only difference between them: the program. Still, researchers must be aware of challenges that can arise after random assignment and can affect whether the study groups remain similar to each other (discussed in the next section).
- **QED**. Because study participants are not randomly assigned in a QED, the treatment and comparison groups might not be the same at the start of the study. Researchers must therefore develop methods to select a comparison group that is as similar to the treatment group as possible and then show that the two groups are similar.

What happens when study participants don't stay put after random assignment?

Random assignment generally results in baseline equivalence between the treatment and comparison groups, however, problems can arise that could compromise that equivalence. Two particularly troubling challenges are:

- **Attrition**, or losing people from the study. Attrition can produce study groups that are no longer similar, even if they were similar before the study began. For example, if a study randomly assigned individuals into treatment and comparison groups but a large proportion of individuals in the comparison group could not be located for follow-up surveying, differences in outcomes between the groups might reflect the fact that individuals in the comparison group who responded to the survey are not similar to those in the treatment group who responded to the survey. Because different focus areas in CNCS have different standards for what constitutes an acceptable level of attrition, practitioners should work with their evaluation partners to understand the level that is acceptable in their area.
- **Reassignment**, or switching study participants from the comparison group to the treatment group (or vice versa). Reassignment undermines baseline equivalence because study participants are usually reassigned for a reason that is likely related to outcomes. For example, if highly motivated students who were assigned to the comparison ask school counselors to get into a program, and the counselor moves them into the treatment group, the comparison group is likely left with a higher percentage of unmotivated individuals than the treatment group.

If attrition or reassignment occurs during an RCT, it could jeopardize baseline equivalence and remove confidence that differences in outcomes between the study groups were caused by the program. To mitigate these concerns, researchers should (1) report the extent of attrition that occurred and demonstrate that they were within acceptable levels and (2) analyze study group participants according to their original group assignment (for example, analyzing the "switched" students as being in the comparison group even if they received the program).

SHOWING BASELINE EQUIVALENCE

If attrition occurs at very low levels and reassignment does not happen in an RCT, researchers can maintain baseline equivalence.

A job training program randomly assigns individuals to a treatment group that receives technical training and services during program enrollment and job search and placement services after program completion. Individuals in the comparison group receive only job search and placement services.

Over 98 percent of the treatment and comparison group members completed a follow-up survey. Although five percent of comparison group members were actually enrolled in the program, they were analyzed as being in the comparison group. Both the low levels of attrition and lack of reassignment means researchers can conclude that the groups had baseline equivalence. [The Further Readings provide details on what constitutes "low".]

If, however, over 25 percent of the treatment and 30 percent of the comparison group did not complete a follow-up survey OR program counselors allowed 15 percent of the comparison group to enroll in the program OR the weakest members of the treatment group/strongest members of the comparison group did not complete a survey, researchers might not be able to conclude that the groups had baseline equivalence.

What happens when random assignment is not possible?

Random assignment is not always feasible, and a QED that shows similarity between the treatment and comparison groups on a variety of characteristics might be the strongest design possible. In studies using a QED, researchers must find another way to construct the study groups and demonstrate that the two groups are similar before the study begins. Typically, researchers look for similarities in demographic, socioeconomic, and sociopsychological characteristics and measures of outcomes captured before the study began (such as test scores, employment, body mass index). But even when similarity on these observed characteristics can be shown, without random assignment, the groups might not be similar on unobserved characteristics such as motivation or attitudes. For this reason we have a little less confidence than an RCT that QEDs demonstrate causality.

We offer three methods that might be used to create comparison groups in QED studies.

1) **Using survey or administrative data.** Researchers can use survey or administrative data—data that is used for recordkeeping by governmental and other agencies—to construct a comparison group that is similar to the treatment group.

   For example, researchers might be able to administer a survey to all applicants to a weight loss program, only some of whom will be selected for program participation. In another example, a school might use their administrative records of students who participated in an afterschool reading program and those that did not.

   Researchers can use such data and propensity score matching (see sidebar) to form a treatment group from individuals who enrolled in the program and a comparison group from individuals who did not. Propensity score matching helps identify program participants and nonparticipants who are most similar to form the treatment and comparison groups. Of note, data on characteristics (such as demographic or socioeconomic characteristics, knowledge or beliefs, or opinions) can be used to show baseline equivalence between the groups.

   Limitation: Although techniques like propensity score matching are often viewed as rigorous alternatives to random assignment, this method requires sophisticated statistical knowledge and appropriate data.

2) **Using a cutoff score.** Sometimes test scores are used for admitting individuals into a program. For example, applicants might have to score 80 percent on a test to be admitted to a program. Individuals who score close to the cutoff score are likely to be similar in every way except program admission: the likely difference is that the group scoring slightly above the cutoff score guessed correctly on a couple more questions than the group that scored slightly below it. Given this similarity, the group that scored just above the cutoff score can form the treatment group and the group that scored just below the cutoff can be the comparison.

   Limitation: Researchers need a very large group of applicants to have enough individuals in the treatment and comparison groups.

3) **Choosing people in similar contexts.** Researchers can use different environments to select comparison group members. For example, researchers might compare academic achievement for students enrolled in an afterschool reading program with students in similar districts that do not offer the program or with students enrolled in the district (in the same grade) during the year before the program began. Alternatively, researchers might use preexisting data to develop a comparison group. For example, outcomes of individuals in a nutrition program for low-income mothers might be compared to individuals in the Current Population Survey or

PROPENSITY SCORE MATCHING

Propensity score matching is frequently used to create a comparison group when random assignment cannot be used.

This technique is intended to mimic random assignment by creating study groups that are similar based on their characteristics that can be captured in the data source.

A propensity score reflects the probability that a person with a given set of characteristics (that are captured in the dataset) will enroll in the program. The score developed from this probability can be used to select a matched group of individuals who are enrolled in the program (treatment group) and individuals who are not (comparison group) and to balance the observed characteristics of participants between the two groups.

administrative data from a program like Women, Infants, and Children. Researchers must use some type of matching technique to establish that the groups had similar characteristics and influences on behaviors (for example, school conditions or other environmental factors).

Limitation: It is often difficult for researchers to establish that the only difference between the treatment and comparison group is the program, given the plethora of environmental and contextual factors that likely exist between the groups.

Having individuals who choose to participate in a program form the treatment group and those who chose not to participate form the comparison group is generally considered to be an extremely weak design. The characteristics that lead individuals to choose to participate makes them different. Often, those who chose to participate in a program are more motivated, have fewer barriers to participation, or exhibit more grit and persistence than those who chose not to participate. Researchers using such a design must demonstrate baseline equivalence between the groups and recognize that some of the unobservable characteristic differences between the groups might be creating estimated impacts.

Which design do I choose?

Which design is best? Random assignment is the gold standard for achieving baseline equivalence—if low levels of attrition and no reassignment can be attained. However, real world considerations often require researchers to conduct a QED and these designs require considerable forethought to establish the baseline equivalence. The best QED study uses techniques, such as propensity score matching or a cutoff score, to develop a comparison group that is similar to the treatment group. Still, both RCT and QED studies must show that members of the treatment and comparison group **for whom outcomes are compared** are similar with respect to their characteristics before the study began.

Further Reading

Home Visiting Evidence of Effectiveness Review, On Equal Footing: The Importance of Baseline Equivalence in Measuring Program Effectiveness (https://homvee.acf.hhs.gov/HomVEE_brief_2014-50.pdf)

What Works Clearinghouse Review

WWC Standards Brief: Baseline Equivalence (https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_brief_baseline_080715.pdf)

Baseline Equivalence: Module 3 (https://ies.ed.gov/ncee/wwc/Docs/OnlineTraining/wwc_training_m3.pdf)

KEY POINTS ABOUT BASELINE EQUIVALENCE

- Baseline equivalence must exist to accurately estimate program impacts.
- When studies use random assignment to form treatment and comparison groups and there are low levels of attrition and no reassignment, researchers can be confident of baseline equivalence.
- Ensuring baseline equivalence before a study begins is not enough. The researchers must show that the individuals in the treatment and comparison groups at the end of the study were similar before the study began.
- Showing equivalence only for characteristics that we can measure—such as age, education, race/ethnicity, and gender, or a score on a pretest—does not ensure baseline equivalence because characteristics that we cannot measure—such as values, motivations, and attitudes—can affect a person's outcomes. Random assignment helps establish this equivalence.

About the Series

The Corporation for National and Community Service (CNCS) supports the scaling of effective interventions that it funds and has engaged Mathematica Policy Research to conduct the Scaling Evidence-Based Models project (contract GS10F0050L/CNSHQ16F0049). As part of that project, Mathematica developed a series of guides to help practitioners collect evidence on their interventions' effectiveness and increase the likelihood of successfully scaling those interventions.

Each guide provides a succinct overview of a topic that can help practitioners. The guides are based on research and practitioners' experiences, but they do not provide exhaustive reviews of a topic. More in-depth articles can be found in the Further Reading section.