Home Visiting Programs

On Equal Footing: The Importance of Baseline Equivalence in Measuring Program Effectiveness

August 2014

OPRE Report #2014-50

To understand the effects of a program, researchers must distinguish effects caused by the program from effects caused by other factors. This effort typically involves comparing outcomes for two groups. The similarity of the two groups before program services begin is referred to as **baseline equivalence**.

In the Home Visiting Evidence of Effectiveness (HomVEE) review, researchers search, screen, review, and rate studies to identify home visiting programs with rigorous evidence of effectiveness. The results may help inform the decisions of state administrators, practitioners, and other stakeholders. This brief explains the importance of baseline equivalence for measuring a program's effectiveness. Baseline equivalence is an essential consideration for HomVEE reviews of certain research designs and may determine whether the study earns a rating of high, moderate, or low.

This brief on research methods and standards was written by Sarah A. Avellar and Jaime Thomas of Mathematica Policy Research. Since 2009, Mathematica has conducted the Home Visiting Evidence of Effectiveness (HomVEE) review under contract to the U.S. Department of Health and Human Services. The purpose of the review is to identify, assess, and rate the rigor of impact studies of home visiting programs for pregnant women and families with children from birth to age 5.

The HomVEE website:
http://homvee.acf.hhs.gov/

Why Is Baseline Equivalence Important for Program Evaluation?

Studies designed to measure a program's effects should include two groups: a **program or treatment group** and a **comparison or control group**.[1] The program group can receive services through the program being evaluated (for example, home visiting). Those in the comparison group (sometimes known as the "business as usual" group) should not receive those services but may receive other services in the community. Without a comparison group, researchers cannot differentiate between changes caused by the program and changes that arise for some other reason. Consider, for example, a study measuring cognitive development for a group of children before and after their participation in a program. To determine how much of the children's development was due to the program (or whether they would have developed as much without it), researchers need to compare them to a group of children who did not participate.

The comparison group represents the **counterfactual**—what would likely have happened to participants if they had not participated. The ideal (and impossible) experiment would include people simultaneously participating and not participating in the program. Such a situation would rule out all other explanations for change, other than the effects of the program. Because this ideal is impossible, researchers aim to form comparison groups that are as similar to the program group as possible before program services begin (that is, at baseline). For example, if the program group in the hypothetical study examining child cognitive development includes only three-year-olds, a comparison group of only five-year-olds is a poor counterfactual. Comparing three-year-olds who participated in the program to five-year-olds who did not participate is a not a good contrast because the groups are developmentally quite different, regardless of program participation. The more similar the groups at baseline—the greater the equivalence—the more likely it is that systematic differences in outcomes can be attributed to the program. (Researchers use tests of statistical significance to distinguish between measured differences that are systematic or significant and those that may have arisen by chance and are not significant.) A study that establishes baseline equivalence on key characteristics is better able to isolate the effects of the program from other factors.

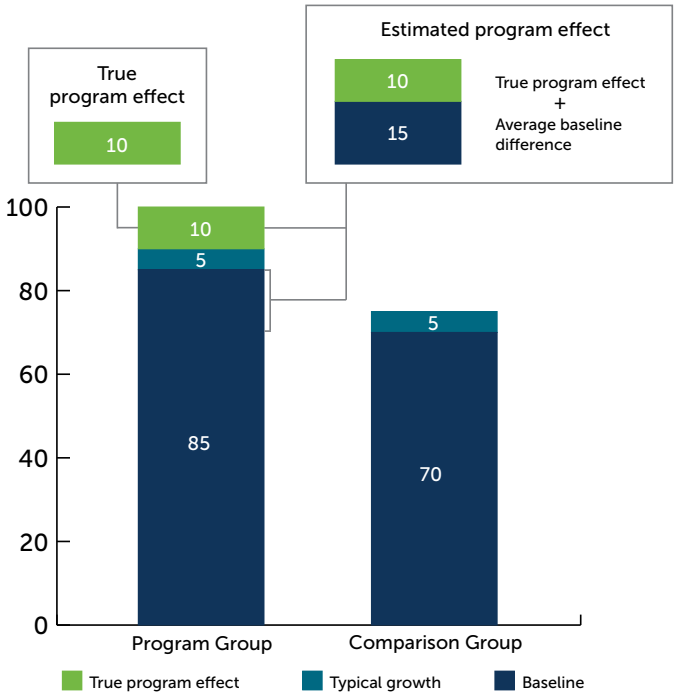
How Can Researchers Be Sure That Study Groups Are Equivalent at Baseline?

The design of the study can create baseline equivalence. In a study with an experimental design, called a randomized controlled trial (RCT), researchers assign people randomly to the program or control group. Random assignment is like drawing names out of a hat to create the program and comparison groups. The key advantage to using random assignment is that, when properly implemented, it creates two groups with no systematic differences at baseline.[2] Members of the two groups would be similar in age, race/ethnicity, education, and other characteristics that are easier to measure. In theory, the groups would also be similar in characteristics that are more difficult to measure, such as parenting style, motivation, or attitudes.

By contrast, a study with a quasi-experimental design does not assign people randomly to the program and comparison groups. Instead, people choose a group for themselves (in other words, self-select) or are intentionally assigned to either group. With this method, researchers can no longer assume that program and comparison groups are equivalent on measured and unmeasured characteristics.

Researchers can minimize the chance of *measured* differences between groups. In a matched comparison group design (MCGD), researchers select people who are as similar to the program group as possible to form the comparison group. For example, if the program group is made up of mothers ages 18–24, the comparison group should include mothers in the same age range.

Figure 1: Baseline differences lead to biased estimates



Key Terms

Baseline: Period of time before the start of program services that are being evaluated

Bias: Erroneously shifting the results in one direction or another

Comparison group: Study participants who should not receive services from the program of interest, but may receive other available services

Counterfactual: What would have happened to people in the program group if they had not participated in the program

Equivalence: Similarity (on average) between two groups

Program group: Study participants who can receive program services; sometimes known as "treatment group"

But researchers conducting an MCGD study cannot rule out the possibility that the two groups have *unmeasured* traits that are not equivalent. For example, although the study participants are the same age, researchers do not know if the program group members are more motivated or determined to achieve the goals of the program. If the program and comparison groups differ in unmeasured characteristics—if they are not equivalent at baseline—then the study's estimates of program effectiveness may not be accurate. The lack of baseline equivalence creates bias: erroneously shifting the results in one direction or another.

Figure 1 shows an example of how bias can occur when the program and comparison groups differ in an important way. In this example, the baseline productive vocabularies of toddlers in the program group (85 points) are larger, on average, than those of toddlers in the comparison group (70 points). Six months later, after the program has ended, both groups have improved simply because the children got older (5 points). The program group has also improved from participating in the program (10 additional points). If researchers measure outcomes at six months, there is a 25-point difference between toddlers in the program group (85 + 5 + 10 = 100 points) and toddlers in the comparison group (70 + 5 = 75 points). But part of this difference is due to the initial (baseline) underlying differences. The true effect of the program is 10 points. Initial differences in the two groups led to a biased estimate of program effectiveness.

In this example, researchers could have measured the baseline productive vocabulary of toddlers in the program group and created more-similar groups. But in practice, there will always be some unmeasured characteristics that may create differences between groups.

What Is the HomVEE Standard for Baseline Equivalence?

The highest HomVEE rating an MCGD can receive is "moderate" because MCGDs by their nature involve some uncertainty about the similarity of the program and comparison groups at baseline. To receive a moderate rating, MCGDs must establish baseline equivalence in (1) race and ethnicity, (2) socioeconomic status (SES), and (3) baseline measures of outcomes, when feasible.[3] Baseline equivalence is established if there are no statistically significant differences between the groups on the three sets of variables.[4] For a study to receive a moderate rating, the researchers must also use statistical methods to increase the likelihood that small differences between groups do not affect estimates of program impacts. Such methods include using baseline measures of outcomes as controls (or covariates) in a regression model (for more details, see http://homvee.acf.hhs.gov/document.aspx?rid=4&sid=19&mid=5).

A lack of equivalence on any of the above variables at baseline may bias the results. HomVEE requires baseline equivalence in race and ethnicity and SES because research shows links between these variables and outcomes such as child health and child cognitive and social-emotional development.[5] That is, race/ethnicity and SES matter because they are closely linked with the outcomes of interest often included in research reviewed by HomVEE. HomVEE requires baseline equivalence in baseline measures of outcomes (when feasible) because they are often the strongest predictors of subsequent outcomes.[6] Collecting these measures is not feasible for all studies, however. For example, in a study of children's vocabulary outcomes, the baseline may occur prenatally, when it is not possible to measure vocabulary. Studies like this must still establish baseline equivalence in race/ethnicity and SES.

The HomVEE review also requires two other elements to establish baseline equivalence of the program and comparison groups:[7]

1. **Equivalence must be established on characteristics measured before the program group has received services.** That is, characteristics used to assess equivalence must truly be measured at baseline. Characteristics may change over time, sometimes as a result of the program itself. For example, a program may offer services to support a family's economic well-being, which could affect measures of SES.
2. **Equivalence must be established on the groups used in the analysis.** It is not sufficient to demonstrate equivalence on the initially formed groups. If people drop out of the study or are not assessed at follow-up, groups that began as equivalent might have quite different compositions by the follow-up. For this reason, MCGDs must demonstrate baseline equivalence for the sample included in the follow-up impact analysis. Information about the baseline equivalence of the initial sample, including those who eventually dropped out, does not satisfy this requirement.

How Can the Baseline Equivalence Standard Be Met?

Depending on circumstances, an MCGD study can use one of several methods to form, or match, program and comparison groups that are likely to be similar at baseline.

First, before collecting study-specific data, researchers can use general information to intentionally select groups from populations that are likely to be similar. For example, researchers may choose to study a program in a rural community with a large Hispanic population because there are other similar communities from which to draw the comparison group.

Second, researchers may use administrative data (such as birth records) that include some families that participated in the program and some families that did not. If such data are available, researchers can select within those groups to form the program and comparison groups. The researchers should select similar subsets from those who participated and those who did not to achieve baseline equivalence.

A third method is available if researchers wish to use a data source that includes program participants but not nonparticipants. In such a case, they can construct a comparison group using a different data source that includes individuals similar to those who participated in the program. For example, if a home visiting program has data for program participants, the WIC or Medicaid database could provide a pool of nonparticipants from which researchers could select a similar comparison group.

With all methods, researchers should carefully consider variables for which it is important to establish equivalence when constructing the groups. The variables needed for baseline equivalence to meet the HomVEE standards are not the only characteristics on which it may be important to establish baseline equivalence. An approach to identifying additional key variables is considering whether a typical reader or practitioner would consider two groups to be similar using common sense, experience, and practice-based knowledge.

Once the variables are selected, the researchers should then assess whether baseline equivalence was achieved. The most common method of demonstrating baseline equivalence is to compare average baseline characteristics for the program and comparison groups. Typically, researchers include a table that compares the mean for each baseline characteristic of each group, along with the results of statistical tests indicating if any of these means are significantly different for the two groups.[8]

Of course, even if the statistical test results suggest that the program and comparison groups are alike at baseline, the risk of differences in other (unmeasured) characteristics remains. Nevertheless, equivalence between program and comparison groups at baseline on key measured characteristics improves researchers' confidence that the comparison group is a good representation of the counterfactual. When baseline equivalence is demonstrated, the potential for biased estimates is reduced.

What Not to Do

Two approaches for selecting comparison groups are described below. These approaches are not advised even if the resulting program and comparison groups match on observed characteristics.

1. Groups may be formed by drawing from those who were eligible for the program: those who chose to participate make up the program group, and those who chose not to participate make up the comparison group. Although the groups may match on observable characteristics, they are likely to differ in other ways, which led to different choices about participation. For example, those who chose not to participate may be less motivated, may have more barriers or challenges in their lives, or may have different attitudes than those who did participate. These differences could affect outcomes regardless of program participation.
2. Groups may be formed using a historical comparison group. For example, if a program begins in a neighborhood in 2012, the program group may be drawn from residents in that year, but the comparison group may be made up of residents from that neighborhood in 2011. The difficulty with using a historical comparison group is that there may be other time-related differences between the groups that affect outcomes, such as other services in the community or public policies. These other differences will be reflected in any estimates of program effects.

Endnotes

[1] Control group generally refers to a group formed through random assignment. The broader term "comparison group" may refer to groups formed randomly or nonrandomly.

[2] Random assignment ensures there are no *systematic* (statistically significant) differences between the program and control groups, on average, but there can be *chance* differences.

[3] This standard also applies to RCTs with high attrition, which occurs when people drop out of the study or are not assessed at follow-up. Although random assignment in RCTs produces initially equivalent groups, on average, attrition can undermine that advantage.

[4] HomVEE uses alpha = 0.05 to determine statistical significance, meaning that p-values greater than 0.05 indicate that the difference between groups is not statistically significant.

[5] See Bradley, R.H., and R.F. Corwyn. "Socioeconomic Status and Child Development." *Annual Review of Psychology*, vol. 53, no. 1, 2002, pp. 371–399; Case, A., D. Lubotsky, and C. Paxson. "Economic Status and Health in Childhood: The Origins of the Gradient." *American Economic Review*, vol. 92, no. 5, 2002, pp. 1308–1334; and MacDorman, M.F. "Race and Ethnic Disparities in Fetal Mortality, Preterm Birth, and Infant Mortality in the United States: An Overview." *Seminars in Perinatology (Science Direct)*, vol. 34, no. 4, August 2011, pp. 200–208.

[6] Hallberg, K., P.M. Steiner, and T.D. Cook. "The Role of Pretest and Proxy Pretest Measures of the Outcome in Removing Selection Bias in Observational Studies." Presented at the Society for Research on Educational Effectiveness conference, March 5, 2011.

[7] The HomVEE standards fit the needs of the HomVEE review; however, there are other methods for assessing baseline equivalence (see, for example, the What Works Clearinghouse standards at http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf, pp. 15–16). Other factors also can influence the quality of quasi-experimental designs, such as the measures used or whether some aspect of the design other than the program lines up with the program or comparison group (sometimes known as a confounding factor). More information on HomVEE standards is available online (http://homvee.acf.hhs.gov/document.aspx?rid=4&sid=19&mid=5). Researchers also may consider other ways to improve the quality of quasi-experimental designs beyond HomVEE standards (see, for example, Shadish, W.R., T.D. Cook, and D.T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. New York: Houghton Mifflin Company, 2002).

[8] For categorical variables such as race/ethnicity, chi-square tests are often used. For continuous variables such as test scores and household income, t-tests may be used.