AmeriCorps

Mathematica®
Progress Together

Employment Brief

Mary Anne Anderson and Nan Maxwell

What Makes for a Well-Designed, Well-Implemented Impact Study

Learn what is needed for a well-designed and well-implemented impact study.

An impact study assesses whether a program improved outcomes for its participants. The study builds an understanding about whether a program unambiguously improved the outcomes that it intended to improve. It can also give insights into whether the program improved other outcomes. Although other types of studies can reveal what factors might be associated with better outcomes, only impact studies can tell whether a program actually *caused* them. Some examples of changes that might be examined in an impact study include increased student test scores, people becoming employed, or people improving their eating habits.

A well-designed and well-implemented impact study contains the following components:

- **Internal validity.** If a study has internal validity it can clearly separate the effects of an intervention from other factors that may have impacted the outcomes. To do this, researchers must carefully construct a counterfactual condition (what program participants would have done without the program). A comparison group is typically used to capture the counterfactual. Internal validity depends both on how well the study's design constructs the counterfactual and how successfully that study design is carried out, meaning how well it is implemented.
- **Independence of findings.** A well-implemented study is not influenced by the individuals who designed or implemented the program or the funders who paid for it. The study should be designed and conducted by impartial researchers to ensure independence.

This guide is intended to help practitioners ensure that their evaluators produce high- quality impact studies. Such studies produce valuable information for practitioners, funders, and other stakeholders, allowing them to understand whether a program is effective and whether it is worth increased funding or scaling to reach more people.

Building internal validity

Can program managers say that their program caused the improvements they see in their clients?

– Yes, if the study has internal validity

Internal validity means that the differences shown between a group of people that participate in a program (treatment group) and a group that does not (comparison group) can be credited to the program and not to other factors. Only a well-designed and well-implemented randomized controlled trial (RCT) can allow researchers to make such statements, although a well-designed and well-implemented quasi-experimental design (QED) can come close:

- A **RCT**—also called an experiment—uses randomization to determine who can enroll in the program (and be in the treatment group) and who goes into the comparison group. It is this randomization that assures that members of the treatment and comparison groups have the same characteristics, both those that can be measured and those that cannot (like motivation or ability).

What is a RCT?

In an RCT, people are randomly assigned to either the treatment group or to the comparison group. Randomization helps to ensure that the people in both groups have the same characteristics—both those that can be seen or measured and those that cannot. The groups are likely to be the same because each person had an equal probability of participating in the program and being in the comparison group. Any differences in outcomes between the groups can be attributed to the program. The RCT is considered the "gold standard" for social and clinical research.

- A **QED** uses a method other than random assignment to form study groups. Even the strongest QED studies, which select people for each study group in a way that makes them as similar to each other as possible before the study begins, cannot control characteristics that cannot be observed. For example, individuals might be selected for the comparison group if they have demographic characteristics similar to individuals in the treatment group, but such matching does not account for the fact that the treatment group consists of individuals who knew they could benefit from the program.

What is a QED?

A QED does not use randomization to assign individuals into study groups. Instead, researchers assign people to the treatment or comparison group using another method and demonstrate that both groups have the same characteristics. The characteristics should be quantifiable, such as demographics and socioeconomic measures, and include measures at for individuals the beginning of the study that will be used as outcomes—such as test scores, employment, and body mass index.

For a QED study, researchers often collect information on people in each group before the program begins and after it ends, (this is called a pre- post design). This design allows researchers to estimate the change in outcomes that can be attributed to the program.

Four challenges jeopardize whether an impact study has internal validity. Each can create differences between the treatment and comparison groups. Such differences between the two groups, and not the program itself, can impact outcomes.

1. **Attrition** refers to losing participants from the study. Although virtually all studies have some attrition, some studies lose enough participants that the treatment and comparison groups are no longer the same. Attrition is particularly important in an RCT because random assignment created study groups that were the same when the study began. If more people leave one study group than the other study group, the people left in those groups at the end of the study might not

have been similar to each other when the study began. For example, less motivated individuals might drop out of the treatment group in a job training program.

2. **Reassignment** refers to actively switching study participants from the comparison group to the treatment group (or vice versa). Reassignment is a major concern for RCTs and undermines validity because study participants are usually reassigned for a reason that is likely related to outcomes. Such switching might occur if, for example, children who applied to a reading program were randomly assigned to the comparison group, their parents demanded that their children get into the treatment group so they can receive the program, and counselors switch them. Even though these children are now receiving the program, the study would need to consider these children as being in the comparison group (that is, ignore the reassignment) when examining the impact of the reading program.
3. **Lack of baseline equivalence** refers to differences between people in the treatment and comparison groups before the study begins. This is a major concern for QEDs. These dissimilarities—and not the program—might create differences in outcomes. Differences can arise if groups are formed in ways other than random assignment. For example, researchers might admit the first 50 individuals into a weight loss program (the treatment group) and the next 50 individuals into a comparison group that does not receive the program. The study's validity might be questioned because the groups might not be similar because more highly motivated individuals often are the first to enroll in a program. For this reason, researchers have less confidence that QEDs demonstrate causality than RCTs.
4. **Confounding factors** refers to the presence of a factor other than the program that could affect outcomes. The presence of a confounding factor makes it impossible to tell whether the program, the confounding factor, or both caused differences in the outcomes between the treatment and comparison groups. For example, researchers examine whether a math enhancement program improved test scores. They randomly assigned students into a treatment group that received the enhanced program and a comparison group that received the regular program. In addition, the researchers assigned one teacher to provide instruction in the enhanced program and another to provide instruction in the regular program. In this case, the teachers are the confounding factor. Because a different teacher instructs the treatment and the comparison group students, we would be unable to tell if differences in test scores were caused by the math enhancement program or the teachers.

Example of a confounding factor

If experienced certified nutritionists run a new program for weight loss while interns use an existing program, greater weight loss among the treatment group might be due to the confounding factor of experience. Nutritionists might be better than interns at working with and educating program participants. The study would be stronger if it controlled for these confounding factors by, say, having both interns and nutritionists teach both treatment and comparison groups.

Ensuring independence

To ensure that the findings from a study are relatively free from bias and subjective judgements, impact studies should be designed and conducted by objective researchers. Although programs might have internal evaluation staff who collect data and conduct studies about the program, such individuals are generally perceived to be biased toward the program, no matter how much they strive to be objective. To ensure an objective assessment, practitioners generally contract with "third-party" evaluators, who are frequently associated with a research and evaluation firm or university. Having a third- party evaluator not only reduces the probability that the researchers' beliefs sway the results of a study, it also helps ensure that the evaluation is conducted by an expert who is well acquainted with the requirements of a well-designed and well-implemented impact study. Such experts generally understand the need for internal validity and bring an outsider's perspective to program conditions

that go unnoticed or unmeasured by internal evaluation staff. Still, hiring and working with third-party evaluators costs money and requires collaboration to ensure the evaluators understand the program. This can be a worthwhile investment— if an impact study is objectively designed and implemented and shows the program to be effective, the program can attract future funders and partners.

What about external validity?

External validity allows researchers to generalize a study's findings to a variety of situations and people and not just to the people in or the location of the study. External validity requires that researchers use high-quality sampling methods and consider who is included in the study—studies generally include only a subset of the overall population and settings in which the intervention is implemented. For a study's findings to apply to similar settings and populations, researchers must ensure that the setting and population studied are typical. The best way to show typicality is for researchers to use random selection. Because it is difficult for a single study to have its findings widely applicable, researchers often replicate studies in different settings and for different populations to demonstrate the intervention's effectiveness in a wide variety of situations. The sidebar provides an example of the limitations researchers face in extrapolating their findings to different situations.

Example of external validity

Researchers want to know how well a reading program works for seventh- and eighth-grade students in Midwestern cities.

For their RCT, researchers randomly select 3 of the 20 districts in which the program is implemented and then take a random sample of seventh- and eighth-grade students from the three school districts. The random selection of both districts and students allows the researchers to say that the study's results probably apply to seventh- and eighth-grade students in the 20 Midwest cities in which the program is being implemented.

However, the study's findings may not apply to students in other grades in those 20 districts, or to seventh- and eighth-grade students in districts outside of the 20 from which the sample was drawn. Further research would be needed to know if the program would be successful in these circumstances.

Further Reading

Clearinghouse for Labor Evaluation and Research, Causal Evidence Guidelines Version 2.1 (https://clear.dol.gov/sites/default/files/CLEAR_EvidenceGuidelines_V2.1.pdf)

Home Visiting Evidence of Effectiveness Review

What Isn't There Matters: Attrition and Randomized Controlled Trials (https://homvee.acf.hhs.gov/sites/default/files/2019-08/HomVEE_brief_2014-49.pdf)

Addressing Attrition Bias in Randomized Controlled Trials: Considerations for Systematic Evidence Reviews (https://homvee.acf.hhs.gov/sites/default/files/2019-06/HomVEE-Attrition-White_Paper-7-2015.pdf)

Home Visiting Evidence of Effectiveness Standards for Random Assignment Studies (https://homvee.acf.hhs.gov/sites/default/files/2019-08/HomVee_Standards_Flowchart_w_Definitions_Random_B508.pdf)

Home Visiting Evidence of Effectiveness Standards for Matched Comparison Group Designs (https://homvee.acf.hhs.gov/sites/default/files/2019-08/HomVee_Standards_Flowchart_w_Definitions_Comparison_B508.pdf)

What Works Clearinghouse Review WWC Standards Brief: Attrition

(https://ies.ed.gov/ncee/wwc/Docs/referencesources/wwc_brief_attrition_080715.pdf)

WWC Standards Brief: Confounding Factors

(https://ies.ed.gov/ncee/wwc/Docs/referencesources/wwc_brief_confounds_101117.pdf)

Reporting Guide for Study Authors: Group Design Studies (https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_gd_guide_022218.pdf)

Reporting Guide for Study Authors: Regression Discontinuity Design Studies (https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_rdd_guide_022218.pdf)

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.

Mathematica, Progress Together, and the "circle M" logo are registered trademarks of Mathematica Inc.