GIRL Center Research Brief
No. 4 April 2019

DELIVERING RESULTS IN GIRLS' EDUCATION: HOW TO EVALUATE WHAT WORKS, WHAT DOESN'T, AND WHAT WE DON'T KNOW

STEPHANIE PSAKI

GIRL Center
FOR INNOVATION, RESEARCH, AND LEARNING

The Girl Innovation, Research, and Learning (GIRL) Center generates, synthesizes, and translates evidence to transform the lives of adolescent girls

popcouncil.org/girlcenter

GIRL Center Research Briefs present new knowledge on issues of current and critical importance and recommend future directions for research, policies, and programs.

Suggested citation: Psaki, Stephanie. 2019. "Delivering Results in Girls' Education: How to Evaluate What Works, What Doesn't, and What We Don't Know," GIRL Center Research Brief No. 4. New York: Population Council.

Feedback? Please let us know how we can make this brief more useful in your work, or share ideas for other resources that would be helpful with GIRLCenter@popcouncil.org.

Support for this brief comes from Echidna Giving.

GIRLS AROUND THE WORLD FACE DAUNTING CHALLENGES WHEN IT COMES TO ENROLLING IN PRIMARY SCHOOL, COMPLETING SECONDARY SCHOOL, AND GAINING THE BASIC KNOWLEDGE AND SKILLS THEY NEED TO BE EMPOWERED, HEALTHY, AND PRODUCTIVE ADULTS.

Many governments and organizations have risen to meet that challenge through policies and programs designed to remove common barriers to girls' education. But as the number of actors in this space grows, and resources to address these challenges remain scarce, it is essential to ensure that investments are targeted toward the most effective policies and programs.

This is not always easy. Organizations may be focused on delivering programs that work but may not have the resources, expertise, or mandate to evaluate what is or is not working. If we want to make the best use of available resources and design the most effective programs, organizations must integrate available evidence into program design and implementation. Those who implement programs must also share their experiences with researchers to ensure that evaluations are relevant, understandable, and useful for future work.

This brief lays out the basics of program evaluation. The goal is to provide practitioners, policymakers, donors, and advocates working in girls' education with basic tools to critically assess and integrate evidence into decisions about program and policy design and advocacy messages.

HISTORY

In 2009, the Population Council published New Lessons: The Power of Educating Adolescent Girls,[1] which reviewed the most common approaches to promoting girls' education, and the available evidence for those strategies. A review of more than 300 programs being implemented around the world revealed the lack of quality evidence in support of much of this girls' education effort. For example, of the 11 most common intervention approaches identified in the review, only two had been proven effective in previous research. Despite this widespread lack of evidence, only 28 percent of the programs reviewed had an evaluation planned, and only 9 percent had an external evaluation planned.

This disconnect between research and practice is not unique to girls' education. That is why the Population Council has been building one of the world's largest bodies of research on programs to improve the lives of adolescents, especially girls. Drawing on evaluations of programs with more than 50,000 adolescent girls and boys living in Latin America, sub-Saharan Africa, and South Asia, we are working with policy makers and practitioners to ensure that investments in adolescents are evidence-based.

A decade after *New Lessons* was published, the Population Council is updating and expanding that work,[2] by: 1) systematically mapping the ecosystem of policymakers, practitioners, researchers, and advocates working in global girls' education; 2) synthesizing the evidence on what works; and 3) identifying opportunities to scale-up successful interventions and investments.

[1] https://www.popcouncil.org/uploads/pdfs/2009PGY_NewLessons.pdf
[2] https://www.popcouncil.org/research/geemap

CORE QUESTIONS WHEN TRYING TO UNDERSTAND WHETHER A PROGRAM ACHIEVED ITS GOALS

1 DID EDUCATION IMPROVE FOR THOSE WHO PARTICIPATED IN THE PROGRAM?

One of the most basic questions an evaluation must answer is whether education improved over the course of the program for participants.

How can an evaluation answer this question?

Collecting data before and after the program was implemented allows us to see what improvements took place, and for whom (see below). This information can be collected by:

- Administering assessments (e.g., testing literacy or numeracy skills)
- Examining school records (e.g., of attendance, test results)
- Asking students, parents, or teachers questions (e.g., *What are the main reasons you did not attend school last week?*)
- Conducting observations (e.g. classroom, household)

What if an evaluation does not answer this question?

It is essential to collect data after a project ends (or after enough people have participated). But sometimes data collection before a project takes place is not possible, due to resource or time constraints. Evaluations that ask questions only after the project ends ("*endline only*") may still provide useful information about the program. In these cases, finding a good comparison group at endline is essential (see question 2). *"Endline only" evaluations may either overestimate or underestimate the effects of a program.*

We refer to improvements in "education" throughout this brief, but this could refer to any program goals, also known as outcomes, including school enrollment, attainment, literacy, numeracy, school violence.

"Self-reported" information [when people answer questions about their experiences, beliefs, or behaviors] can often be misleading because participants may feel pressure to give an answer that they think program implementers would like [e.g., Did you enjoy the program? Did it help you?], or participants might not know the answer [e.g., Did your literacy improve because of the program?]. Whenever possible, it is best to combine these types of questions with more objective approaches, like assessments of skills.

2 DID EDUCATION IMPROVE FOR THOSE WHO DID NOT PARTICIPATE IN THE PROGRAM?

To know whether a program is effective, we need to know what would have happened in its absence. This is known as a *counterfactual*. For example, if test scores increase by the end of a program, we need to know how much of that increase is due to the program, and how much of it would have occurred anyway, even without the program. It is particularly important when working on a topic such as education, and when working with young people, because there are likely to be improvements over time due to many other factors (e.g., regular school, life experience, cognitive development).

How can an evaluation answer this question?

The goal is to collect information on a group of similar individuals (or similar groups, such as schools) who did not receive the intervention (*control group*), or who received a different intervention (*comparison group*). The best way to do this is by using a random method to select who participates in the program (see question 3). But there are other options for finding a comparison group if *randomization* is not possible, such as:

- Students from other schools in the same district
- Students who are a year ahead or a year behind in school
- Students in the same class who did not participate
- Schools from neighboring districts

Whatever the method, finding a group that is as similar as possible to those who participated in the program is key. Seemingly small differences between groups, such as varying distances from roads, can distort the findings in important ways, making the effects of the program less clear.

What if an evaluation does not answer this question?

Even without changes in policies or interventions, many young people will develop stronger literacy skills, or progress through school, or acquire more knowledge. Therefore, observing improvements among those who participated in a program, without a comparison group, can be misleading. *Evaluations that lack a good comparison group tend to overestimate the effects of programs.*

"Selectivity" is one of the most common threats to program evaluations. It occurs when those who participate in programs are different from those who do not participate. For example, program participants may come from wealthier households with more educated parents, they may themselves be more motivated to learn than their peers, or they may come from schools where teachers are working to address gender-related barriers to education. The simple act of joining the program signals that these participants may have more opportunity or motivation than their peers who do not join. Including only the most eager students will only tell you how well the program works for the most eager students but may not tell you how well it works for those who may be most in need.

3 WHO PARTICIPATED IN THE PROGRAM, AND WHY?

As described, to understand the effects of a program, the group participating in the program must be as similar as possible to the group that is not participating.

How can an evaluation answer this question?

The safest way to ensure groups are comparable is through *randomization*. Researchers use a random method (e.g., a coin flip) to decide who joins a program (or is invited to join) and who does not. This can be done for individuals or groups (schools, communities, etc.). Researchers then collect data on both groups, preferably before and after the program takes place. If randomizing students or schools to participate in a program is not possible, or not desirable, researchers can also randomize the timing of a program in different communities, so that everyone has a chance to participate eventually. In this case, those who participate in the program later can serve as the control group for those who participate earlier.

What if an evaluation does not answer this question?

Although it is the gold standard, randomization is costly and time-consuming and may be undesirable in some circumstances. There are some ways that researchers can try to address lack of randomization in their analyses:

- Assess whether those who participated more (e.g., attended more regularly) have better outcomes than those who did not (dose-response relationship). This will give some idea of whether participating in the program is related to improvements in education.
- Explore how participants are different from their peers who did not participate. Are participants wealthier? Did they have stronger school performance? This will provide some insight into how well the program might work for others.
- Explore how similar individuals compare. Identify pairs or groups of students who are similar in every way except that some participated in the program and some did not and compare their levels of education after the program.

Evaluations without randomization may show that programs are more effective than they really are, because the most eager participants are likely to join, and those same participants are also most likely to experience improvements. *Evaluations without randomization tend to overestimate the effects of programs.*

Although we focus on using quantitative information about whether a program achieved its intended goals, there are other important questions to ask about all programs and policies, including:

- **Targeting:** Ensuring that a program is reaching the intended audience, e.g., girls most at risk of dropping out of school.
- **Monitoring:** Regular collection of data on program activities (e.g., number of people trained) to ensure a program is going as planned.
- **Costing:** Tracking information about the costs of implementing a program, which can be combined with evaluation findings to understand cost-effectiveness of activities.

FIVE EVALUATION FAQS

1. Will this program work for other people? Even when an evaluation is done well and a program is found to be effective, that does not mean it will be effective in other places, with different people, or at another time. This is called *generalizability*. When interpreting the results of an evaluation it is helpful to ask what is different about this group that might influence whether the program is effective. For example, do they live in a country where the government is especially supportive of girls' education? Or do they live in a community where child marriage is common?

2. Which parts of the program are most effective? Many policies and programs have multiple parts. For example, a program might include community engagement meetings, scholarships, and teacher trainings. An evaluation that finds the whole program to be effective does not necessarily tell you which part is effective, which may be important information for decisions about whether and how to reach more girls with the same program (*scale-up*). If information about which part of the program is most effective is important, it must be integrated into the evaluation design from the beginning.

3. Why should we separate results by sex if we're not implementing a "gender" program? Even if a program or policy is designed to help both girls and boys, there might be differences in whether and how it works for each group. For example, a school feeding program for all students might have a bigger effect for girls if parents in poorer households are prioritizing education for boys. And even if there are no differences in the effects for boys and girls, that is still useful information for decisionmakers.

4. What if evaluations find conflicting results about the same program? Often, different evaluations of similar programs will find seemingly contradictory results. This could be due to differences in the program design or implementation, in context or participants, in analyses, or in many other factors. When possible, look for a systematic review, which summarizes findings across many evaluations to come up with an estimate of how well the program works on average. Also, try to find evaluations of the program done in similar contexts to the ones where you will be working.

5. What if we can't conduct an evaluation? Not all organizations are interested in or able to evaluate their work. Not every program needs to be evaluated, but all programs should take steps to ensure they reach the right people, and provide the right services. For example, when piloting a new idea that lacks evidence, it may make sense to try it out on a small scale to demonstrate feasibility – that is, is it possible and do participants like it – before engaging in a larger evaluation. Or, if an evaluation is not planned, then it is even more important to ensure that a program is designed and implemented based on existing evidence of what works.

KEY EVALUATION TERMS

Baseline measures: Measures of outcome-related variables (e.g., level of literacy, grade attainment, experience of violence) taken before a program is implemented. *Endline measures* of the same outcome-related variables are taken after a program is implemented.

Comparison group: A group of people (or schools, communities, etc.) who are not exposed to a program, and who are compared with the group exposed to the program. Sometimes the comparison group receives no program at all (control group), while sometimes the comparison group receives the standard of care or a different program.

Counterfactual: The outcomes (e.g., level of literacy, grade attainment, experience of violence) that would have happened without the implementation of the program.

Evaluation: The systematic assessment of a program or policy.

Generalizability: Also known as external validity, it is the extent to which the results of an evaluation can be generalized to other times, other people, other treatments, and other places.

Randomization: Most commonly used in randomized controlled trials, a method of randomly assigning people (or schools, communities) to program and control groups and comparing the groups in terms of outcome variables.

Selection/selectivity: A challenge sometimes faced in evaluations where differences between the program group and the control group before the program is implemented could account for observed differences in outcomes between the program and control group.

Self-reported data: Information that is reported by study participants, usually in response to a survey. This is in contrast to other forms of data, such as biomarker data (e.g., weighing someone on a scale or blood tests) or assessments of skills (e.g., literacy, numeracy).

Systematic review: A structured comparison of evaluations that is intended to distill common themes or summarize evidence that pertains to a research question.

ADDITIONAL RESOURCES ON EVALUATION

Check out the following resources for more information:

Abdul Latif Jameel Poverty Action Lab (J-PAL). n.d. "Introduction to Evaluations." https://www.povertyactionlab.org/research-resources/introduction-evaluations

Frankel, Nina and Anastasia Gage. 2007. *M&E Fundamentals: A Self-Guided Mini-Course.* https://www.measureevaluation.org/resources/publications/ms-07-20-en

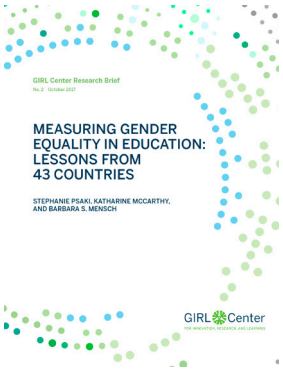
McDavid, J.C. and L.R.L. Hawthorn. 2006. *Program Evaluation & Performance Measurement: An Introduction to Practice.* Thousand Oaks, CA: SAGE Publications.

Shadish, W.R., T.D. Cook, and D.T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Belmont, CA: Wadsworth Cengage Learning.

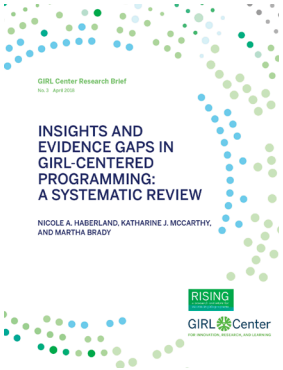
United Nations Evaluation Group. 2005. "2005 Standards for Evaluation in the UN System (updated 2016 Norms and Standards are available)." http://www.uneval.org/document/detail/22

OTHER RESOURCES FROM THE GIRL CENTER

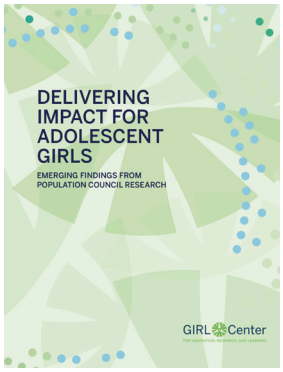
Chae, Sophia and Thoai D. Ngo. 2017. "The Global State of Evidence on Interventions to Prevent Child Marriage," GIRL Center Research Brief No. 1. New York: Population Council.



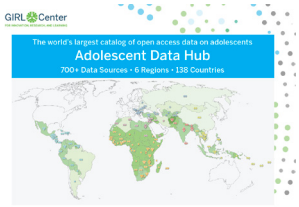
Psaki, Stephanie, Katharine McCarthy, and Barbara S. Mensch. 2017. "Measuring Gender Equality in Education: Lessons from 43 Countries," GIRL Center Research Brief No. 2. New York: Population Council.



Haberland, Nicole A., Katharine J. McCarthy, and Martha Brady. 2018. "Insights and Evidence Gaps in Girl-Centered Programming: A Systematic Review," GIRL Center Research Brief No. 3. New York: Population Council.



"Delivering Impact for Adolescent Girls Emerging Findings From Population Council Research," 2018. New York: Population Council.



The Adolescent Data Hub is a unique global portal to share and access data on adolescents living in low and middle-income countries. It is home to the world's largest collection of data on adolescents and serves as a resource to facilitate data sharing, research transparency, and a more collaborative research environment to drive continued progress for adolescents.

GIRL Center

FOR INNOVATION, RESEARCH, AND LEARNING

The Girl Innovation, Research, and Learning (GIRL) Center generates, synthesizes, and translates evidence to transform the lives of adolescent girls

popcouncil.org/girlcenter