HARVARD Kennedy School
MALCOLM WIENER CENTER
for Social Policy

What is Measurement and Evaluation?

The Importance of Anchoring Evidence Generation in Your Theory of Change

Julie Boatright Wilson
Harry S. Kahn Senior Lecturer in Social Policy

WHAT IS MEASUREMENT AND EVALUATION?
THE IMPORTANCE OF ANCHORING EVIDENCE GENERATION IN YOUR THEORY OF CHANGE

Julie Boatright Wilson
Harvard Kennedy School

Everyone wants you to produce evidence about the effectiveness of your non-profit organization or the program you run. Your funders want evidence that you are accomplishing what you said you would and not wasting their money. And if you have more than one funder, it is likely that each is demanding a different type of evidence. If the situation in your country is like that in the United States, your government wants all sorts of evidence about how you use money and what you do. Many of your managers probably want evidence that the organizational changes you are proposing will make a big enough difference to be worth the work they will need to do. Even some of the people to whom you provide services want evidence that what you are doing or asking them to do will work. And because you are here, we know you want to keep learning what works so you can improve what you do and how you do it in order to support more children and families.

But what do we mean by evidence?

If we are honest with ourselves, most of us would have to admit that we are confused about what constitutes evidence and uncertain how to produce and use it. To complicate it further, all of us have seen data we don't trust or data that are "cherry picked" to make a political point. We have also seen high-priced scientific studies that seem to carry a lot of weight. So we are struggling not just with what constitutes evidence but also with what constitutes good evidence. And with whether we can produce this evidence ourselves or need to hire high-priced researchers to do it for us.

Part of our challenge is that as managers we have many needs for evidence. Some evidence would be very useful for understanding the effectiveness of how we manage our organization while other types or forms of evidence would be useful for understanding whether what we do makes a positive difference.

A second part of our challenge is that some of the evidence we really want will take years to emerge. We have a lot of questions about the long-term effectiveness of our programs but we also need to know if our programs are being run as designed and whether the short-term impacts they aim for are occurring.

Finally, the actual process of gathering data to generate evidence can be a challenge. Some data are potentially easy to gather while other data are extremely difficult or impossible to generate. And many of us have staff who are terrific at delivering services but don't know much about how to provide accurate data on what they are doing or for whom they are providing services.

These challenges are real. It is difficult to know what an overall evidence generating strategy for our organization should include, where we should start, and how we should implement it. But non-profit management has never been a career for the faint of heart. We can do this. What we need is a framework for thinking about evidence and strategies for generating it. And as managers, if we want a high performing organization, we need a strategy for transforming our organizational culture into one that is data driven. The purpose of this memo is to help you move forward on these goals by laying out a framework for thinking about various types of evidence.

The complexity of defining success in non-profits

We should start by acknowledging that your jobs are complex. One of the things we know is that those of you who run non-profit organizations face a lot of pressures that business people don't face. Three of them are particularly important for today's discussion.

First, for-profit businesses have a straight-forward metric for measuring their success – profit vs. loss. This means that if you run a for-profit business and people don't like your product, you get pretty quick feedback as sales decline. On the other hand, if your profits are large, you can feel confident that your customers value what you produce. Most non-profit organizations don't have an easy metric for measuring their success. For example, how do you measure the value of an art gallery? In addition, non-profits don't generally have rapid internal feed-back loops. We don't have the equivalent of quick information on profits to let us know if we are on the right track.

Second, all for-profit businesses can be measured by the same outcome – profit. We might be interested in market share or sales growth as well, but profit is the driving metric. For non-profits, in contrast, no common metric is available. We might like to see a bit more sharing of outcome measures among non-profits in a given sector, but even there it can be difficult to find a common outcome that is measured in the same way by each organization.

Third, for businesses, those who fund the organization are also those who use its services. Although some businesses might be eligible for subsidies of various sorts, most of the money generated comes directly from those who purchase the goods or services produced. This is not the case for most non-profit organizations. Those who provide our funding are not usually the individuals who use our services. Thus, one of the big challenges non-profits face is figuring out how to effectively convince funders that we are good stewards of the resources they have given us and simultaneously demonstrate that our services are having a positive impact on those who use them.

So, where do you start if you want to measure the effectiveness of your non-profit organization? We need to figure out both how success should be defined and then how to measure it.

Question Zero

The place to start this process is to clarify what it is your organization is trying to accomplish. To do this, we focus on something we call Question Zero. Question Zero is the challenge of stating in as close as possible to ten words – or less – what the goal of your intervention is – what you want to be held accountable for accomplishing. Why the emphasis on ten words or less?

Most non-profit organizations have mission statements that lay out very lofty ambitions. This enables these organizations both to motivate internal actors to keep on doing what needs to be done in the face of limited resources and other hardships and to build support on the part of those outside our organization who might potentially be allies in achieving these goals. But these mission statements, so

important for building support for our efforts, may not provide much guidance to us in measuring our effectiveness. For example, take the case of the Educational Volunteers Foundation of Turkey. This organization provides after-school programs to low-income Turkish school-children and uses volunteers as instructors. This is a complex undertaking and we are inspired by their ambitions as laid out in their mission statement:

> Create and implement educational programs and extracurricular activities for children aged 7-16, so that they can acquire skills, knowledge and attitudes supporting their development as rational, responsible, self-confident, peace-loving, inquisitive, cognizant, creative individuals who are against any kind of discrimination, respect diversity and are committed to the basic principles of the Turkish Republic.

But we also recognize that it will be exceedingly difficult for TEGV to measure the impact of its activities on each of these outcomes. For example, how might TEGV measure the effect of participation in afterschool programs on the development of a more peace-loving nature or on a commitment to the basic principles of the Turkish Republic? This is not to say that these are not worthy goals. Rather, it is to say that they are challenging to measure. It is even more challenging to assess what part of the increase in one's peace-loving nature can be attributed to participation in a TEGV after-school program. We need something more specific if we want to measure the effectiveness of TEGV's efforts and hold it accountable over the short- as well as long-term for its use of resources.

This is where Question Zero becomes very important. If we step back a bit and think about what TEGV does and why, we might come up with a Question Zero like the following:

> **Improve long-term child outcomes through volunteer-led after-school education enrichment programs**

This is a very clear statement of what TEGV does and hopes to accomplish by carrying out these activities. It doesn't define long-term child outcomes the way the mission statement does. Nor does it define "enrichment." But it provides a clear picture of the goals and the mechanism for achieving them.

Developing your Question Zero is hard work and normally takes time. In the process you are likely to discover that individuals in your non-profit who are responsible for different functional parts of the organization have different ideas of what the organization's Question Zero is. We can illustrate this by going back to the example of TEGV.

The Question Zero I laid out is a very clear statement of what TEGV does and why. But it is a statement of what TEGV does from the perspective of its educational mission. TEGV has a secondary mission to increase the capacity of the voluntary sector in Turkey by recruiting and training volunteers and deploying them to do important work that could not otherwise be done. From that perspective, a Question Zero might be:

> **Increase Turkish volunteerism through opportunities to lead after-school enrichment programs**

A former student reported that when he returned to his organization after learning about Question Zero, he gathered his senior staff, gave each a piece of paper, explained the concept , and asked each to write down his or her understanding of the organization's Question Zero. He also asked that the

respondents not put their names on their papers. He then gathered and shuffled the papers, re-circulated them, and asked each person to explain the Question Zero on the paper he or she was given. It turned out that each person's Question Zero was slightly different.

Why should this matter? Why should you start here? And why do you care if your others in your organization are in agreement? Because, if you are not clear about what you are trying to accomplish, you will never really know if you have been successful in achieving your mission. You need to be very clear what your Question Zero is. And if you have more than one Question Zero, you need to be clear about that and make sure you are addressing each of them.

Theory of Change

Once we are clear about what we want to accomplish, we need to articulate the path we will take to reach that goal as well as define what we mean by success. We will start with the path. We call this path or road map our Theory of Change. A Theory of Change is a series of IF…, THEN…. statements that articulate in detail what we will do. For TEGV, as for many organizations, there are likely to be several theories of change because of its multiple missions – think of our two Questions Zero – and because of the diversity in the population of school-children for whom it provides services. This makes sense because the developmental needs and appropriate activities for the youngest segment of school-children TEGV serves are very different than those for the oldest. But we might start with the following basic theory of change for school-child involvement:

> **IF** we create safe and inviting environments for afterschool activities, **THEN** school-children will feel comfortable coming to our after-school programs (or, their parents will feel comfortable letting them come to our programs);
>
> **IF** school-children feel comfortable coming to our after-school programs, **THEN** we will be able to offer them activities that interest them;
>
> **IF** we offer activities that interest school-children, **THEN** they will want to keep coming to our after-school program;
>
> **IF** school-children keep coming to our after-school program, **THEN** we will be able to build enrichment activities into the afterschool programs that attract them;
>
> **IF** we can build enrichment activities into the afterschool programs that attract them, **THEN** we will be able to help them develop skills in X that they otherwise would not have had the opportunity to develop;
>
> **IF** we can help them develop skills in X, **THEN** they will be better prepared for Y

There is a lot embedded in those statements and yet, if you run an organization, you have already figured out that your theories of change need to be much more detailed than what I have presented. To the extent possible, your theory of change should draw on evidence. In the case of school-age programs, we have a great deal of evidence about child and adolescent development that can guide us in determining what skills and capacities young people should have developed at certain ages and what types of activities help develop these skills and capacities. It will take time to develop your theories of change. TEGV would also need to develop a theory of change for recruiting, training and retaining

volunteer teachers. For this goal, there is less evidence-based practice on which to draw, so the full Theory of Change might have to be developed over time as the organization figures out what works.

You should also expect that your theories of change will evolve both as the environment around you changes and as you learn more about the effectiveness of your programs for various subgroups of the population you serve. It is this sort of evidence that will bring the specificity to your theories of change that will enable you to define "success."

Logic Model

As we said, the theory of change is your road map. It lays out your best guess, based on evidence and common sense, of what the path to accomplishing your Question Zero goals will be. But your theory of change does not tell you in detail what resources you will need in order to accomplish your mission. It does not lay out the details of the activities you need to undertake to be successful. And it does not delineate the specifics of what these activities will produce. Those details are addressed in the Logic Model, depicted below. The logic model lays out the resources you need to implement your intervention, the activities you will undertake, and what the immediate effects of these activities should be. It also pushes you to articulate in a measureable way what the longer-term outcomes of your efforts should be.

Logic Model

I will illustrate this with two examples. First, if we think about creating safe and inviting environments for afterschool activities, we need to have a good sense of what we mean by "safe" and what we mean by "inviting." Safe would include spaces without obvious hazards such as harmful substances or substandard construction as well as freedom from harassment or harm by peers, teachers, or individuals not involved in the program. The definition of an inviting environment is likely to be very different for school-children of different ages. This says a lot about the type of space and equipment we need to procure as well as the type of supervision and amount of supervision we need to have in place. The lesson here is that in order to measure whether our after-school environments are safe and inviting, we need to define those terms in detail.

The inputs for providing safe and inviting environments would include not only the size and types of space and numbers of personnel with specific skills for working with school children but also the financial resources needed to procure the space and staff and the type of staff needed to undertake such activities as routine cleaning, inspections and maintenance.

Our outputs are the immediate results of these activities. In this case, we would want to gather information on such things as maintenance activities, complaints about unsafe conditions and injuries, and bullying episodes. This type of data or evidence is usually gathered by counting and categorizing. These counts are often referred to as performance metrics. You as managers convert performance

metrics into performance management tools when you discuss them with your staff and use them to make changes in the ways staff carry out their jobs or are rewarded.

But those metrics only measure the safety of school-children who came. We speculated that if we created a safe and inviting environment, all youth would feel comfortable coming to our program. How would we know if there were some groups of youth who were not participating because they did not feel safe or welcome? This is a difficult question because we normally can't count what didn't happen. We might start by comparing the characteristics of the school-children coming to our program with the characteristics of our target population. Are any groups – characterized by age, ethnicity or religious affiliation, sex, geographic location of residence, school performance, etc. – under- represented? If so, how would we find out why? We know that to answer this question we would have to find a way to gather information on perceptions of safety and interest among students who came and those who didn't. This is doable, but will take time and resources.

Performance metrics are a form of evidence. They are particularly useful if we are measuring the "right" performance because they can assure us that we are implementing our programs with fidelity. But they are just the first step. We also want to generate evidence that participation in our afterschool programs makes a difference in the lives of these students. We can illustrate this through another example.

Imagine that TEGV has created an afterschool program focused on building robots out of Legos – the kind of robots that can perform simple tasks like picking up objects and transporting them to another location. The program is designed for 12-14-year-olds with the goal of increasing their interest in science and math. The theory of change is that if they begin to see science and math as fun and useful they will be more likely than they otherwise would have been to take additional science and math courses, study harder and perform better in these courses, and they will be more likely to choose to pursue careers that involve science and math than they would have been if they had not participated in our after-school program. This is asking a lot of one 6-week program, but we will come back to that issue later.

If we go back to our logic model, we would start with the inputs. Of course we need the Legos and other equipment as well as a space that has access to electrical power, tables for building and testing robots, and enough security to ensure that the robot projects will be undisturbed when the students are not there. And we need the money for this. But we also need to recruit volunteer teachers with enough science and math background to direct this program and with enough understanding of the developmental needs of this age group and enough interpersonal skills to manage the program. And, of course, we need to recruit students.

The program process involves dividing the students into small working groups and, during their daily sessions after school every day for the six weeks of the program, teaching them about the basics of robot construction and providing opportunities for them to build, experiment, test and rebuild their robots. It may involve demonstrations of group products or even competitions. And it certainly involves teaching some basics of working in groups and helping students with group problem solving.

What should you measure to test the effectiveness of this program? We might start with the immediate outputs. First, what are the attendance patterns? Once they started, did the students keep coming? How engaged were they when they came? Did they actually build any robots? Did they have fun? Did the groups manage to get along and work out any differences? There are a lot of useful output measures we could gather, so we need to figure out which ones are important.

But your goal is to do more than just provide an after-school activity to fill their time. You want to influence their behavior. Your outputs, a little farther away in time, are also important. You probably want to know if the participating students began working harder in math and science classes. If so, you would expect that their grades would go up so you probably want to look at their grades before and after participating in your program. You might want to know if, when they had a chance to select courses, they chose to take extra math or science courses. You might want to know if they are more interested in a career in a math- or science-intensive field or now have a better idea of what that type of career might be. And you might even want to know if this pattern of interests and behavior persisted for a few years.

Suppose you followed these students, gathered data from their school transcripts to check the courses they took and their grades and interviewed them about career plans. Suppose you found that they were indeed taking more math and science classes than their peers and getting better grades. And suppose you found that their interest in careers demanding math and science backgrounds persisted. Can you take credit for this?

The answer is: we don't know. Maybe you can. Maybe you can't.

The reason we don't know is that we don't have any measure of the counterfactual – we don't know what would have happened if they had not participated in this wonderful 6-week afterschool program in robotic construction. Let's imagine three different scenarios.

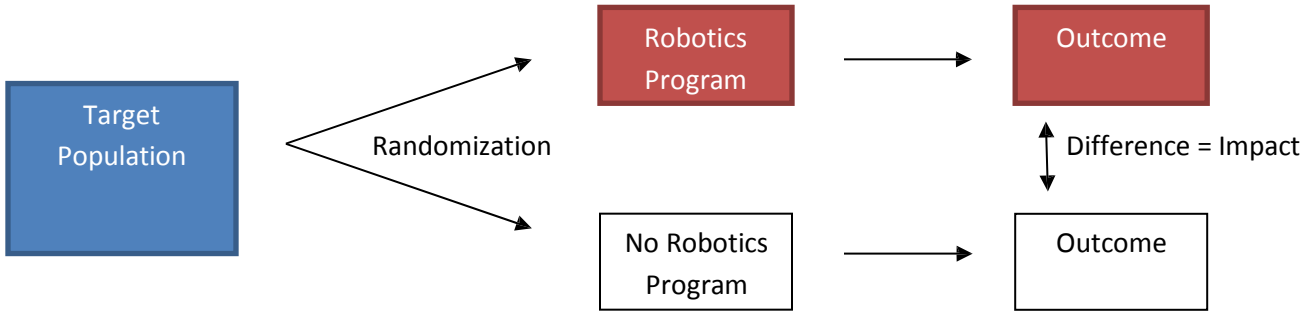
- It could be that only the students who were already really interested in and good at science and math signed up for you robotics program. If so, you know they had a good time and, while they were in your care were not off doing things that would get them in trouble. But you probably can't take credit for their later academic success or career interests.

- On the other hand, suppose the participants were students who had always done well in science or math but never really liked it or never thought about a career in a math- or science-intensive field. If after participating in your program, they changed their views, signed up for more science and math courses than they had planned to, and were thinking more seriously about math- and science-intensive careers, you would feel comfortable that you had made a difference.

- Yet a third group might be students struggling in math and science courses who were about to give up because they thought the material was too challenging and they couldn't figure out how working on these topics would help them when they finish school. They participate in your program, figure out that math and science really can be fun, and begin to try to figure out how they might be able to do more of it. The volunteer tutor, in talking with them, helps them think this through and gives them tips on how to study math and science. Their grades go up and they take more courses than they otherwise would have.

In the second and third examples it is highly likely that your program made a real difference. We call the difference your program made the "impact of your program." But you can't prove your program had an impact unless you have a counterfactual – a group of students who are like the participating students in every way except that they didn't participate in your afterschool robotics program.

It is the "alike in every way except that they didn't participate in your program" that is challenging. Where do you find this "alike in every other way" group? It is easiest if many more students sign up for your program than you are able to enroll and you hold a lottery to randomly select which students get into the program. Everyone who signed up was equally aware of the program and equally motivated to participate. You can then compare those who participated with those who did not, even following them for several years to gather information on grades, courses taken and career plans. Some of the non-participants will find other science- or math-related things to do but others in this group will not. The fact that some participated in other science or math activities is normally fine because the question is: what would have happened if your program had not been offered. We would worry a bit if the other programs they participated in were also your programs, but we will come back to that question later.



But what if you are able to accommodate every interested student in your afterschool program on robotics? Then what should you do? With whom should you compare the participating students? Should you just compare them with students of the same age in another afterschool program? You can, but it won't be very meaningful. Suppose you were to compare this group of students with another group of students the same age who took an afterschool program in dance? We would have a lot of reasons to think that there might be some fundamental differences between students who choose to take a dance class and those who choose to take a robotics class and that this difference is what would influence the amount of math and science they chose to take and how hard they worked at those subjects – not your robotics program.

It is not always easy to find an appropriate counterfactual. In some cases it is possible to estimate the impact without a counterfactual. For example, there are sophisticated statistical techniques available to attempt to construct a comparison group. There are sophisticated research designs in combination with statistical techniques that provide a great deal of analytic power for measuring impact. The important point is that you can only be certain that participation in your program is what made the difference if you can compare participants with a group that is as similar to it as possible.

A randomized experiment like the one I described is a very powerful tool, but it has some limitations. The most important limitation is what we call generalizability. Your experiment with the robotics class might show that participation leads to a statistically significant increase in performance in math and science courses and career choices for students like those in your program. It does not mean that you would necessarily see the same impact in another group of students – older students, younger students,

or even students of the same age in a different country.[1] You would have to run a similar experiment to see if you got the same results.

There is one other important point about experiments that is particularly important for managers of programs and organizations. What if you run an experiment and find that your program, the one you carefully designed and thought through, made no difference in student performance in math and science or subsequent career plans. Should you conclude that the program was a failure?

Maybe. But maybe not.

It could be that your Theory of Change is correct and this program is one you should support, but you didn't implement the program as it was supposed to be implemented. You didn't have the right equipment or you didn't have enough equipment. A flu epidemic or weather disaster may have kept students and volunteer teachers from attending regularly. Carefully monitoring the implementation of your program – the day-to-day operations – and assessing the fidelity with which your program was implemented is called process evaluation. The right performance metrics combined with observation will help you answer the question of whether you actually implemented the program as intended. If you didn't, you can't really know if it would have worked or not. All you know is that it did not work in the manner in which it was implemented.

But suppose you ran the program precisely as intended and you found no impact. It could be that you have the wrong Theory of Change. There are ways to increase interest in math and science, but this may not be one of those ways. Or it could be that the basic idea is a good one, but it takes longer than six weeks working on robotics to really influence 12- to 14-year olds. Your next steps are reviewing the literature, talking to experts in the field, and experimenting with different versions of the programs.

The point is that you can never stop asking the "why" and "how" questions.

<u>One final point</u>

I want to go back to the example of the TEGV mission statement that was discussed earlier.

> Create and implement educational programs and extracurricular activities for children aged 7-16, so that they can acquire skills, knowledge and attitudes supporting their development as rational, responsible, self-confident, peace-loving, inquisitive, cognizant, creative individuals who are against any kind of discrimination, respect diversity and are committed to the basic principles of the Turkish Republic.

These are important goals. They are also very large goals – the type of goal that no single organization can accomplish on its own. One of the values of a mission statement that sets broad goals is that it may inspire other organizations that could be part of achieving these goals to join the effort through their own activities. But that is very different from being clear about your role in achieving these goals and implementing your role effectively.

[1] The technical terms for these are internal validity and external validity. Internal validity refers to our confidence that the differences we observed between the two groups could not have occurred by chance more than, for example, five times out of a 100 times running the course. External validity refers to the extent to which we can say that such a difference would be observed for other groups or in other settings.

Likewise, some of TEGV's experiments in after-school programming might get picked up and incorporated into the required school curriculum. This raises the question of how we could measure our collective impact on the educational outcomes and career options for Turkish school children. We don't have a lot of good models or methods for doing this yet. But this is clearly where we need to go if we want to achieve the missions of non-profit organizations and scale up effective interventions.

Conclusion

Effectively generating and using data from "real-time" programs is difficult. It takes strong management, committed staff, accurate record keeping, and sensible data analysts who understand the importance of using data to better manage individual cases as well as how to aggregate data in ways that let both managers and the public at large know what is being accomplished. It also takes careful planning and meticulous attention to detail. It demands treating individual cases with confidence while sharing data on operations so that those involved can work together more effectively. It means being accountable on many levels. This is challenging work but it is very important for achieving our mission and giving the public a sense that their resources are being used wisely.