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
John F. Kennedy School of Government, Harvard University

If you manage a nonprofit organization, everyone wants you to produce evidence about the effectiveness of your organization or the specific programs you offer. Funders want evidence that your organization is accomplishing what you said it would and not wasting their money. And if you have more than one funder, it is likely that each is demanding a slightly different type of evidence. Various government departments want evidence about how your organization uses money and what you do as well. Many of your managers and front-line staff want evidence that organizational changes you are proposing will make a big enough difference to be worth the work they will need to do to implement them. Even the people to whom you provide services want evidence that what you are doing or asking them to do will work.

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 of our organization's effectiveness and are uncertain about how to go about producing and using it. 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 scientific studies that seem to carry a lot of weight but also a hefty price tag. So we are struggling not just with what constitutes evidence but also with what constitutes good evidence that we can afford to produce and that will be useful to us.

Part of the challenge is that managers have many needs for evidence. Some evidence would be very useful for understanding the effectiveness of how well our organization is managed while other types or forms of evidence would be useful for understanding whether what the organization does makes a positive difference in the lives of those it serves.

A second part of the challenge is that some of the evidence that would be useful may take years to emerge. We have a lot of questions about the long-term outcomes for people who participate in our programs; we hope their lives will be better than they otherwise would have been had they not participated in our activities. But as managers, we also need to know if our programs are being run as designed and whether the short-term impacts we are aiming to achieve 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 do not know as much about or feel comfortable providing data on what they are doing or for whom they are providing services.

These challenges make it difficult to know what an overall evidence generating strategy for an organization should include, where to start, and how to implement it. But non-profit management has never been a career for the faint of heart. Managers can generate evidence on the effectiveness of their

organizations. What is needed is a framework for thinking about evidence and strategies for generating it. And as managers, if you want a high performing organization, you need a strategy for transforming your organizational culture into one that is data driven

The complexity of defining success in non-profits

We need to start by acknowledging that the job of managing a non-profit is complex. Non-profit managers face a lot of pressures that people who manage for-profit businesses don't face. Three of them are particularly important.

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 an art gallery generates? In addition, non-profits don't generally have rapid internal feed-back loops. They don't have the equivalent of quick information on profits to let them know if they 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 or other measures 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 across non-profits in a given sector, but even in that situation it can be difficult to find a common outcome that could be 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 funding are not usually the individuals who use the services that funding generates. Thus, one of the big challenges non-profits face is figuring out how to effectively convince funders that they are good stewards of the resources they have given and simultaneously demonstrate that their 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? You need to figure out how success should be defined for your organization and then how to measure your success in achieving 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 the organization who might potentially be allies in achieving these goals. But these mission statements, so important for building support for an organization's efforts, may not provide much guidance to for

measuring effectiveness. As an example, take the case of the Educational Volunteers Foundation of Turkey (TEGV). This organization provides after-school programs to low-income Turkish school-children and uses volunteers, primarily college students, as instructors. This is a complex undertaking and the organization's ambitions as laid out in their mission statement are inspiring:

> 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 could be attributed to participation in a TEGV after-school program. TEGV needs more specific goals to measure the effectiveness of its efforts and hold itself 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.

The Question Zero laid out above 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 second Question Zero might be:

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

TEGV has two Questions Zero and as a result is trying to maximize performance on two agendas. This is a very important insight. In this case the two goals support one another, but it is possible that an organization will have competing Questions Zero.

Developing an organization's Question Zero is hard work and normally takes time. In the process a manager is likely to discover that individuals 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 the example of a manager from another organization who convened his senior staff, explained the

concept of Question Zero, gave every staff member a piece of paper and had each write down his or her understanding of the organization's Question Zero. In this case, his organization had only one Question Zero. He also asked that the staff not put 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. His conclusion was that his senior staff was not in agreement about the goals of the organization. For a manager, this is important information.

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 goals. You need to be very clear what your Question Zero is.

Theory of Change

Once your organization is clear about what it wants to accomplish, it needs to articulate the path it will take to reach that goal as well as define what it means by success. The path or road map is an organization's Theory of Change. A Theory of Change is a series of IF…, THEN…. statements that articulate in detail what the organization will do. For TEGV, as for many organizations, there are likely to be several theories of change because of multiple missions – think of the 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 student 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 activities that attract them;
>
> **IF** we can build enrichment activities into the afterschool activities 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

A lot is embedded in this series of statements and yet anyone who manages an organization will quickly figure out that a theory of change needs to be much more detailed than what is presented here. Indeed, an organization may have "cascading" theories of change. For example, it might be useful to

develop a theory of change for designing and administering specific after-school enrichment activities and another for recruiting students.

To the extent possible, a 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 managers in determining what skills and capacities young people should have developed at certain ages and what types of activities are likely to be most effective in helping children and youth develop these skills and capacities. But TEGV would need to develop a second theory of change for its second Question Zero – recruiting, training and retaining volunteer teachers. For this goal, there is likely to be less evidence-based practice on which to draw, so its full theory of change might need to be developed over time as TEGV figures out what works and what does not work.

An organization's management should expect that its theories of change may evolve over time as the environment in which it operates changes and as the organization learns more about what is most effective in specific programs for various subgroups of the population it serves. This is the type of evidence that will bring specificity to an organization's theories of change and help define "success."

Logic Model

The theory of change is an organization's road map. It lays out management's best guess, based on evidence and common sense, of the most effective path to accomplishing its Question Zero. But the theory of change does not provide any details about what resources are needed in order to move forward along this path. It does not lay out the details of the activities an organization needs 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 is a framework for identifying what resources will be needed to implement a program, what activities will be undertaken, and what the immediate effects of these activities should be. In addition to pushing management to be specific about the here and now, working through a logic model pushes management to articulate in a measureable way what the longer-term outcomes of the organization's efforts should be.

Logic Model

We can 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 quite different for students of varying ages. The logic model framework pushes TEGV's

management to be specific about the type of space and equipment needed as well as the type and amount of supervision necessary to insure safety. In order to measure whether TEGV's after school spaces are safe and inviting, it needs to define these settings 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 administration, cleaning, and maintenance.

The process or activities are the details of what will be done – what the specific afterschool activities will be, how they will be managed, etc. The outputs are the immediate results of these activities. Very often measuring outputs involves counting and categorizing. In this case, TEGV would want to gather information on such things as maintenance activities, complaints about unsafe conditions and injuries, and bullying episodes. It would also want to count the number of activities undertaken and student attendance. 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 assess performance or make changes in the ways staff carry out their jobs or are rewarded. In thinking about what to measure, it is important to ground performance metrics in the Theory of Change.

The metrics listed above measure the safety of school-children who came to TEGV's programs. However. the Theory of Change posited that if TEGV created a safe and inviting environment, all youth would feel comfortable coming to their after school programs. How would TEGV know if there were some groups of youth who were not coming because they did not feel safe or welcome? Or because they did not find the environment inviting? This is a difficult question because we normally cannot "see" or count what did not happen. TEGV might start by comparing the characteristics of the school-children attending its programs to the characteristics of the 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 TEGV find out who is not coming and why? To answer that question the organization would have to find a way to gather information on awareness of its program on the part of those who did not attend as well as perceptions of safety and interest among the students who came and those who did not. If those who did not attend would have participated if they had known about the program, TEGV might choose to focus on advertising. If those students did not come because they did not think they would be safe or feel comfortable, TEGV has a different set of issues to address. This type of research is very useful for improving program design and operation but takes more time and resources than needed for the performance measures discussed above.

Performance metrics are a form of evidence. They are particularly useful for assessing whether a program is being implemented with fidelity. But as important as this is, such metrics are just the first step. TEGV also wants to generate evidence that participation in its afterschool programs is making 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 more likely to choose to pursue careers that involve science and math than they would have been if they had not participated in the 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 TEGV needs 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 they need the money for this. But TEGV also needs 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, TEGV needs to recruit students.

The program process involves dividing the students into small working groups and, during their 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 TEGV measure to test the effectiveness of this program? They 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 TEGV could gather so they need to figure out which ones are important.

But TEGV's goal is to do more than just provide an after-school activity to fill students' time. TEGV wants to influence their behavior. This means that outcomes, a little farther away in time, are also important. TEGV probably wants to know if the participating students began working harder in math and science classes. If so, they would expect that participating students' grades would go up so TEGV probably wants to look at their grades before and after participating in the program. TEGV might want to know if, when participating students had a chance to select courses, they chose to take extra math and science courses. They might want to know if the participants 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 they might even want to know if this pattern of interests and behavior persisted for a few years.

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

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

The reason we don't know is that TEGV does not have any measure of the counterfactual – they do not know what would have happened if participating students 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 the robotics program. If so, TEGV knows they had a good time and

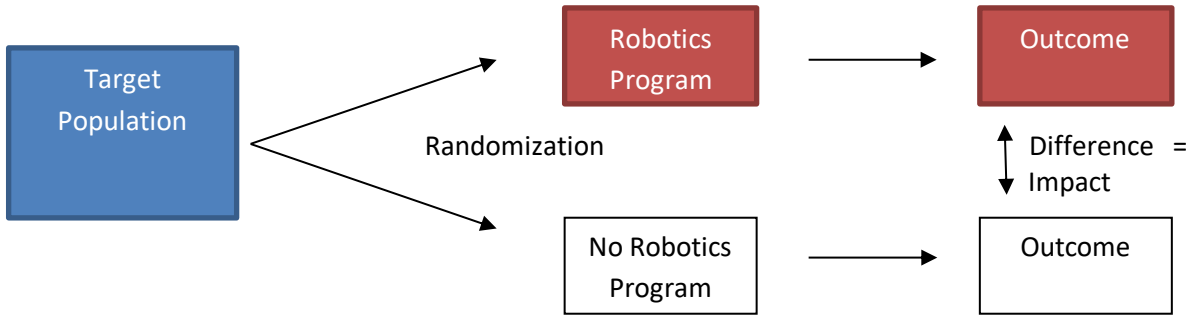
that while they were participating in the program were not off doing things that would get them in trouble. But TEGV probably cannot take credit for their later academic success or career interests.

- 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 the robotics 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, TEGV would feel comfortable that it 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 could not figure out how working on these topics would help them when they finished school. Suppose that through participating in the robotics program, they figured out that math and science really could be fun and began to think they might want to do more of this kind of work. Suppose the volunteer tutors, in talking with them, helped them think this through and gave them tips on how to study math and science. And then suppose their grades went up after participating in the program and they took more math and science courses than they otherwise would have. TEGV would want to take credit for this change.

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

It is the "alike in every way except that they didn't participate in the program" that is challenging. Where does TEGV find this "alike in every way" group? It is easiest if many more students sign up for the program than TEGV is able to enroll and it holds 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. TEGV 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 will not. The fact that some participated in other science or math activities is normally fine because the question is: what would have happened to these students if your program had not been offered. If your program had not been available, interested students would have found those other programs. We might worry a bit if the other programs they participated in were also TEGV's programs, but we will come back to that question later.

But what if TEGV is able to accommodate every interested student in its afterschool program on robotics? Then with whom should it compare the participating students? Should it just compare them with students of the same age in another afterschool program? TEGV could do that, but it might not be very meaningful. Suppose they compared this group of students with another group of students the same age who took an afterschool program at the same time which focused on 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 in later years and how hard they worked at those subjects – not the 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 comparison groups. 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 TEGV can only be certain that participation in its program is what made the difference if it can compare participants with a group that is as similar to it as possible on all dimension except one – they did not participate in the robotics program.

A randomized experiment like the one just described is a very powerful tool, but it has some limitations. The most important limitation is what we call generalizability. TEGV's experiment with the robotics class might show that participation leads to a statistically significant increase in performance in math and science courses and changes in career choices for students. It does not mean that we 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] TEGV or other organizations would have to run a similar experiment to see if they got the same results with different groups of students in different settings.

A second point about experiments is that they tell us that "on average" participants in our program performed better on our outcome measures than those who did not participate. But "on average" may mask a great deal of heterogeneity. Some participants might have gotten a boost from participation in TEGV's program while others might have gotten no boost or even been turned off to science and math. Some studies have found that in the United States many adolescent girls who have talent in science and math do not pursue careers in those fields. Imagine if TEGV discovered by looking more deeply into its program that participation did not make much difference for boys but had a huge positive impact on the interest in math and science careers among girls. This would be an important finding.

[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.

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

Maybe. But maybe not.

It could be that the Theory of Change was correct and this program is one TEGV should support, but that the organization did not implement the program as it was supposed to be implemented. Perhaps they did not have the right equipment or they did not have enough equipment. A flu epidemic might have kept students and volunteer teachers from attending regularly. Carefully monitoring the implementation of a program, sometimes called process evaluation or implementation evaluation, is important. The right performance metrics combined with observation will help you answer the question of whether your organization actually implemented the program as intended – or as researchers would say, with fidelity. If you did not, you cannot really know if it would have had a positive impact or not. All you know is that it did not make a difference in the desired outcomes in the manner in which it was implemented.

But suppose TEGV ran the program precisely as intended and found no impact. It could be that they 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. TEGV's next steps would be to review the literature, talk to experts in the field, and experiment with different versions of the programs – testing, modifying, testing – until they get it right. Or it could be that positive effects occur only for those school children who participated in a series of programs that enabled them to explore a wide range of interests and develop a diverse set of skills over a long period of time. This is a more complicated hypothesis to test, but it can be done.

The point is that as a manager you should never stop asking the "why" and "how" questions. And you should never stop asking for evidence and mapping that evidence back into your theory of change.

One final point

As discussed earlier, TEGV has a noble and visionary 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.

These are important goals. They are also very large goals – the type of goal that no single non-profit 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. Or they might copy and extend the reach of TEGV's programs. For example, after seeing the evidence of the effectiveness of some of TEGV's after school programs, Turkish primary and secondary schools might incorporate some of these activities into their curriculum.

Or other non-profit programs might expand access to TEGV youth to activities that complement what TEGV is doing.

This raises the question of how we could measure our collective efforts in improving the educational outcomes and career options for Turkish school children. We do not have a lot of good models or methods for assessing collective efficacy yet. But this is clearly where we need to go if we want to achieve the missions of non-profit organizations. And such an endeavor is likely to have a complex Theory of Change!

Conclusion

Effectively generating and using data from "real-time" programs is very important for those of us managing non-profit organizations or programs. Our tasks are too large and too important and our resources normally too small for us to do otherwise. Moving forward to develop an organizational culture that effectively generates and uses evidence will take strong management, committed staff, timely and accurate data, and sensible data analysis. And it will demand an organizational culture that is always asking questions: Did this work? For whom? What is it about the program that caused the changes we observed? With whom or where was it effective? When? Why? At what cost? And how could we make the program even more effective?