**Performance Measurement and Evidence in the Social Sector:**

**Proving and Improving**

**September 2015[[1]](#footnote-1)**

***Introduction***

Increasingly managers are being asked to produce evidence about the effectiveness of their organization’s activities. Funders want to know that an organization is accomplishing what it said it would and not wasting money. Likewise, governments want a wide range of evidence about how organizations use money and what they do. Those who work in organizations often want evidence that the organizational changes a manager is proposing will make a big enough difference to be worth the new responsibilities they will need to take on. Even the people to whom an organization provides services may want evidence that what it is doing or asking them to do will work. Finally, effective managers and their staff want to keep learning what works so that they can improve what they do and how they do it. Performance measurement in the social sector is both about *proving* that your efforts are successful and *improving* your activities.

Determining what constitutes evidence for efforts focused on creating social value and driving social change can be challenging; developing strategies for producing and using this evidence can be difficult and confusing. Part of the challenge is that as managers we have many different 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 challenge is that some of the evidence we want may take years to emerge. We have a lot of questions about the long-term outcomes for people who participate in our programs. But as managers, we also need to know if our programs are being run as designed and whether the short-term outcomes 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 while staff might be terrific at delivering services, they might not know much about how to provide accurate data on what they are doing or for whom they are providing services.

In this note we propose a four-step framework for thinking about what constitutes evidence of effective social programs and what strategies we can use to generate it. Our goal is to help managers create high performing organizations whose organizational culture is data driven and whose goals include both proving the social value they create and improving their programs and impact.

***The complexity of defining and measuring success***

While all for-profit businesses have a straight-forward metric for measuring their success – namely profit, organizations focused on creating social value do not. A for-profit business gets pretty quick direct feedback from its customers: if customers perceive value in the product or service offered, market share, sales and profits all increase; if not, sales and importantly profits, decline. Most non-profit organizations and government programs don’t have an easy and quick metric for measuring their success. For example, how do you measure the value of an art exhibit or a new park? For businesses, those who fund the organization are also those who use its services – the customer. Social value producing entities might not have this direct feed-back loop that allows them to tap into stakeholder sentiment. Those who fund the service organization may not be the same people who use its services. And while all for-profit businesses can be measured by the same outcome (profits), nonprofits and government agencies have no such common metric. So, where do you start if you want to measure the effectiveness of your social change efforts? Before we can measure effectiveness, we need to be able to define success.

***STEP 1: Question Zero***

The place to start this process is to clarify what your organization or program 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, and what you want to be held accountable for accomplishing. Throughout this note, we will be using the example of Educational Volunteers Foundation of Turkey (TEGV), a nonprofit organization that provides free after-school programs to low-income Turkish school-children using volunteers, primarily college students and young adults, as instructors. This is a complex undertaking and we are inspired by their ambitions as laid out in their mission statement to:

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 or that there might be a connection between what TEGV does and the subsequent characteristics and behavior of the children and youth to whom they provide services. 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 concise 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 piece of evidence for assessing overall organizational effectiveness.

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. In this case, his organization had only one 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. It also turned out 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 aligned and clear about what you are trying to accomplish, you will not know what to measure and you will never know if you are successful. You need to be very clear what your Question Zero is.

***STEP 2: 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 to achieve our Question Zero. For TEGV, as for many organizations, there are likely to be several theories of change because of its multiple objectives – 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 welcoming and safe settings, **THEN** children and youth will feel comfortable coming to our after-school 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 them, **THEN** children and yout will want to keep coming to our programs;

**IF** school-children keep coming to our after-school program, **THEN** we will be able to build a series of enrichment activities 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

Many theories of change will be a lot more detailed than the one illustrated above. Their role is to articulate the causal relationship, often building on implicit assumptions, between the actions we take and objectives we seek. 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 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.

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.”

***STEP 3: Logic Model***

A theory of change is your road map. It lays out your best guess, based on evidence, common sense and assumptions, of what the path to accomplishing your Question Zero goals is. But your theory of change does not tell you in detail what resources you will need in order to move forward along this path. 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 captured in the Logic Model, depicted below. In addition to pushing you to be specific about resources and activities, working through your logic model pushes you to articulate in a measureable way what the longer-term outcomes of your efforts should be.

**Logic Model**

*Inputs/Resources Process/Activities Outputs Outcomes*

*Needed*

Building on the TEGV example above, we might think about the inputs and activities needed to create an environment in which students feel comfortable coming to our afterschool activities. The creation of this comfortable environment can help determine the type of space and equipment we need to procure, the type and amount of supervision we need to have in place and the financial resources required for both. The key here is that in order to measure whether our after-school environments are comfortable (something defined in our theory of change as important to achieving our Question Zero), we need to define this in detail.

Outputs are the immediate results of our activities and very often measuring outputs involves counting and categorizing. In this case, we would want to gather information not only on the number of children attending our programs but on outputs that help us track the extent to which we are creating a comfortable environment: attendance, facility maintenance activities, complaints about unsafe conditions and injuries, and bullying episodes. These counts are often referred to as performance metrics and managers can convert performance metrics into performance management tools for staff, using them to make changes in the ways staff carry out their jobs or are rewarded. Performance metrics are a form of evidence and are particularly useful because they can help us assess whether 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.

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 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. These are our *activities*.

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 the goal is to do more than just provide an after-school activity to fill their time. We want to influence their behavior. *Outcomes*, a little farther away in time, are also important. We 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 we probably want to look at their grades before and after participating in your program. We might want to know if, when they had a chance to select courses, they chose to take extra math and science courses. We 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 we might even want to know if this pattern of interests and behavior persisted for a few years. Suppose we 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 we found that they were indeed taking more math and science classes than their peers and getting better grades. And suppose we found that their interest in careers demanding math and science backgrounds persisted. Can we take credit for this?

The answer is: we don’t know. Maybe we can. Maybe we can’t.

***STEP 4: Understanding Impact***

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 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 who are about to give up because they think the material is too challenging and they can’t figure out how working on these topics will 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 this. 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 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 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.

Outcome

Robotics Program

p

Target Population

 Randomization Difference = Impact

Outcome

No Robotics Program

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 subsequently 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 program participants with a group that is as similar to it as possible on all dimension except one – they didn’t participate in your program.

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? You should not necessarily conclude that this is the case.

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 kept students and volunteer teachers from attending regularly. Carefully monitoring the implementation of your program is sometimes 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 – or as researchers would say, with fidelity. 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 – testing, modifying, testing – until you 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 of success that can help you both prove and improve your programs.

***Conclusion***

Managers of organizations whose mission and mandate is to create social value and change, often (like TEGV) have broad 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, schools might incorporate some of these activities into their curriculum. Or other programs might expand access to their activities that complement what TEGV is doing. This raises the question of how we could measure our collective efficacy in improving the educational outcomes and career options for Turkish school children. We don’t have a lot of good models or methods for assessing collective efficacy yet. But this is clearly where we need to go if organizations working in the social sector want to achieve their missions.

Effectively generating and using data from “real-time” programs is very important for those of us managing organizations or programs focused on creating social value. Our tasks are too large and too important, and our resources normally too small, for us to do otherwise. But 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 was it about the program caused the changes we observed? Where was it effective? When? Why? At what cost? And how could we make the program even more effective? It requires generating evidence and using performance measurement to both prove and improve our work by developing and articulating a Question Zero, Theories of Change and clear Logic Models.

1. This note was written by Julie Boatright Wilson and Nathalie Laidler Kylander in consultation with MLD faculty and the MLD Research and Curriculum Development team to serve as an introductory conceptual framework for performance measurement and management at the Harvard Kennedy School. [↑](#footnote-ref-1)