# Evaluation Issues: A Practical Discussion for Small Agencies and Not-for-Profit Organizations[1]

Special Topics : Working Paper #2012-01
August 2012

www.rit.edu/cpsi

Center for
Public Safety
Initiatives

John Klofas, PhD
Professor and Director, Center for Public Safety Initiatives
Rochester Institute of Technology
John.klofas@rit.edu
585-475-2423

Janelle Duda, MSW
Assistant Director, Center for Public Safety Initiatives
Jmdgcj@rit.edu
585-475-5591

Center for Public Safety Initiatives
cpsi@rit.edu

[1] We are grateful to the agency and not-for-profit community members who gave us the idea to write this and to all those who commented on an earlier draft of the paper.

The goal of this paper is to be helpful to small agencies and not-for-profit organizations as they address the issues of program evaluation. Today almost anyone affiliated with a program or intervention -of nearly any kind- is likely to find themselves dealing with the issue of evaluation. No doubt many program leaders are anxious to document their program’s effectiveness; for others, tightening budgets may drive efforts to assure continued support from funders and fund raisers. And, who is not tempted by the spreadsheet on their computer. "Data" sounds and feels easy, so why not do so? Then, of course the call for evaluation is fueled by the constant barrage of "evidence-based" or "best practice" programs that get called-for and, occasionally, cited in the journals, newsletters and, now, webinars. Finally, it is easy to be drawn in by the sophisticated and complicated evaluations that get covered in a minute or two on NPR or the evening news. The siren call of evaluation is difficult to ignore.

But hazards can lie ahead if you answer that call. Evaluation can be complicated. The results may be disappointing, especially if expectations are unrealistic. Regardless of results, critics may lay in wait. Evaluation is not for the timid- a lesson too often painfully learned.

But even with those warnings you may find yourself propelled forward, into the breach. That is not necessarily a bad thing, if you can keep the right frame of mind. The important question is: how can you get the most out of moving ahead? That is what we want to help you answer with this brief paper. In doing so our goal is not to address the technical questions; there are better places for that, and the truth is, despite their "technical" label, those are the easy ones. You can find answers to those questions in program evaluation textbooks and websites. The more difficult and, we think, the more important issues, are conceptual. The right perspective in an evaluation can help you develop a process of self-guided analysis, of useful questioning of what you do, and of constant improvement so that whatever program you oversee, or whatever service you provide, will be as sound as possible and always striving to be better still.

Our approach in this paper will be to try to anticipate the questions that are, or at least might be, on your mind and to provide as clear an answer to them as we can. Of course, in the interests of full disclosure, we must admit that our answers suggest greater clarity and certainty than we actually have. This should be seen as both a cautionary note and as encouragement to think things through carefully on your own.

**1. What do you mean by evaluation?**

First, get it out of your head that it is a study, or a thing of any kind. Sure you can approach it that way and wind up with a great thumbs-up, or a less great thumbs down, for your program. But no one learns much from those kinds of exercises- or perhaps, the lesson is often not an informative or welcome one. Instead, it is more useful to recognize evaluation as a process. In that process you can continuously clarify and sharpen your mission, improve measurement of your processes and their outcomes and respond to what you have learned only to reexamine it later, seeking a new and better response.

Our preferred view, then, is to see evaluation as a key process in an organization which continuously shapes and reshapes itself based on data. This is a problem solving approach to evaluation and we recommend it. The questions that follow and, of course, their answers, all flow from that perspective.

But let's not move too quickly. Our view of evaluation in a learning organization is not the only one and if you buy into it you should also consider what you are giving up. The critical question to be addressed is always “evaluation for what?” We believe in evaluation for continuous program improvement. But, if you think of evaluation as a simple assessment of effectiveness there are other issues to consider. In that context, evaluation is always about the distribution of resources. Evaluations demonstrating effectiveness can be used to garner resources. Evaluations that fail to demonstrate effectiveness can, and probably should, be expected to fail to attract resources or can, and probably should, lead to their withdrawal. However, that is not the end of the story. Take this path and you must also ask the corollary question, “effectiveness for whom?” That should make clear the importance of purpose: who is the piper who gets to pick the tune? Who will determine what rate of recidivism is good or bad? How will a graduation rate be judged? What will your outcomes be compared with? These are the practical consequences to thinking this way; context is everything. How you, and those around you, think about evaluation is the most important evaluation issue you face.

## 2. What will you evaluate?

This question may sound silly, but the most difficult problem faced by evaluators is often figuring out what the program actually is. That is, it is very common that people running programs, and who are interested in their evaluation, cannot describe specifically how their program is supposed to work. Their hearts may be pure and their intentions good, but they may not be able to tell you precisely what they think their program is supposed to change or how it is supposed to bring about such a result.

Sure that sounds strange, but as you know, in the world of human services needs are great, demand for services seems unlimited, and the consequences of not providing services can be tragic; it’s easy not to connect the dots. But the "need to do something," while not an unreasonable foundation for a program, doesn't necessarily lend itself to evaluation. And too, "meeting the needs of individuals" is generally not a sufficient explanation of what that "something" is. How needs are identified, the range of needs that might be recognized, and the range of specific services used to meet them, are all important to the evaluation process.

An important point, however, is not just that you need to be specific in how your program is supposed to work, but also in what it is supposed to do. That is, evaluation almost always involves some assessment of outcomes. Specifying what you are trying to change is of critical importance. And that means both being able to explain the program goal as a concept that makes sense, and also considering how that general purpose will be translated into measureable outcomes.

Take, for example, the common use of recidivism rates as outcome measures for treatment interventions in criminal justice. The idea of falling back or returning to crime may easily be seen as a reasonable indicator of program success or failure for individuals. But there are many ways to measure recidivism, many factors outside of a program itself that will influence the outcomes, and few clear guides to assessing whether any given rate of recidivism should be regarded as good or bad.

Using arrest as an indicator of failure will produce much higher recidivism rates than conviction, and longer follow-up periods will always reveal higher failure rates than shorter time periods. Although it may seem obvious that the key to success is producing lower rates of recidivism than would occur without the program, how can you know what the "naturally" occurring rates would be? You could find yourself taking pride in the low recidivism rates of the graduates of your program that treats elderly female homicide offenders, but if you are treating that population to reduce their recidivism you are wasting your time and money- they don't need your program. But a 50% failure rate of young male property offenders three years after their release from prison would likely mean your program is a solid success.

So the lesson is: you have to do your homework. You have heard of the concept of a logic model and you need to do one. You need to specify exactly what you are trying to change, for what specific population of program participants’, and precisely what steps you will take to try to accomplish that result. Yes this is big picture stuff. It is really about theory; theory about what causes a problem and what can fix it. But in practical terms, the questions are: how is the program supposed to work and what will it do if it does work?

**3. How can you keep the big picture in mind?**

Outcomes are important, as is the design of the program that you expect to affect them, but in our favored approach to evaluation, you also want to pay attention to implementation. You will want to understand why something works and what works best. And, if things don't work out as well as you hoped, since our goal is always improvement, you also want to know why- was it the idea or the execution? So along with outcomes there are two other key things to pay attention to: assessing the structure and the processes of the program.

First, in the structure component of an evaluation you want to ask whether the program is actually set up to do what you want it to. This sounds like an obvious issue but this is where a lot of programs collapse. Do you have enough staff? Are they in the right positions? Are they the right staff, with the right training and backgrounds? If you are trying to deliver outreach services, do you have enough outreach workers to make a difference or is the demand so overwhelming that their work cannot be focused enough to make a difference? If caseloads are the method of service delivery, what is their actual size? What about when you consider vacations and other time off?

Second, the process component of the evaluation extends these sorts of questions. Is the staff there when you need them? Does their 9 to 5 with weekends-off schedule make sense? What are they actually doing when they are working? For example, a common finding in studies of small classroom and small caseload programs has been that, if expectations are not clear, teachers and caseworkers may do pretty much what they did with larger classrooms and larger caseloads. In sum: execution matters.

These issues, from the design of programs to their structure and processes, all address the issue of program fidelity. In the end you want to do an evaluation in which you can fully describe what it is your program actually does, how it squares with the plan, whether that plan was yours from the start or if you borrowed it from someone else. Knowing all this will help you learn as much as you can from the evaluation and it will also allow others to learn from your experience, just as you should be trying to learn from the programs of others. This is how we will all make progress together.

## 4. What is research design and how do you get ready for it?

If you are focused on outcomes, then when all is said and done there are two major questions an evaluation will try to answer: did the things the program seeks to change actually change? And, if they did, did the program cause the changes?

Answering the first question usually requires you to identify changes over time. In some cases this means measuring what the program is intended to change before and after delivering program services. That is your basic pre-test/post-test research design. When it's possible, a stronger design involves collecting information at regular intervals for a period of time before and after the program intervention. This design relies on what evaluators call *time-series data*. With such data, you can see if and when things change and for how long these changes last. How long a time frame do you need? That will depend on what you are trying to do, but between 18-months and three years of monthly observations before and after an intervention program starts may be reasonable for analysis.

But seeing change does necessarily mean you caused it to happen. One of the common problems in evaluating crime reduction programs has been that changes that seem to result from the program also occur in areas where no program has been implemented. This can happen when a program is started at a time when crime starts to fall or is in decline for reasons not related to the program. It is a more common evaluation problem than you may think. The rub is that it is often spikes in crime that prompt the need and interest in starting crime reduction programs in the first place. In many cases spikes are followed by returns to more normal levels. A program can look like it worked but it might simply have caught the downward trend. The problem is so common that it has a name- *regression to the mean*. It is one of the first things evaluation researchers look for.

Even when things look good, you still have to connect the dots and show that it was the program that brought about the change. That generally means the use of comparison groups - or *control groups* if you want the technical term. The idea is that you need to find groups who are similar to your program group but didn’t get the program. For people, this can mean similar ages, genders, education and other background characteristics. For places, it may mean matching areas on such things as social characteristics, wealth and poverty, and crime. If your group shows changes and the others do not, then you are in a better position to argue that the program caused the change- if not, then it’s back to the drawing board.

What all of this means to you is that you need to start thinking about what information you need to collect and when you need to collect it. Some measures can be pulled together long after the fact. Crime rates are a good example of such a measure. Most police agencies today should be able to provide data for specific offense types well into the past and they can often do so even at the neighborhood level. For other measures - especially at the individual level - things can be much more difficult. If you are tracking test scores, or problem solving ability, or time watching TV, or most other things involving individual behavior, you need to know what these were like before your participants started a program. That can be difficult. This is where pre-testing with survey instruments can be helpful.

Through all of this try to remember that no one knows as much about your program, its participants or the area that it is in, than you. Maybe it will help you to work with a researcher to figure some of this out, but you should always lead the way. Know what you want to measure, know its strengths and weaknesses, and start working through the process of collecting the information as soon as you can. Analysis will come later. You don't want to leave these decisions to someone else, even if they are skilled researchers, and you don't want to ignore this until late in the game. Never underestimate how much you know about your program and how important that understanding will be to your evaluation.

## 5. What are you going to measure?

There are a few issues researchers worry about when it comes to measuring things and you should be aware of them as well. The two most critical concepts are known as *validity* and *reliability*. The *validity* of a measures deals with the degree to which the measurement accurately reflects some underlying concept. So, the relevant question is something like this: how good as a measure of program success is an improved attitude, or not being arrested for six months or, for that matter, anything else you are tempted to choose.

*Reliability* deals with whether others will see things the way you do and that they will do so over time. That is, do other ways of assessing things lead to the same conclusion? Will measuring success with self-reported behavior surveys or with official measures give you the same results and will those results be consistent across time? These issues are important because you can't always pick perfect measures. Instead you will just have to make the best choices you can but being concerned with validity and reliability can help guide those choices. This is one reason it is often good to use tested and proven ways of defining and measuring outcomes. Things like re-arrest, conviction, graduation rate, or employment status are all well-known concepts. In many cases the validity and reliability of such commonly used outcomes have already been widely recognized; but again, it is important to note if these are indeed accurate measures for your program.

But there are also costs to any choice you make. Here is an example to consider. Recidivism is a well-known concept, often seen as valid and reliable, even if there is not necessarily much agreement on its details, specifically the criterion for failure and the length of time you should track someone to see if they fail. But the most important issues may be more basic than that. Most lives are more complicated than the limited choices described as success or failure. It is clear, for example, that some parolees commit serious crimes but are never arrested. Others avoid arrest despite desperate lives of homelessness and drug addiction. Yet all would be counted among the successes. And too, some parolees live mostly crime-free lives but find themselves arrested for minor offenses and are, therefore, counted among program failures.

It is easy to see that much of life occurs along its margins, with marginal success and marginal failure common but mostly uncounted. This is a big problem for evaluation, especially for evaluation that wants to be as informative as possible. That is why interviews, case notes, narratives and detailed life histories of program participants can be very important. So don't think of measuring outcomes as only numbers and statistics. Think of measurement, instead, as the ways you describe the reality of the lives of people who go through your program. The concerns with validity and reliability still apply but that is a much more useful approach when the goal of evaluation is continuous program improvement.

Finally, there is one more measurement-related issue to wrestle with, and it is an important one. How will you know if your program has been successful? After all, there aren't many miracle cures out there. Hard work is hard work. You should be skeptical of anyone who says the answers are easy. So, the practical question is: how good is good and how bad must it be to be judged unsuccessful? Statisticians have a way of answering that. They call it *statistical significance* and it has to do with the likelihood that the results you find could have happened simply by chance. If you beat chance, even by what seems like a little bit, you have been successful. Those small differences may not look important to some audiences but if your sample is large enough, they are evidence that the program is working. So it is not at all uncommon for graduates of sound intervention programs to have success rates that are a few percentages better than those who never went through the program. Those differences count, not just in the evaluation report but more importantly, in the human suffering avoided and the costs that are saved. Don't let anyone underestimate their value. As we noted earlier, if you improve on what would have happened without you; that is success. If you use evaluation to keep improving, that is even better.

**6. How will you deal with the question of "Best Practices?"**

Today the discussion of "best practices" or its close cousin, "evidence-based practice," is almost as common as the discussion of evaluation itself. The idea makes perfect sense: let's go with programs that have proven themselves, especially those already supported by existing evaluations. But this simple idea gets a lot more complicated when you look more closely. First, we really don't know that much about what works. For many reasons, few programs have been tested. In most social service intervention areas there is not much of a compilation of programs known to work. Our knowledge of things that don't work is a bit better, but there is generally no menu of program designs to select or to avoid. And even among those interventions that seem promising based on strong evaluations, few have been repeated or replicated; and those that have often either failed or revealed complications that limit their usefulness in new locations. The classic example in criminal justice is programs of mandatory arrest by police in cases of domestic violence. While originally linked to reductions in further violence and additional calls to the police, those findings were not replicated in all the cities that were studied and, in some places, domestic violence incidents actually increased following arrest. Findings like that are a caution against universal application and a call for attention to the differences across communities.

It is not that there hasn’t been good work done to compile what we know and don't know about what works and what doesn't, at least under some circumstances. An ongoing helpful effort in criminal justice is put together by the US Bureau of Justice Assistance and is available on the web at www.crimesolutions.gov . There have been lots of similar efforts in a wide variety of fields. Anyone involved in developing or assessing intervention programs should consult those in their subject area. But even with informative reviews of intervention strategies, the real problem is that things often don't go according to plan, as much as we would wish they would. The work you do is more complicated than rocket science. Don't deny it. If it weren't, social problems would have been solved before a man landed on the moon.

So the idea is that you shouldn't just pull a program off the shelf and expect it to work even if someone has identified it as a best practice. The program will need to match the problems you have and the resources you have *and* you will need to evaluate it as it has been implemented with your own clients in your own community. Fidelity to an established program design is important, but so is learning from your own experience. Paying attention to all of that is the best way to use the knowledge base that currently exists and the best way to add to it.

But, let's be careful here. The point is not that you don't need to pay attention to the research on existing programs- quite the opposite is true. A sound foundation in theory, expressed in a clear model of program logic, and understanding the experience of programs that came before yours, will be invaluable to your efforts. Just bear in mind that situations are never the same as the ones that came before them. Program models must change and adapt to improve, and part of your obligation should be adding, not just to your knowledge, but also to the knowledge base of others who follow behind you. Best practices won't remain “best” for long, and a commitment to improving your program over time through evaluation will serve you better than the unquestioning adoption of a program from somewhere else. Remember this isn't rocket science. Things don't change as quickly up there as they do down here. Adopting "best practices" is indeed related to problem solving approaches to program design and implementation, but they are not the same thing.

**7. Summing up: When it is all put together, the steps below make sense to us.**

1) Start thinking about evaluation as early as you can. The very best time is well before a program starts, but it is never too late to get serious about it.
2) Find out as much as you can about what has been done already. What theory guides practice in the area? What is the state of best practice and evidence-based practice in your field?
3) Find a research partner or at least someone with research expertise that you can talk to. Look at a local college or university. Look for someone who has technical skills and who also recognizes your program skills and knowledge. Take him or her to lunch. A lasting relationship could be very beneficial.
4) Tell everyone you talk to that your goal is to use evaluation to continuously improve things not to give the program a thumbs up (or, I suppose, a thumbs down). Mean it.
5) Start your work by writing up a complete logic model of the program. Be specific. Write down what are you trying to change and how will you do it.
6) As soon as you think you have any idea of what you are doing and why, start to collect data that is relevant to what you are planning to measure. Worry about analysis later.
7) Set the wheels in motion. Start by assessing program structure. Is it set up to do what you want?
8) Now look at process. Is it actually functioning in a way that could produce good results?

9) Now it is off to the races. Use the data to assess outcomes.

10) Write it all up- completely. Be as thorough as you can and look at everything, every way you can. Graph the results.    Pictures help.

11) Talk about it. Engage everyone you can- staff, clients, board members, interested citizens, even your mother.    Encourage everyone to help you understand it and to act on it.

12) Figure out the lessons from your evaluation. Write them up and write up the action plan that flows from them.

13) Implement it.

14) Go back to the top of this list and start again.