The Goldilocks Challenge: learning to structure a 'right-fit' and effective monitoring and evaluation system

A written capture of the webinar delivered by Dean Karlan, Professor of Economics, Yale University and President, Innovations for Poverty Action. Please note that we had technical difficulties at the start of the webinar, Therefore the written capture of the webinar begins at slide 9.

(Slide 9)

There is some strong evidence. Maybe there was a study once in some other place, but we are not quite sure whether it applies to us here in Ontario. Maybe it’s similar to things that were done elsewhere, but not exactly the same, and so we don’t feel quite comfortable saying, oh, this evidence elsewhere is just so spot on to what we’re doing that we can use it. Instead, we say, no. I want to know that this implementation and the way we designed it here, this adaptation is working. And so then the question is, okay. Well, can you generate it? And that’s not an obvious yes in all cases. I’m going to talk through some examples. But the question is can you do a credible impact evaluation? I’m using our term that I’m going to come to later, credible, that’s part of the CART principles.

And again, if you cannot credibly answer the question, did you cause a change to happen in the world, then we go in this No box and we say then fine. Just do the monitoring. The last thing you want to do is waste money on impact evaluations that are not going to be convincing to anybody because of methodological issues. But if you can generate it, well that’s the world in which you do both monitoring and impact evaluation. We are in the bottom right square. And that’s a world in which not only can you find out the impact of what you’re doing, but it’s done well and actually do it in a way that helps other people learn from what you’re doing and scale up the policy elsewhere. So that’s our basic flow chart for deciding do you do monitoring or monitoring and evaluation.

Notice that there’s really no end point here that is impact evaluation and no monitoring, or no monitoring, no impact evaluation. Monitoring is something that for the most part should be easily built in to just about any program even at the minimal level, and it’s just an accountability exercise to make sure that things were being done that were supposed to be being done.

(Slide 10)

When is an impact evaluation appropriate? This is going to be a little bit on the long here side. One issue is sample size. This is often the deal breaker in a lot of situations. And when we say sample size part of what we mean here is that there’s enough separable units. By separable units, what I mean is if you are running a program that is – I will use an extreme example – about trade policy, that is a very hard thing to separate people who are - or communities that are affected by the policy from communities that are not going to be affected from the policy. So you need some way of separating what we call treatment and control in order to go – you know, not just how has the program changed from the people who did experience the program, but
how lives would have changed had the program not been experienced, not been part of people’s lives. And so that’s what you need, you know, some separation.

A training program is a very easy example of when you can get easy separability because a lot of times training programs, there might be a thousand spots for a training program and there might be 1,500 people that apply and there’s a scarce resource so it’s only a thousand people that can get it and then you can do a random assignment where a thousand get the training and 500 do not, or maybe get it later or something of that nature. But these are separable units in the sense that they – unless you worry about people getting this training and then immediately teaching everybody around them – this includes the control group – these are separable units. We could do an entire week on what that means, and so I’m going to skip out there and stop and just to say that that is a technical criteria that is absolutely critical and, you know, and we can dig into it if there’s Q & A on it, or if not, that is the type of thing that we then spend a lot of time talking about in more extended sessions.

A second is that the timing is right and this goes back to the – a little bit like the tinkering conversation, the tinkering point I made earlier that you don’t want – you know, you want the program to be well enough to find but it does need to be that there’s some expansion of the program happening. Something is happening to grow the program, and as part of that growth, that is when you can build in the creation of control groups so that you can understand what the impact of your program is. You know, that is usually not too much – it usually does not weed out too many programs because for the most part, people usually are not that interested in understanding the impact of something that is not growing. If it is growing, it means there is some enthusiasm for it, some ideas to put this and to, you know, to try to scale this up, and then there is a natural constraint that then one is using that natural constraint of the growth of the program to try to construct treatment and control groups. If the program is shrinking, then it is usually not going to be something that generates lots of enthusiasm for finding out whether it is working.

The last is important, and obviously is a hand raising point, but it should be that there’s a knowledge gap worth filling. That this is actually an important question, not just for the – not just for the context of this program, but that for others, and that is really the gap. There have not been 17 studies of exactly this program that already tell us what the impact is. And so then the 18th doesn’t really add to our knowledge. And that’s an importance to consider, because we do not want to see too many evaluations that are then – then this money spent on evaluation that could be spent on programs.

(Slide 11)

There’s some practical issues. It does require a budget and capacity to do a proper impact evaluation. Often non-experimental methods cost just as much as a randomized trial. The thing to remember is that a randomized trial is actually not any more expensive than a non-randomized study that still goes out and does surveys. The thing that is expensive about doing research is surveying. A lot of times people think randomized trials are expensive, and they are when there’s heavy surveying involved, but the randomization is not what’s expensive. It is just the idea of going out and surveying people to find out what is happening in their lives. That costs money. There are many things out there now though where you got the data and administer the data that come out of some either banking data, to see savings levels, things like
that. And then there’s limited or no survey data and then randomized trials are very cheap. One thing that is important to remember is that a randomized trial is not what makes things expensive. With the cost driver for impact evaluations is whether or not there are going to be surveys done to ask people about their life and how it’s changed. Or you don’t ask people how their life has changed. You just ask people how their life is and then you compare those who got a program to those who do not.

Another practical issue is that the partner cares deeply about knowing their impact. This is not to be taken lightly. There’s a recurring theme of my failures as a researcher is when I engage in projects and the partner is not all that enthused about doing it. It is not only unpleasant to do, but more often than not, does lead to a failed project. That things just go amok. It is important that everybody is excited about knowing ‘what are we doing’, ‘what is the impact we’re having’.

(�Slide 12)

Then it’s not appropriate, as I mentioned, when the units are separable and the sample size is too small or there’s macro policy, when spillovers are rampant. Spillovers would be a training program in which we think that people who get trained are going to teach everybody around them. Or it could be if it’s a public health issue with communicable diseases, then it’s not to say you can’t do it, but you have to think a little bit more creatively about how one is going to set up a randomized trial. Or sometimes there’s a resource constraint that just can’t be randomized very well. A refugee camp. We have deal with a lot of situations in developing countries where there’s refugee camps and these are very hard settings to set up randomized trials. And the ethnical constraint that’s most important is when we know the answer already, so we don’t keep running randomized trials on vaccines because we know they work and continuing to evaluate would be unethical. So that it's important to understand, what’s the knowledge gap? That should drive the ethics decision.

(�Slide 13)

Back to some of the big picture concepts here. The core problem we’re looking at is too little data, too much data and the wrong data. Now how do we strike this?

(�Slide 14)

Here’s a great example of what I mean by the management problem of too much data. A very simple study done in a high-end supermarket in California with jams. People were given little samples in the grocery store and in one setting they had a choice of six jams in the other one they had a choice of something in the order of 24 jams.

(�Slide 15 & 16)

They counted how many people stopped and of course when you have 24 jams a lot more people stopped and tasted the jams, than the six jams. But the kicker is how many purchased.
And when there were 24 jams up for display, taste any of them you want and then buy, lots of people stopped, and nobody bought, or very few. When there are only six jams available for tasting, then fewer people stopped but a lot more people purchased as you can see from the graphs.

The point being here that when you give people too many choices, that sometimes you just get stagnation. You get people being overwhelmed and then no choice is made at all. Whereas with fewer choices, sometimes you actually see more action taking place. Too much data has the same exact problem. When you overload people with data, there is overwhelming amount of information to process and sometimes that is just as well. I will have to think about this. Then nothing happens. And so maybe, there are situations where fewer data are actually better. Of course that’s aligned with – particularly when more data are actually poorly analyzed.

(Slide 17)
Finding the right fit is a, you know, a very simple graph. Getting that maximum point is not – there’s no magic point that I’m saying this is it, but you know, the point is there are diminishing returns after a certain point and there’s obviously increasing returns going from zero.

(Slide 18)
This is what leads us to our principles. Credible, actionable, responsible, transportable. By credible what we are talking about is collecting accurate, high quality data and analyzing them appropriately. This is really two parts.

There is the accurate high quality data. There are clearly important method issues about how you collect data, how you ask people if you’re doing a survey what their income is and consumption, and there’s good ways and bad ways of doing that. There is ways that can generate lots of buy in and it’s important to pay attention to these details. There is administrative data that could be useful but might not be so accurate. If it is tax data, for instance. A common issue with tax data is to try to understand what is happening in the informal economy. Maybe it is picking up just the formal sector and if there is lots of informal activity, it is inaccurate to what you want to know.

Analyzing them appropriately goes back to what I was referring to earlier. If you’re just following people before and after a program, and calling that your impact, then you’re analyzing them inappropriately because that does not tell you the impact of a program. That is just – that could be driven by macro-economic conditions. It could be caused by from just getting older and wiser. It could be caused by changes in the ages of their children and what implies for their economic activities. Lots of things that change over time. We want to know how things changed over time compared to how they would have otherwise. When we see impact evaluations that just say, well, 80 percent of our people are in a job that tells us absolutely nothing. I want to know what proportion would have had a job had this program not existed.

Actionable. What I am going to do is just talk through these slides and then we are going to post the full thing on, but I am not going to go through each of the actual slides associated with each of these. Actionable refers to having a plan that, depending on the outcome of the data
collection and the analysis, there’s different actions that one can take and one of the things that we often see is that data are collected and if someone sat down and said up front before the data are collected, okay. What are you going to do if you get these data and the answer is above 12? If the answer is nothing differently than at below 12, then you probably should not be collecting the data. It is not – what good is it really doing? So actionable is about making sure that there are different decisions that you’re going to make depending on what those data said, and that requires both believing the data, believing and having confidence in the credibility. It also requires having an information system so that you get that data in a timely fashion, that it doesn’t come to management three years after the program is finished and there’s nothing left to do and now it’s just a point of mental exercise.

The third point is responsible. Responsible is about thinking what the costs are of getting these data. Do the benefits of this data collection outweigh the costs? Data are not free, both in terms of the cost of collecting them, nor the cost of thinking about them. Think about the jam analogy. It costs management time to sift through these data and sift through the analysis of these data. So being responsible about the data collection is about recognizing that every dollar spent on data collection is a dollar not spent on a program. Is that the responsible choice?

Transportable is about asking oneself are we filling a knowledge gap that is useful, not just for us, for others. There are situations in which transportable might be argued as not useful. So a lot of monitoring exercises, for instance, by their very nature, are not very transportable, and that’s fine. So we think of transportable as a when you can think hard about the transportability and if there’s something you can do to improve the transportability, what it means is that your impact is not just going to be on your participants but it’s going to be on future participants and maybe people in other areas that are – and policymakers in other areas that are dealing with the same struggle and giving them more information on how they can improve what they’re doing, improve the impact of what you’re accomplished. That is part of what transportable is.

(Slides 19, 20 & 21)

Here’s a cute slide we have about credible. This is a good example of data analysis. So shaved heads goes up, sales goes up. We found this correlation in the data. Everyone take a razor. Start shaving your heads. Obviously this is – the point is this doesn’t actually cause sales to go up, it’s just that, you know, these are two independent trends that happened to go in the same direction.

(Slide 22)

This is a particularly important one when we are talking about whether data are credible. The impact – when we talk about impact, we are talking about causality. We are telling how an organization has changed the world around it, and that means answering this question in bold. What happened with the program, minus what would have happened without the program is the impact. One of the most common flaws we see are groups that say, well, we did a job-training program and 80 percent of the program in our program got jobs at the end. That is a really bad way of measuring that, because you don’t know what proportion would have had jobs. Now of
course you can filter in, at the beginning there is zero and then they go to 80, but the question is what would have happened?

Imagine two programs. One says 80 percent, and the other says 40 percent of our people got jobs. From that alone, would you want to make any claim as to which one is a better program? Absolutely not. What if the one that's doing 40 percent is finding harder to reach people? Maybe they're dealing with people who have been recently incarcerated and so of course it's going to be harder, or maybe it's not something that's so easily absorbed, but it's just an organization that is more dedicated to working with hard to work with households, and helping them get job training and get jobs. You cannot compare 80 percent to 40 percent and claim one is better than the other.

(Slide 23)

What you would need to do in each of those cases is set up a randomized method where they have some process of doing an intake that is larger than their capacity to actually deliver services to. That allows you to create a control group in which some people get the training and other people do not, or get it later. Now I understand what is happening in terms of my impact. A lot of times, as we know, people do cycle in and out of jobs. When you see someone who's unemployed get part of a job training program, just knowing that they got the job later doesn’t tell you the job training caused that to happen, or maybe just, you know, if they’d spent that time searching for a job they would have gotten one. Some proportion will and knowing what that proportion is, is critical. That is basically what a randomized trial is aiming to do. That is really critical to knowing which policies are actually working so that you don't make false decisions about which policies to scale and which policies not to.

(Slide 24)

As an example, we did have a comparison of a few different methods. This was impact of a remedial education program in India, and this program had a few different methods of measuring impact, and then a randomized experiment. What we found, for instance, was just following children before and after, what we called pre-, post, radically overestimated the impact of this program by fivefold. If we just compared those who participated and those who did not, it actually looked like a bad program, like lower, but that is wrong because what happened was it was choosing children who were doing poorly. So of course comparing those who were in the program to those without made it actually look like the program was bad. We actually do get a fairly close estimate when we look at what's called the Difference-in-Difference. The regression results actually make it go away, but the proper analysis for the randomized experiment lead to the most accurate estimate. In some cases, you will get -- you might get by loss a result that is close with some other method. One thing that is important to understand when we can do that and when we cannot because it does make things easier if you can. There are examples, for instance, with boating, where we have seen there is no method that got it right.
One thing to remember is a lot of times the randomized trials are referred to as the gold standard, and that’s not really the right way of saying it. The right way of saying it is that randomized trials may be the gold standard for measuring impact, but they are not the right fit for every organization. There are certain circumstances like I described, when you cannot do them, and it is very important to remember that, to not just think, oh, randomized trial. That is the best. Let’s go do that. But do it in settings where it can and should be done – which is a lot of places. Don’t get me wrong, but it’s not everywhere. So we don’t like – I don’t like using the phrase randomized trial is the gold standard because it makes it sound like in any setting if we randomize it, that’s going to be better. And there are situations where randomizing it would be a really stupid thing to do.

The two line summary of the talk is data, have a plan, make sure it’s credible, actionable, responsible, and ideally, but not always, transportable. And that’s part of what we mean by the second line, which is probably a good one.

So time and time again we’ve seen a lot of groups have a data plan but then it doesn’t adhere to these principles. That’s the punch line of this talk is to have a plan, preferably a good one. Now we can do Q&A.

Katharina Wolfe: That was absolutely fantastic and if I can give a virtual applause, I think everyone around the table – there are over 90 people on the line – giving you a thorough applause. Thank you so very much. Very insightful and informative. I think we have a few minutes for some Q&A.

Answer: (Answering question from online chat requesting the resources available.) Two answers. There’s a lot of resources on IPA’s website, as well as the Neo Poverty Action Line website. I would start with some of those. There’s actually a course done at MIT on randomized trials that’s available. We also conduct executive education programs. Those are posted on the Neo Poverty Action Line – Action Labs’ website. And last but not least is one thing that we’re happy to do that I know has been part of the conversation is – depending on what the interest is and the fit - we have done specialized shorter courses where we come in and can dig into details on a few things over a few days for groups if there’s enough people within a group. We’re happy to talk about that if that’s something of interest to people.

Question: We have a question here about strategies for collecting information to demonstrate impacts beyond randomized control trials. Are there other tactics that can be used?

Answer: The short answer is yes. They do often require more expertise and depending on the circumstance, we’ve seen many situations when alternative methods lead to evaluations that introduce too much bias where it might not be advisable. There are situations where it’s better off to say, you know what? We’re not going to have impact evaluation. We’re just going
to rely on monitoring data for accountability purposes. It’s hard to answer that question in the abstract, but I’ve personally been in many situations where sometimes if there is a non-randomized method that could be credible enough, depending, it requires more expertise than a randomized trial to do. There are strategies that use kind of ancillary econometrics but require more assumptions. So it requires more judgments and expertise to make those assumptions. Abstractly speaking, yes, but most people who’ve done randomized trials have been involved at some point in their life in a non-randomized study that had the right institutional path that allowed them to feel comfortable that they weren’t leading people down a bias path and giving people results that lead to bad decisions afterwards.

Question: Well, I'm going to take a stab at a question here, Dean. It's Katerina. We go out and talk to a lot of various organizations about the local poverty reduction fund, which is really about encouraging the use of evaluation for poverty reduction initiatives across the province. One thing that we heard a lot was around the delicacy of individuals who live in poverty and how do you develop a control group with a vulnerable population. Is there an experience on some other project you’ve worked at internationally that you could speak to about that matter?

Answer: There’s two basic strategies typically employed. The first thing to remember is that just about most all programs that I’ve experienced, there’s some capacity in the program. They are going to out and they’re going to deliver a program to 500 households and so it’s embracing, recognizing that capacity and then working the design around that. So there is a training program and 500 people are going to be viable for it. Then let the program roll out to 500 people. Nothing has changed on that and so there’s the same number – the randomization does not reduce the number of people that are allowed to participate in the program. All the randomization does is say we’re going to keep the application process open, so to speak, until 1000 applicants are made. You’re embracing that capacity of the policy itself and then randomizing within that. The second thing to note is that in a lot of cases, there isn’t a control group. It’s just two different programs. There’s two different alternatives for how to do something. And you’re saying, well, I don’t know whether to do A or B given our program and I want to learn what the impact is of this extra piece that’s a part of this program because it’s a really expensive piece, so is it worth it? That’s a situation where there’s no control. There’s 1000 people that are in a program and 500 are getting Program A and 500 are getting Program B and you’re learning which program is better. And you know, a third thing to remember is that there are programs that have no impact. I always find it ironic when groups are concerned about the control group rather than the treatment group because there’s a lot of programs out there that if it’s not going to work are really bad for the treatment group and the control group should be the happy one. So this goes right to the point of when there’s a knowledge gap, who are you – what are you – what’s the aspiration here? If the aspiration here is to maximize the benefits of people who receive services from your program, that should include people this year as well as next year and the year after. Putting blinders on and not figuring out what the impact is does real harm to potential people in the future who are going to be treated by an ineffective program. I think are quite ambiguous when there’s – when there’s genuinely a knowledge gap, which there should be playing on your – it’s, you know, it’s not obvious which way to go and who one should feel is being, you know, at risk or at the treatment of the control. And that just goes to the heart of when there’s a knowledge gap of saying, you know what. We have a responsibility to our future beneficiaries of this policy to find out if the program’s actually working
or not before we waste more people’s time and take their energy away from something more effective.

Question: Are there good methods for monitoring root causes when using activity research or participatory methods?

Answer: I think there’s – for instance, feedback is often solicited in a participatory sort of way. Focus groups where you’re engaging people and getting them to take a step back and [inaudible] to try and generate [inaudible]. Participatory methods encompasses a lot of different things. So that’s just one example.

Cindy Perry: My name’s Cindy Perry. I’m from the Ministry of Community and Social Services and just on behalf of my Ministry and the Poverty Reduction Strategy Office, I wanted to thank Professor Karlan for donating his time to share his expertise and enthusiasm about structuring and monitoring an evaluation system that’s a right fit. So thank you also to all of the participants in this Webinar for taking the time to learn more about this topic as we all work toward building systems that will support continuous learning and improvement, and please remember to check out the online toolkit that was prepared by Professor Karlan and his team. We look forward to reading applications to the Local Poverty Reduction Fund from many of you in the coming weeks.

Katharina Wolfe: So with that, Professor Karlan, thank you very much again, and I think we can close the Webinar.

Professor Karlan: Great. Thanks everybody.