October **2018**

A primer about monitoring and evaluation, for funders and implementers

*Caroline Fiennes**
Natalia Kiryttopoulou
Marc Maxmeister

**corresponding author: caroline.fiennes@giving-evidence.com*

keystone
accountability for social change

Giving Evidence
Enabling **giving** based on sound **evidence**

Contents

Introduction

This document was written for a client, a funder, who was relatively new to monitoring / evaluation / learning. We are publishing it because we feel, and hope, that the material is useful for a wider audience of funders and implementers.

It is designed to explain what monitoring is, and what evaluation is, and how they differ.

We structured our thinking into a **four-level framework**. This simply splits out the various questions about monitoring and evaluation (note that, as explained below, monitoring and evaluation are two completely different things, even though they are often conflated):

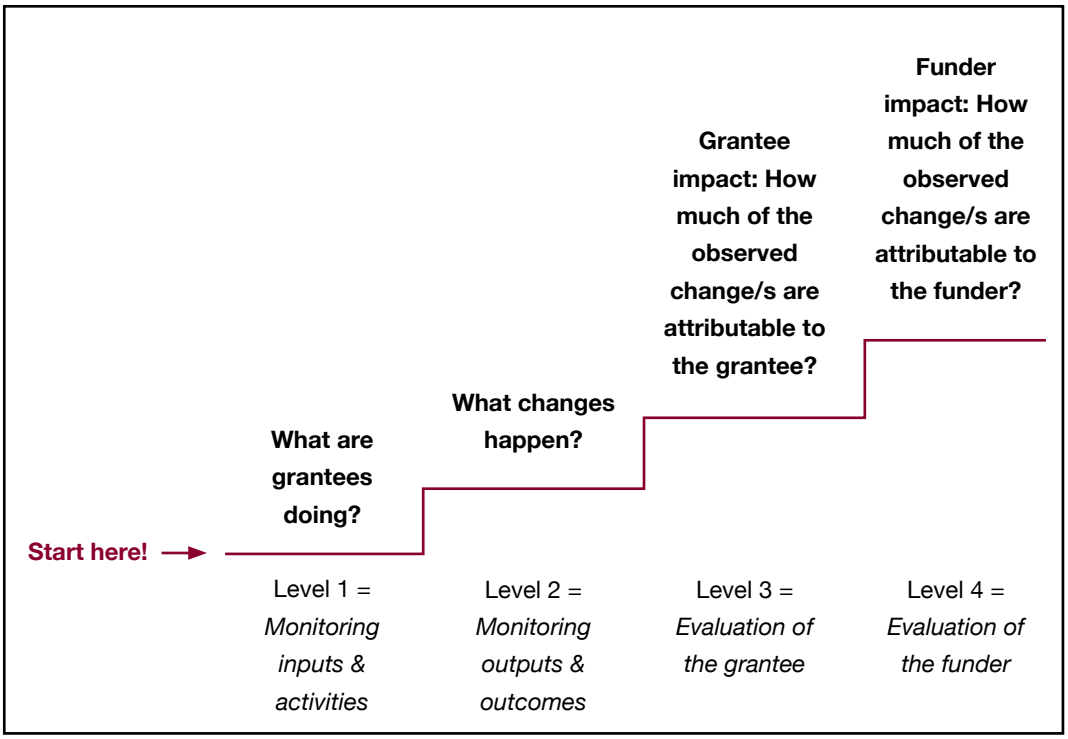
Level 1: dimensions of the grant; **inputs (such as grant size) and grantee activities**

Level 2: tracking **changes around the grantee**, e.g., increase in number of jobs, change in grantee partner revenue, number of workshops run

Level 3: **evaluating grantees**: i.e., establishing what of those changes result from (i.e., are attributable to) the grantee partner

Level 4: **evaluating a funder**: i.e., establishing what of those changes result from (i.e., are attributable) to the funder

We present these four levels as a ladder, because the issues at Level 1 are simpler than those at Level 2, and so on, both in terms of the types of data / analysis needed and the conceptual complexity.



To be clear:

1. An indicator is at Level 1 if you can could gather the data for it simply by knowing what is going on ***inside the grantee***, e.g., the number of loans that it made, its revenue, its management competence.
2. If the indicator requires knowing about changes ***around the grantee*** but outside the grantee - such as number of jobs (beyond the grantee's own payroll), number of homes built, laws which have changed - then it is Level 2.
3. Reliable work on Levels 3 and 4 unavoidably involves having a comparator.

Levels 1 and 2 are monitoring; Levels 3 and 4 are evaluation.

Many funders have information Levels 1 and 2, very little convincing at Level 3, and almost nothing at Level 4.

Distinguishing between monitoring vs. evaluation

The terms 'monitoring' and 'evaluation' are often used interchangeably, and / or run together as though they are one thing. They are in fact two completely different things. We use them to mean:

- **Monitoring** counts (or measures) the inputs into some work (such as cost, people involved, materials and equipment used), outputs (e.g., number of workshops run, number of leaflets distributed, number of children vaccinated), and / or outcomes (e.g., incidence of measles, number of people who vote, pollution levels).

- **Evaluation** is "a serious attempt to establish causation". In other words, evaluation aims to show whether /when the inputs cause the outcomes. Does increasing the amount of input increase the amount of outcome? Do the outcomes observed arise because of (are caused by) the inputs, or are they caused by something else?

This is a standard distinction. For example, to quote from the Hewlett Foundation's 'evaluation principles':

What Is Evaluation?
Evaluation is an independent, systematic investigation into how, why, and to what extent objectives or goals are achieved. It can help the Foundation answer key questions about grants, clusters of grants, components, initiatives, or strategy.

What Is Monitoring?
Grant or portfolio monitoring is a process of tracking milestones and progress against expectations, for purposes of compliance and adjustment. Evaluation will often draw on grant monitoring data but will typically include other methods and data sources to answer more strategic questions.

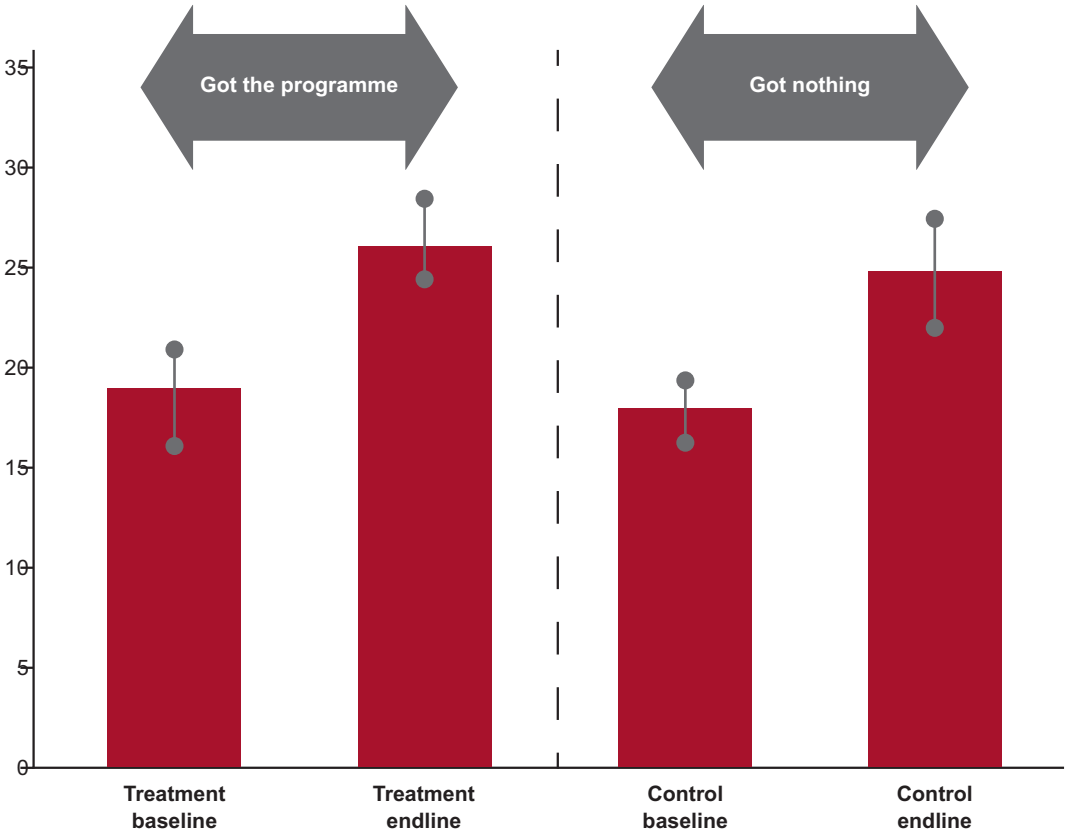
It is not enough just to understand what happened.
We also need to understand why it happened.

Monitoring is basically counting, whereas evaluation is science.

Good evaluation matters. A clear illustration comes from a remedial reading programme in South Africa. Monitoring data shows the children's reading ability beforehand versus afterwards. It increases during the programme. That might lead us to think that the programme works.

But careful examination with a randomized controlled trial shows that children who do not do the programme also make progress during it: in fact, they make precisely the same amount of progress. The evaluation ('serious attempt to establish causation') shows that the programme achieves nothing: the apparent improvement is simply due to the passage of time (see below).

Figure 1: Reading levels before and after reading programme[i]



There are other forms and purposes of evaluations, e.g., to understand how project partners or grantees feel about something.

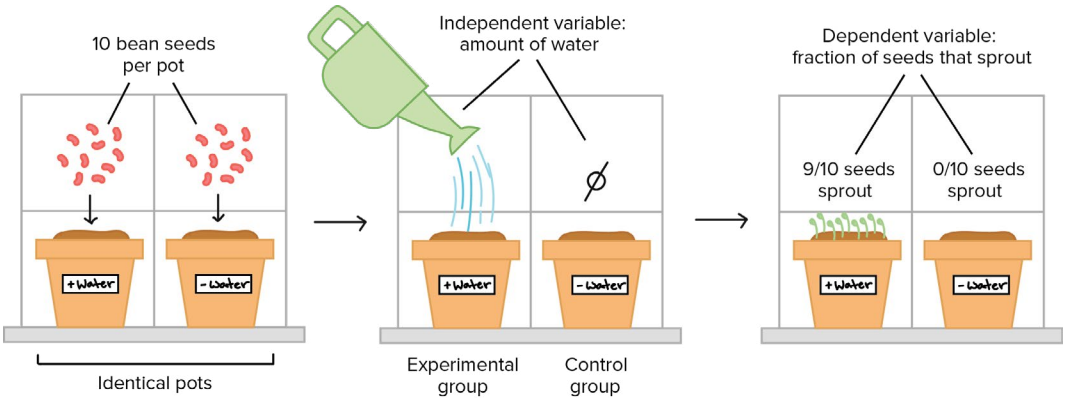
Impact evaluation requires a comparator

The impact (=effect) of a programme is the things that it changes in the world. That means, the things that change which would not have changed without it: in other words, the change beyond the change which would have happened otherwise.

An org / programme's impact
=
the change attributable to the org / programme
=
the difference between what happened with the org / programme and *what would have happened anyway*

It follows that a key question in evaluation is 'what would have happened otherwise?' You need a comparator group which don't get the programme. By analogy, to see the effect that watering has on seeds ('to understand causation'), we would need to have a group of seeds which do not get watered (their growth will be 'what would have happened otherwise'). We measure the progress of both using some 'ruler' (in this example, counting the number which sprout), and compare them:

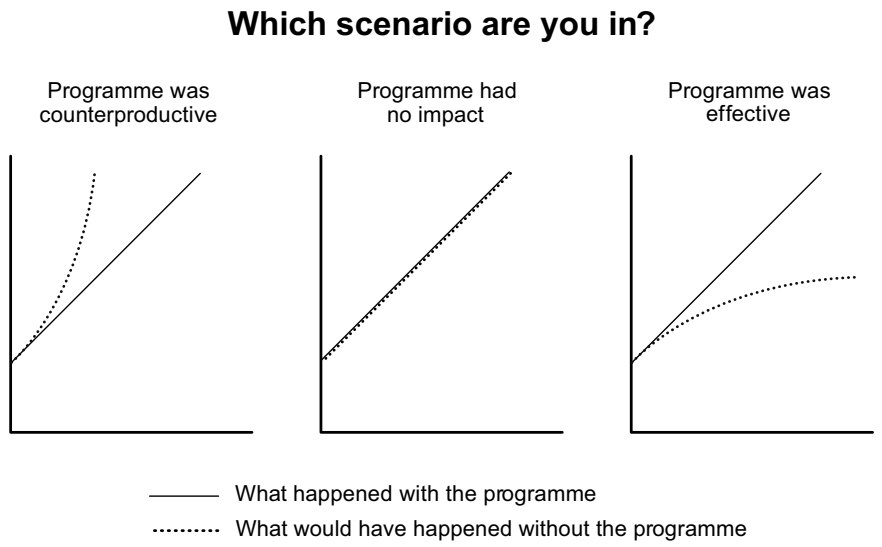
Figure 2: Illustration of importance of control groups in establishing causation



It follows that it is not possible to evaluate a treatment (here, watering) simply by collecting loads of data about the sample (here, seeds) which get the treatment: evaluation relies on having data about a sample which does not get the treatment.

It is quite possible that a programme (of which funding is an example) has a positive effect, or no effect, or has a negative effect. Without knowing what would have happened otherwise, there is no way to know which of these three scenarios you are in:

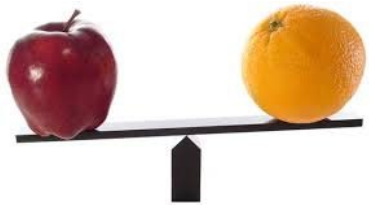
Figure 3: Illustration of importance of counterfactuals in establishing causation



We rehearse all of this here because it is very common for foundations to gather masses of data about organizations that they fund, but not have any data about comparator organizations that they do not fund. This prevents them seeing (evaluating) the effect of their funding.

We have seen many funders in that category. They have data about their grantees, but do not regularly collect data about non-grantee partners. The data about grantees do not show anything about the effect of the funding - because it provides no insight about 'what would have happened otherwise'. In the analogy, much monitoring is a way of counting sprouts; but cannot indicate what would have happened without that funder's watering.

The comparator needs to be a fair comparator



Suppose that in the experiment above, we put all the biggest seeds into the group which get the water. They then sprout better than the smaller seeds which do not get the water. We then cannot know whether the differential growth rate is due to (i) the seed size or (ii) to the watering. In other words, we would be unable to distinguish between:

i. a 'treatment effect', i.e., an effect of the treatment, or to

ii. a 'selection effect', i.e., some characteristic in the sample (here, size) which lead to them ending up in the treatment group.

In anything involving people choosing to apply to some programme or to ask for something, there is danger of a 'selection effect'. Below are some illustrations.

Fitted For Work (FFW) is an Australian charity which helps women into work. By way of demonstrating FFW's effectiveness, it reports that "75 percent of women who received wardrobe support and interview coaching from FFW find employment within three months… In comparison…about 48 percent of women who rely on Australian federal job agencies find work within a three-month period."

The comparison isn't valid, and doesn't demonstrate anything about FFW's effect. This is because women who get FFW's support differ from those who don't in (at least) two respects.

- First, they found out about FFW and chose to approach it for help. It's quite possible that the women who do this are better networked and motivated than those who don't. That would be a 'selection effect' in the women which FFW serves.

- Second, of course, the women who come to FFW get FFW's support. This is a 'treatment effect'.

The comparison doesn't show how much of the difference is due to the selection effect versus how much is due to the 'treatment effect' i.e., to FFW's support.

This isn't to say that FFW's programme doesn't work. Rather, it says that these data don't show whether it works or not.

This isn't just theory. Microloans to poor villages in Northeast Thailand appeared to be having a positive effect when analyzed using readily-available comparators. But these analyses didn't deal with selection bias in the people who took the loans. A careful study[ii] which did correct for selection bias and looked at how those people would have fared anyway found that loans had little impact. They had no effect at all on the amounts that households save, the time they spend working, or the amount they spend on education. It was only the most motivated people who sought and took the loans, and it turned out that they would out-perform their peers anyway. Here, the selection effect concealed that the treatment (the loans) in fact had no effect.

There are instances where a selection effect is so strong that it conceals that the treatment is in fact harmful. People die from this. Here's medical doctor and author Ben Goldacre on one example:

> *"We used to think that hormone-replacement therapy reduced the risk of heart attacks by around half, for example, because this was the finding of a small trial, and a large observational study. That research had limitations. The small trial looked only at "surrogate outcomes", blood markers that are associated with heart attack, rather than real-world attacks; the observational study was hampered by the fact that women who got prescriptions for HRT from their doctors were healthier to start with. But at the time, this research represented our best guess, and that's often all you have to work with.*
>
> *When a large randomized trial looking at the real-world outcome of heart attacks was conducted, it turned out that HRT increased the risk by 29%."*

Randomizing to get a fair comparator

Goldman Sachs has a programme called 10,000 Women which supports female entrepreneurs. Apparently '70% of [its] graduates surveyed have increased their revenues, and 50% have added new jobs.'[iii] Goldman Sachs obviously thinks this is impressive because it took out full-page adverts in magazines to announce this. Should we be impressed?

You can now see that this is pretty hopeless. For one thing, this isn't even before / after data: it's just 'after' data. It doesn't say that '30% of them were growing their revenues before, whereas 70% are now'.

For another, no control or comparator is given. We're not told that '70% of graduates increased their revenues, whereas only 20% of other businesses did in the same period.'

And there's a third major problem which is also pretty ubiquitous. Let's wonder for a second what those women would have achieved anyway. It's not hard to imagine that the kind of women who get themselves onto a Goldman Sachs programme are just the kind of go-getters who would do well in virtually any circumstance. That is, this programme may well attract and select people who are atypically entrepreneurial: the results that Goldman seems to be claiming may arise simply from **selection effect.**

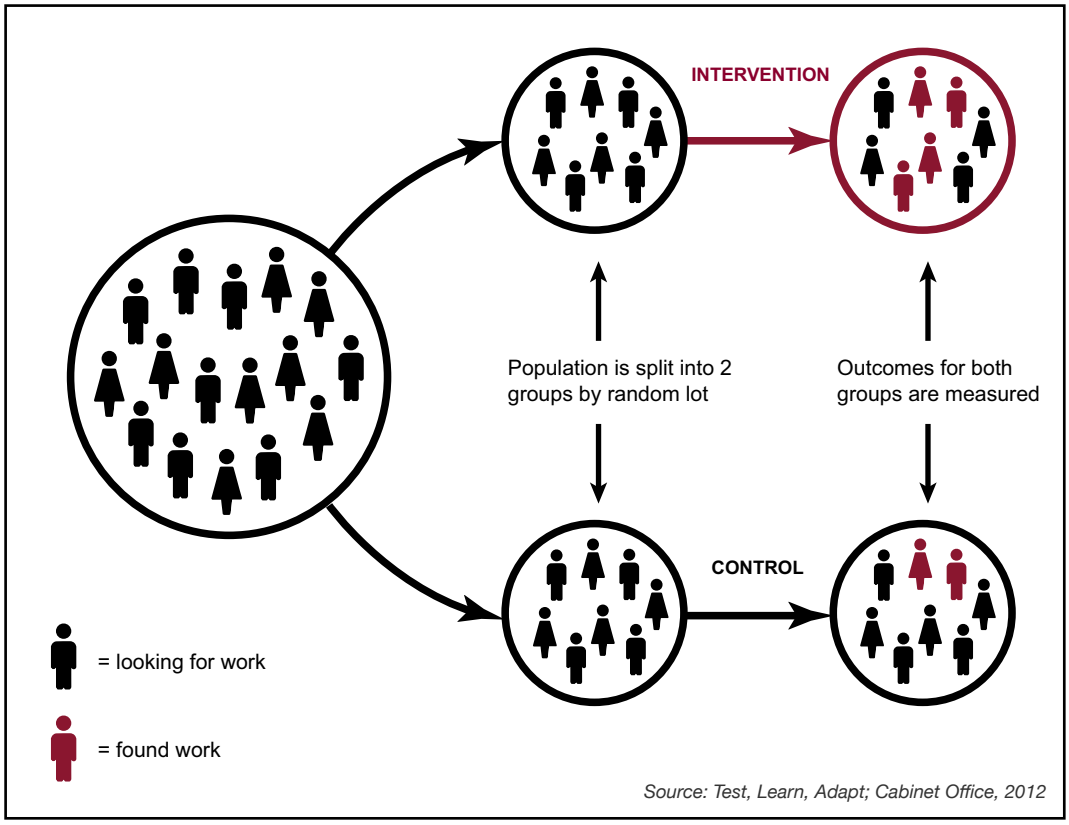
To 'establish causation', i.e., to see the effect of the programme, a researcher would need two sets of female entrepreneurs who are identical in every respect. She would put one set through the 10,000 Women programme and see how much better they do in their careers than a set which doesn't (this latter set is the control group, which will show what those women would have achieved anyway).

Irritatingly though, we can't create groups this way because people don't come in handy matching pairs: they have all manner of quirks and experiences and attitudes and individual traits which might affect their performance (introducing other possible causes). However, if the researcher takes a large enough group of women all of whom are eligible for the programme and divides them randomly, it's reasonable to expect those quirks and individualities to even out between the two groups. The randomness of the division removes the selection bias, leaving the programme itself as the sole difference between the groups. It isolates the effect of the programme and therefore comparing the groups' results will show the effect of the programme. Voilà.

The experiment we've just created is a **randomised controlled trial**. Executed properly, they do isolate the programme from all other possible causes and thereby show what would have happened without it[1]. This is why they're often called the 'gold standard' of evaluations (for single studies of impact), developed for pharmaceutical drugs trials and now increasingly used to provide robust and reliable insight elsewhere.

1 The trial described here would test the impact of the programme relative to doing nothing. In fact, we're generally not choosing between doing something and doing nothing, but rather seeing whether a new / / proposed programme is an improvement on what is already being done. So more useful is for one group to do the new / proposed programme and the other to do the best programme already available.

Figure 4: Example RCT (back-to-work programme)



For example, randomized controlled trials (by Innovations for Poverty Action and J-PAL) showed the usefulness of lentils in getting children immunized in India, and the relative cost-effectiveness of various programs to decontaminate water in Kenya. IPA and J-PAL also use them to test programs around improving sexual health, reducing corruption, and even post-conflict peace-building. They deploy proper scientific method – the great intellectual achievement of the modern era – to some of the most pressing social problems of our modern era. Randomized controlled trials (RCTs) have been used to study the effect of work to counter disadvantage among children in the USA, which is currently creating interest in the UK, and reducing child mortality in Uganda.[iv]

PICO: Population, intervention, comparison, outcome

All clinical trials can be read as investigating some PICO: the **o**utcome that some **i**ntervention(treatment) creates in some **p**opulation, compared with some **c**omparison treatment. In Ben Goldacre's example above:

- Intervention: hormone-replacement therapy
- Population: women*
- Comparison: no hormone-replacement therapy
- Outcome: heart attacks

P	I	C	O
Population Patient Problem	Intervention or Exposure	Comparison	Outcome
Who are the patients? What is the problem?	What do we do to them? What are they exposed to?	What do we compare the intervention with?	What happens? What is the outcome?

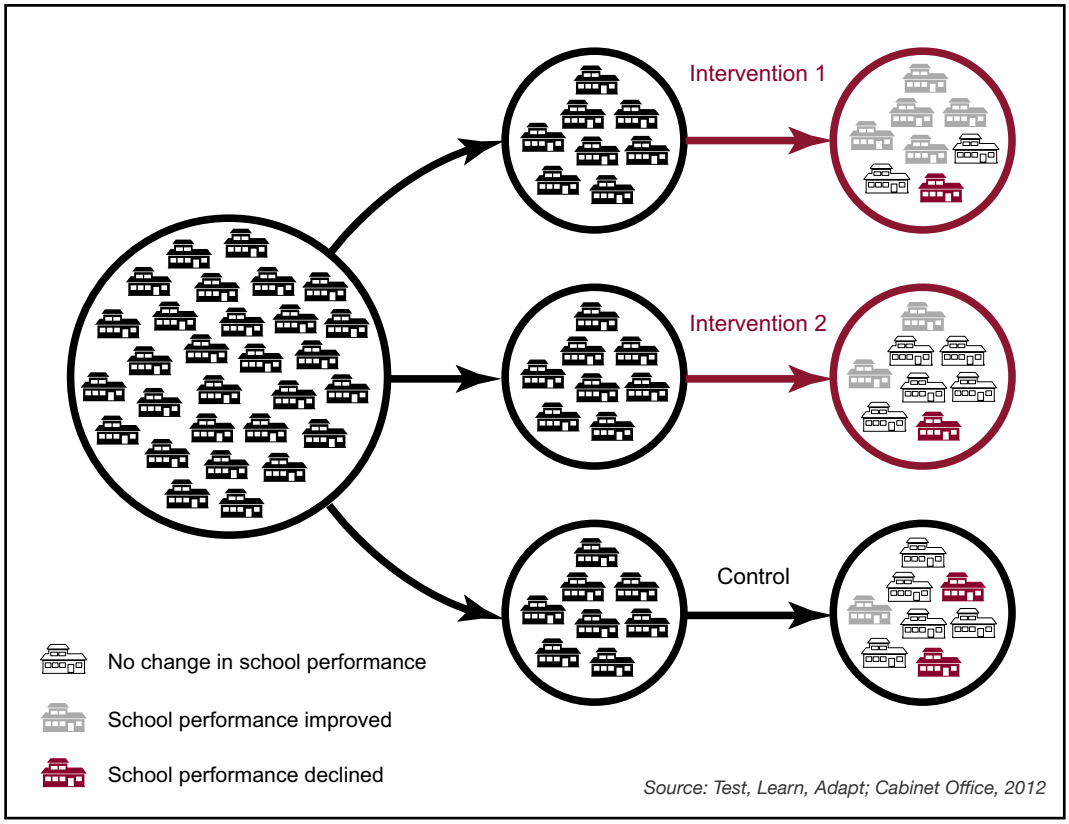
*The original studies were misleading - fatally so - because they failed to distinguish between pretty healthy women vs. less healthy ones, who are a meaningfully different population.

We recite this here because it underscores the importance of a comparator. PICO is taught to all medics. *This intervention in these people produces that outcome, compared to what?*

Which brings us to the point that **it is not the case - as sometimes claimed - that the comparator is invariably nothing (no treatment)**. Social programs, like doctors, are rarely choosing between providing something vs providing nothing at all. More often it is between providing Programme A vs Programme B, or full-service vs. partial service (e.g., funding a lot vs. funding a bit). In fact, the danger of using 'null comparisons' was explored in a medical journal editorial about medical trials which do that, with the unusually ardent title of 'Blood on Our Hands: See The Evil In Inappropriate Comparators'.

RCTs which compare programmes with each other work like this:

Figure 5: RCT of two educational interventions



All evaluations need a good ruler

All rigorous evaluations rely on data about performance, e.g., students' learning levels, the number of heart attacks, the school performance, the number of sprouts, how many women are employed, the business' revenue. Evaluators will examine those data for before the programme and after it (and possibly during), and for each of the various groups in the trial. They will measure performance using some kind of scale.

These measurement scales we can call 'rulers'.

A ruler is not an evaluation method. A ruler does not, of itself, 'establish causation'.

Good rulers (measures) are necessary but by no means sufficient for evaluation. Rulers (measurement scales) can be used for measuring the amount of inputs and / or the amount of outcomes. That is monitoring. As discussed, evaluation is different from this by also looking at comparator groups.

Many funders have rulers, but no evaluation tools (i.e., tools at Levels 1&2, but nothing at Levels 3&4). That means that those funders have no evaluation system.

Notice that rulers vary in how good they are. For example, some scales for measuring, say, a person's self-confidence are unreliable because the questions they ask are understood differently by different people: their 'inter-rater reliability' is low. (The analogy might be a stretchy ruler, which doesn't reliably indicate an object's length.) There are now many reliable, tested, stable 'rulers' (measurement scales) for a huge range of phenomena. Inventing one's own ruler is generally a bad idea.

Level 4: Funding agencies do basically two things

Funders have no arms and legs: the 'actual work' of changing things on the ground is done by the organizations that they fund. A funder's impact is vicarious, through its grantee partners.

Consequently, it is important to distinguish between the difference created by

- The grantee partner (This is what Level 3 does)
- The fund itself (This is what Level 4 does)

Funders sometimes claim that their impact is the total impact of all their grantee partners. However, this is patently not true because:

- At least some of those grantee partners would have achieved at least some things without that funder
- Some grantee partners may have achieved more without that funder, particularly if the funder is burdensome. (In other words, the grantee partner's achievements may be despite the funder. We certainly know of examples of this.)
- The funder may also have an impact on organizations that it does not fund, e.g., through its application process. Again, that can be positive (e.g., if the process helps applicants to clarify their goals) or negative (e.g., if it just creates deadweight administrative work with no benefit)

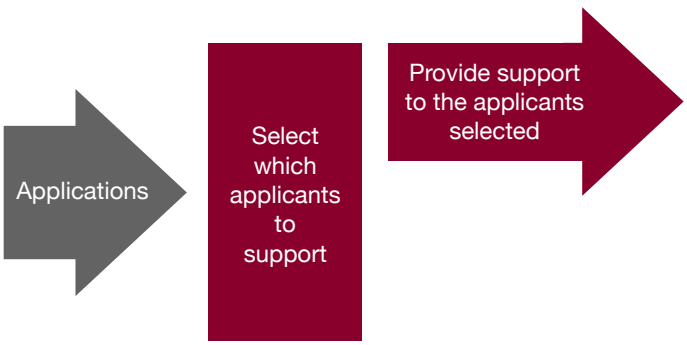
In terms of the impact of a funder - as distinct to it grantee partners - it is useful to distinguish between the two main things that a funder does:

(i) select which organizations to support, and

(ii) support them.

The first is a selection decision, the second is the funder's 'treatment'. They are quite different.

Figure 6: Grantmakers' two primary roles



It is possible to evaluate both a funder's selection process and (separately) its 'treatment'. Funders need to decide which (if either) it wants to understand.

In both cases, the rigorous way to evaluate them is to compare what happens with what would have happened otherwise.

Does your selection process add value? / Evaluating a selection process

Is your current process better than if selected between applicants at random? This is not flippant suggestion.

A large research funder, let's call them X, told us recently that most proposals that get through to its second round eventually get funded by somebody, even if rejected by X. It had examined the success of research proposals (in terms of citations etc.) that it funded, and compared it to the success of research proposals that it rejected. The answer was… nothing at all. That implies that X's selection process - which probably costs several million dollars a year - is no better than random.

The work of Nobel laureate Daniel Kahneman and his colleagues is awash with examples of human judgement being consistently worse than random. For instance, "Many individual investors lose consistently by trading, an achievement that a dart-throwing chimp could not match."[v]

The purpose of a selection process is (presumably) to choose the organizations likely to do best / likely to benefit most from the funder's support. In that case, the way to evaluate that selection process is to compare:

- the outcomes of organizations which that selection process rejects, with
- the outcomes of organizations which that selection process selects - but without giving them the 'treatment' of the funding and whatever else the funder provides {because providing the treatment would intermingle the selection effect and treatment effect, and so prevent identification of the value of the selection process.}

This is perfectly possible, though unusual. Notice that **evaluating a funder's selection process involves gathering performance data on organizations that it does not support.**

Obviously out-performing "a dart-throwing chimp" is a pretty low bar (which some clearly fail to clear). There is a next level of question around whether the selection process justifies the costs that it creates, those costs being to both the funder and all the applicants (successful and rejected). For that, you would need to ascertain the total cost of the application process, both internalized to the funder and externalized elsewhere.

Given all this, there is a growing evidence-base around the value of allocating grants at random.

Does your treatment (support) add value? / Evaluating a funder's support

The way to evaluate a funder's support is to compare:

- the outcomes of organizations that the funder selects but to which it provides no support, with
- the outcomes of organizations which that funder selects and to which it does provides support.

Again, this **involves gathering performance data on organizations which the funder does not support.**

A variant would be to compare:

- the outcomes of organizations which that funder selects and to which it provides a little support, with
- the outcomes of organizations which that funder selects and to which it provides rather more support.

This is a non-null trial like the education one shown in the diagram above. Importantly however, the funder would have to decide the level of support to each grantee partner at random: if it gave most support to those which seem to need it most, then it is again intermingling a selection effect ('you look ill') with a treatment effect ('here's lots of medicine'). Again, this is perfectly possible, though we do not know of any funder which yet does this.

Note that the debate around the ethics of trials like this is well-developed and essentially settled: if the answer to a research question (such as the value of a funder's support) is not rigorously known and the potential harms of the research are small, then it's fine.

Many funders could get a (non-randomized) approximation of this analysis by looking at the performance of grantees which get a lot of support with that of grantees which got less. A paper in the scientific journal *Nature*[vi] in 2017 by one of us (Caroline Fiennes) looks at how to assess a funder's work, as distinct to that of its grantee partners. GlobalGiving recently completed a three-year impact study which does this.[vii]

To conclude

There is a great deal more to say about how to do monitoring well and how to do evaluation well, in various circumstances, and how to deploy learnings from them. Much has been written about them.

We hope that this brief paper has clarified the difference between monitoring versus evaluation, and explained role and importance of each and why funders evaluating their own effectiveness is quite different from evaluating that of the organisations that they fund.

References

i Data refer to 100 schools in Pinetown District. 95% confidence intervals are indicated. Source: Fleisch, Taylor, Schoer, and Mabogoane, 2015
ii https://www.adb.org/sites/default/files/publication/28306/wp009.pdf
iii Goldman Sachs advert in Stanford Social Innovation Review, Summer 2011, back inside cover
iv Allen, G., Smith, I.D., 2008, 'Good Parents, Great Kids, Better Citizens', Centre for Social Justice, https://www.centreforsocialjustice.org.uk/library/early-intervention-good-parents-great-kids-better-citizens, accessed 12 October 2018
v *Thinking Fast & Slow*, Farrar, Straus and Giroux (2011)
vi Available at www.giving-evidence.com/nature. See Appendix, p20
vii https://www.globalgiving.org/learn/ggtestlab/globalgiving-impact-study/

Appendix

WORLD VIEW *A personal take on events*

We need a science of philanthropy

Billions of dollars are being donated without strong evidence about which ways of giving are effective, says **Caroline Fiennes**.

Philanthropists are flying blind because little is known about how to donate money well. Facebook founder Mark Zuckerberg's US$100-million gift to schools in Newark, New Jersey, reportedly achieved nothing. Some grants to academic scientists create so much administration that researchers are better off without them. And some funders' decisions appear to be no better than if awardees were chosen at random, with the funded work achieving no more than the rejected.

The recipients of funds are increasingly scrutinized, but the effectiveness of donors is not. Funders are rarely punished for under-performing and usually don't even know when they are: if the work that they fund helps one child but could have helped ten, that 'opportunity cost' is felt by the would-be beneficiaries, not by the funder. The same is probably true of agencies that fund research.

I founded an organization that promotes charitable giving based on sound evidence. I am acutely aware of how scant the evidence is about which ways of giving work best. The solution lies in more research on what makes for effective philanthropy. A 'science of philanthropy' could enable more to be achieved with the tens of billions given each year by foundations and other donors and funders.

Only a handful of studies have been done on donor effectiveness. The Center for Effective Philanthropy in Cambridge, Massachusetts, found that the time spent on proposals for, and the management of, ten grants of $10,000 takes nearly six times as long as the time spent on one grant of $100,000. The London-based consultancy nfpSynergy found that UK charities value £2 ($2.6) of unconditional funds as much as £3 of conditional funds, suggesting that attaching strings to donations reduces their value. And the Shell Foundation found that three times as many of its grants succeeded when the charity was heavily involved in creating and managing the work than when it had funded work based on a proposal from a non-profit.

Establishing the effectiveness of a donor is not straightforward. After all, donors have diverse goals, from funding basic research to testing interventions, providing services or promoting social policies. Nonetheless, answering three questions can provide useful insights for any donor. First, how many grants achieve their goals? (I call this the donor's hit rate). Second, what proportion of funds are devoted to activities such as preparing proposals or reports for the donor? Third, how satisfied are the recipients with the donor's process? Logging the goal of every grant and tracking whether these goals were met would be a big step forward.

Several fundamental questions about effective giving have yet to be studied. An obvious one is the role of grant size. Intuitively, larger grants should enable more impact and be proportionally less expensive to manage. But my organization's analysis of ten years of grants by ADM Capital Foundation in Hong Kong (published this month) found that grant size didn't seem to affect success. Similarly, a study of the impact of arthritis research found that large grants were no more consequential than small ones, possibly because smaller grants were awarded for different types of work. Another key issue is whether a broad or narrow scope makes funders more effective. The dominant theory in business is that specialization boosts success; nobody knows whether (or when) that is true in philanthropy.

Other unanswered questions concern the appropriate duration of grants, whether funders do better operating alone or in partnership with other funders, how involved donors should be in the work that they support and how donors should find recipients. Is it better to open applications to everyone, or to approach prospective grantees?

RECIPIENTS OF FUNDS ARE INCREASINGLY SCRUTINIZED, BUT THE EFFECTIVENESS OF DONORS IS NOT.

How to select recipients also needs study. Almost all funders make their decisions subjectively, either by soliciting the opinions of experts about a proposal or by interviewing applicants. Research on everything from picking stocks to student admissions shows that humans show weaknesses and biases in allocating scarce resources. The role of biases in awarding philanthropic funds has not been examined. One funder of academic research found that shortlisting applicants based on objective criteria was a better predictor of success (measured by scientific publications) than interviews were. Such findings are intriguing, but still too indiscriminate to yield broad implications.

When medicine became a science, health and longevity increased. Similarly, a science of philanthropy could reveal principles about which ways of giving are most successful. To move in this direction, every funder should gather data about its performance on the three metrics I outline, and share these data with researchers. Analyses should be done by researchers, not by the funders or by the recipients. The analyses could be retrospective, for example, by assessing how performance and recipient satisfaction have varied with grant duration or with how recipients were selected. Or it could be prospective, for instance, a funder could deliberately make some grants large and others small, and invite researchers to investigate how grant size affects hit rate and the cost of managing funds.

Such studies will of course require resources — from research councils or philanthropic funders. Although that might initially reduce the resources for the work being funded, it stands to improve the effectiveness of that work overall. More evidence about how to fund well could also increase the amount that donors are willing to give. ■

Caroline Fiennes *is the founder of Giving Evidence in London. She writes the How To Give It column in the* Financial Times.
e-mail: caroline.fiennes@giving-evidence.com

8 JUNE 2017 | VOL 546 | NATURE | 187