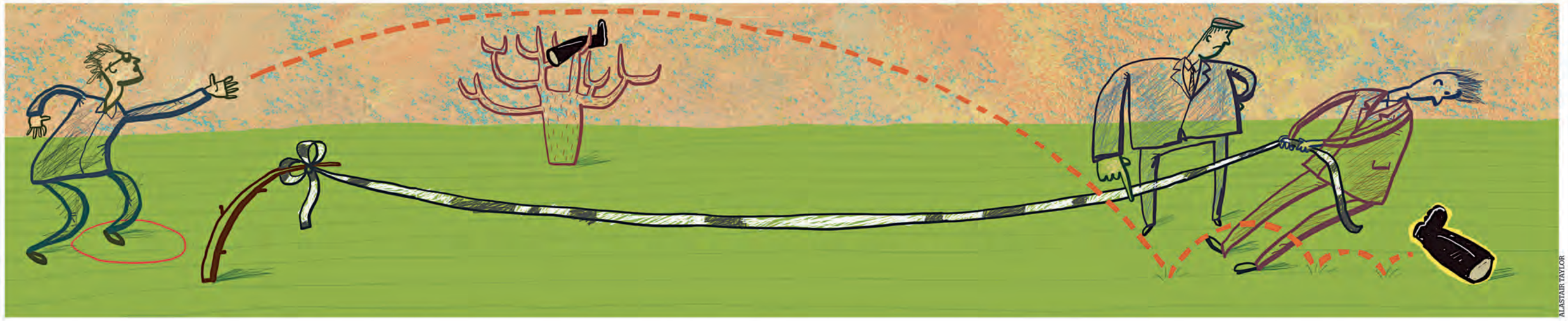
thirdsector.co.uk

ANALYSIS



ALASTAIR TAYLOR

We don't all need to throw wellingtons

Too much evaluation is a waste of time and money: charities should cite relevant research but not always produce it themselves. Caroline Fiennes explains why

Stand over there while I throw this wellington boot – I want you to see how well I throw it. Oops, that throw wasn't very good! Let's not count that one. Ah, the second was better. OK, now my assistant will measure how far it went. No – him, not you. It's actually quite hard to measure it properly – the tape has to be taut, so I have to secure it in the ground *here* – and I've not learnt to do that properly. Anyway, a bit of slack is all to the good! We'll use this tape-measure we made – it has a special unit of distance that we invented.

This, I suspect, is uncomfortably close to how monitoring and evaluation by charities work. Charities are judged on evaluations they produce themselves, for which they design measures, and they decide whether and what to publish. It appears not to help them much. If the aim is to improve the decisions made by charities, funders and policy-makers by providing reliable evidence about what's worth doing and prioritising, then much of it fails; it's just too ropey.

This needs to stop. It wastes time and money, and – possibly worse – it steers people towards bad decisions. My aim here is not merely to complain, but also to present some evidence about how monitoring and evaluation currently works, and to make some suggestions for a better set-up.

Why are we evaluating?

When asked in 2012 what prompted them to measure their impact, 52 per cent of UK charities and social enterprises talked about the requirements of funders, despite the fact that they existed for social purposes.

A study by two US universities indicates the incentives that influence charities' evaluations. The universities contacted 1,419 microfinance institutions offering to evaluate their work. Half of the invitations referenced a study by prominent researchers indicating that microcredit was effective. The other half referenced another study, by the same researchers and using a similar design, which indicated that microcredit had no effect. The letters suggesting that microfinance worked got twice as many positive responses.

Of course. The MFIs are selling: they are doing evaluations in order to bolster their case to donors, so it's little surprise when evaluations that don't flatter aren't published. I withheld unflattering research when I ran a charity. Withholding and publication bias are probably widespread in the voluntary sector, preventing evidence-based decisions from being made and wasting money.

Bad research methods

If charities want (or are forced by the incentives set up for them) to do evaluations that are flattering, they are likely to choose bad research methods. If you survey 50 random people, you'll probably hear representative views. But if you choose which 50 to ask, you could choose only the cheery people. Bad research is also cheaper – surveying five people is cheaper than surveying a more statistically significant 200. A charity head told me of a grant from a UK foundation "of which half was for evaluation. That was £5,000 – I told it that was unfair. We couldn't do decent research with that."

Charities' research is often of poor quality. The Paul Hamlyn Foundation graded the quality of research reports it received from grantees over several years as "good", "OK" or "not OK". The scale it used was more generous than that used by medical researchers. Even so, 70 per cent were "not OK". Another example is the Arts Alliance's library of evidence gathered by charities that use the arts in their criminal justice work. About two years ago, it had 86 studies. When the National Offender Management Service looked at that evidence for a review that had a minimum quality standard, how many studies could it use? Four.

One charity chief executive my organisation interviewed recently blurted it out: "When I first started in this sector, I kept talking about evaluation. But a senior person in the sector told me: 'Don't worry about that. You can just make it up. Everybody else does. At the very least you should exaggerate – you'll have to in order to get funded.'"

"Ask an important question and answer it reliably" is a central tenet of clinical research. But that isn't what happens in our sector. On reliability, much research by charities fails. This is inevitable, because investigating causal links is hard and most charities don't have the necessary skills. Given that 1,475 NGOs work in criminal justice in the UK alone, you wouldn't want them all to hire a researcher. And on importance, research by charities often seems to fall short – 65 per cent of chief executives of US foundations say that generating meaningful insights from evaluations is a challenge.

The collective spend on evaluation in the US is 2 per cent of total grant-making. That proportion of UK grants would be £92m. That would be enough for several pieces of reliable research, but split between multiple organisations in pieces of £5,000, it is unlikely to generate much that's robust. To enable evidence-based decisions, evaluations must enable comparisons to be made. So it's no good if everybody designs their own tape-measures – and yet a survey of 120 UK charities and social enterprises found more than 130 measurement tools in play. We need standardised metrics. This doesn't mean some impossible, universal measure of human happiness, but methods could be standardised within specialisms, such as particular types of mental health care, or job creation.

> **"I do not want my flight to be part of a trial to prove the aeroplane will stay up; I want to know that's been established already**

Cite research – don't necessarily produce it

When I get into an aeroplane, I do not want my flight to be part of a trial to prove that the plane will stay up; I want to know that's been established already. If an intervention is innovative – a new drug, for example – then obviously it won't yet have been evaluated fully, but it's reasonable to ask that the practitioner can cite some evidence that this intervention isn't bonkers: maybe it's a variation on a known drug, or other research suggests a plausible causal mechanism. We should do more of this in our sector. We should expect organisations to cite research that supports their theory of change; but we don't need every single organisation to produce research.

Imagine you're considering starting a breakfast club at a school. Should you do an evaluation? No. The first thing you do is look at the literature to see what's already known about whether it works. To be fair, the literature is currently disorganised, unclear and tough to navigate, but ideally you would look at research by other charities, academics and others.

continues overleaf

thirdsector.co.uk

ANALYSIS

continued from previous page

If that research is reliable and shows that the clubs don't work, then obviously you stop. If that research is reliable and shows that clubs do work, then crack on. The evaluation has been done already and you don't need to duplicate it: by analogy, we don't expect every hospital to be a test site. You can simply cite that evidence and monitor your results to check that they're in line with what the trials predict. If they're not, that suggests a problem in implementation, which a process evaluation can explore (see "What is evaluation...?", facing page).

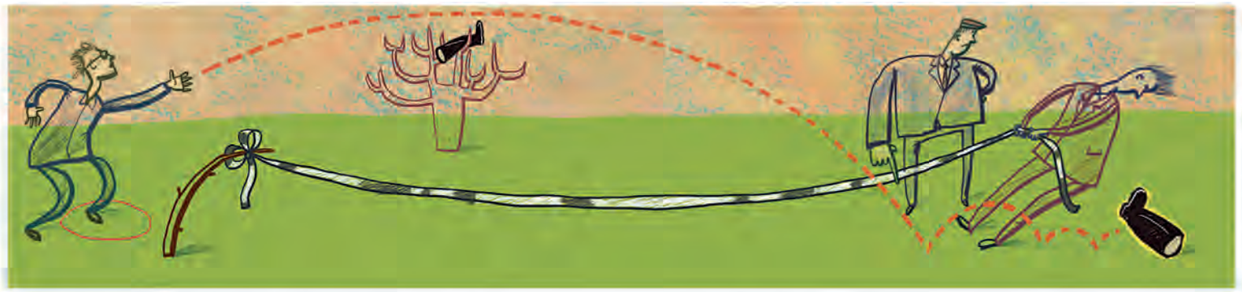
This, of course, is different from what happens now. In the model I'm suggesting, there will never be a rigorous evaluation of your breakfast club, just as most cancer patients will never be in a rigorous trial, and you never want to be in one by an airline. But you *will* have a sound basis for believing that your club improves learning outcomes, and you won't have spent any time or money on evaluation. Of course, this model requires funders, commissioners, trustees and others to sign up to tested methods and not produce their own research models. They would look for evidence before they fund, rather than looking only at the monitoring and evaluation that emerge afterwards. The Children's Investment Fund Foundation, for example, reviews the literature relevant to any application it considers. Under this model, many fewer evaluations happen. Those few can be better.

If your literature search finds no evidence because it's a novel idea, then look at relevant literature, run a pilot and, if it works, decide whether to do a proper evaluation. Questions will then arise as to who does that and who pays for it – and a few points become clear.

First, the evaluation shouldn't be funded by the charity. If it's for the public good, other people will also use it, so it's unfair to tax the first mover by making them fund it. In international development, many institutions want to use reliable evaluations, but few are willing to pay for them, so many of them are funded centrally through the International Initiative for Impact Evaluation, known as 3ie – essentially a pooled fund from the Bill & Melinda Gates Foundation, the Hewlett Foundation and the Department for International Development.

Second, the budget for the evaluation has nothing to do with the size of the grant. If the question is important, it should be answered reliably, even if that's expensive. If adequate budget isn't available, don't evaluate at all. A bad answer can be worse than no answer and is just wasteful.

Third, the evaluation shouldn't be conducted by the charity. The obvious answer is academics – but, sadly, their incentives aren't always aligned with ours; their funding and status rest on high-profile journal articles, so they might not be interested in the question, and their product can be impenetrably theoretical and might be paywalled. Several people have suggested that young researchers – PhD and post-doctorate students with suitable skills – might be the answer: some system to introduce them to charities whose work, genuinely, needs evaluating. Project Oracle is doing this with some charities in London.



> **"If there is no evidence, that does not prove a programme won't work – but it should put us on our guard**

Reliable evaluation is essential

Does all this preclude innovation? No. You can tell that it doesn't because the model in which charities cite research, but don't always produce research, is essentially what happens in medicine, where there's masses of innovation. In fact, reliable evaluation is essential to innovation because reliable evaluations show which innovations are worth keeping.

They also show what's likely to work. Few things are totally new: most build on something already known. Suppose that you have a new programme for countering workplace gender discrimination and it relies on magic fairies visiting people at night. Well, that's interesting, because there's no evidence of magic fairies in the whole history of time. Thus there's no evidence to support the notion that this programme will work.

By contrast, suppose that your programme assumes people will follow the crowd, shy away from complicated decisions and are weirdly interested in hanging on to things they already own. Those three traits of human behaviour are very well established – Daniel Kahneman was awarded a Nobel prize for proving the first of them, and substantial evidence for them all is in his book *Thinking, Fast and Slow*.

At the outset, you won't have any evaluations of your particular programme, but you can cite evidence that it's not bonkers. We're not talking here about proof, clearly, but rather about empirically driven reasons to believe. What gives you reason to think that it will work? What else is similar and works elsewhere? What assumptions does the programme make about human behaviour, or organisations or political systems, and what evidence supports those assumptions?

Hence, the "cite research, don't necessarily produce research" model reduces the risk of funding or implementing innovations that fail, and thereby of wasting time, money and opportunity. It allows us to stand on the findings of many generations and disciplines and, therefore, see better whether our innovation might work. We might call this "evidence-based innovation".

If there is no evidence, that doesn't prove the programme won't work – but it should put us on our guard. The Dutch have a great phrase: "comply or explain". If your innovative idea doesn't comply with the existing evidence, then you have more explaining to do than if it does.

For example, to improve exam results, various economists handed schoolchildren a $20 note at the exam hall door. It sounds crazy. The students were told that they would have to hand the money back if they didn't do well. Now, suddenly, it sounds sensible. This innovation is informed by Kahneman's finding that people will work hard to retain something they already own – harder than they would work to gain that thing in the first place.

Context is, of course, important. Perhaps the evidence came from a time or place that is materially different and hence does not apply – or, at least, requires a bit of translation to the here and now. Innovations might be evidence-informed, rather than proven. And once your new gender programme is running, we need to see whether it's really working – not just whether it looks as if it's working. For that, we need rigorous evaluations. ■ ***Caroline Fiennes is director of Giving Evidence and author of It Ain't What You Give***

What is evaluation, what isn't, and what to do when? The example of a school breakfast club

"Evaluation is distinguished from monitoring by a serious attempt to establish causation," says Michael Kell, chief economist at the National Audit Office. This means that evaluation is not needed all the time. For service delivery, the types of research that are useful at various stages of a programme's development are as follows, taking the example of a school breakfast club:

Stage of programme development	Purpose of the stage, and useful information to gather	Application to breakfast club
Pilot	● Establish if the programme is feasible, if there is demand, the resource requirements (time, people, cost), management challenges and costs. ● Type of research: monitoring.	How much cereal is needed? Do children and parents want it? How many staff and how much time are needed to wash up? How much does it all cost?
Test	● Now tha t the programme is stable and manageable, investigate whether the inputs cause the intended outcomes. ● Type of research: evaluation, ideally rigorous – for example, with an equivalent control group – and conducted and funded independently. Most programmes need several evaluations, in diverse circumstances.	(How) does a breakfast club improve learning outcomes?
Scale up/Delivering services	● Now the programme is known to be effective and can be scaled up. We don't need to evaluate it again, so we can just monitor it to ensure that it's working as expected. ● Type of research: monitoring.	● Are the changes in learning outcomes in line with results from the trials? If not, something may be awry in the implementation. ● Monitor views, beneficiary uptake, measurable results (such as test scores) and cost.

● Monitoring and evaluation of **research and development** work and of **advocacy** work rather differently.
● This table does not look at **process evaluation**, which is separate (and highly useful). That aims to understand whether the intervention was actually delivered as planned; variations in cost, quality, staffing and so on; and to identify operational improvements.