# Improving the efficiency of evidence-based interventions

## The strengths and limitations of randomised controlled trials

**Mark Tomlinson, Catherine L Ward and Marguerite Marlow***

markt@sun.ac.za

http://dx.doi.org/10.4314/sacq.v51i1.5

*Globally, randomised controlled trials (RCTs) are increasingly seen as the gold standard of programme evaluation, representing the best way to determine whether new interventions are effective – but they are not without limitations. In this article, we discuss the phases of scientific discovery and the research standards that are necessary before scaling up interventions. We also outline the core characteristics of RCTs, such as randomisation, efficacy and effectiveness, and discuss the benefits of using the RCT as the standard of intervention evaluation. We discuss how 'realist' evaluation contributes to what policymakers need to know in order to make a decision about an evaluation and alternatives to the RCT, such as stepped wedge, regression discontinuity, non-randomised cohort, and time series designs.*

Evidence-based medicine aims to make clinical practice more scientific and empirically grounded in order to achieve safer, more consistent and cost-effective care.[1] It helps ensure that interventions are backed by evidence of sufficient quality to justify investment in implementation and scale-up. Since its introduction in the 1970s, the term 'evidence-based intervention' has moved from being an intellectual curiosity to a central component in conversations about health or behavioural interventions. There have been substantial successes with evidence-based medicine and policy development, but they are not without critics.[2]

* Mark Tomlinson and Marguerite Marlow are with the Department of Psychology, Stellenbosch University. Catherine Ward is with the Department of Psychology and the Safety and Violence Initiative, University of Cape Town.

Globally, randomised controlled trials (RCTs) are increasingly seen as the gold standard of programme evaluation, representing the best way to determine whether new interventions are effective.[3, 4] Evidence-based medicine is built upon the foundation of the RCT. It is rare, particularly in clinical practice, for evidence other than that emanating from an RCT to be considered of sufficient evidentiary standard – despite the fact that a great deal of clinical practice remains based on professional experience and observation. Others argue that the 'hegemony' of the RCT marginalises intervention types that do not lend themselves to an RCT design.[5]

In this article, we discuss the phases of scientific discovery and the research standards that some argue are necessary before scaling up interventions. We also outline the core characteristics of RCTs,

such as randomisation, efficacy and effectiveness, and discuss the benefits of using the RCT as the standard of intervention evaluation. Finally, we will juxtapose this with a discussion of the limitations of RCT and how other methods can be used as a way of testing interventions.

## How and why is evidence built?

### Efficacy and effectiveness

If policymakers propose to invest in a violence prevention intervention (a parenting programme, a life skills curriculum, reducing access to alcohol)[6] then one of the central questions should be: does that intervention achieve the outcomes that are expected of it, so that it will be a worthwhile investment of taxpayers' money? The purpose of an efficacy trial is to answer precisely that question: did the intervention make a difference, and how sure can we be that it was the intervention (and not something else) that made the difference? This is a question of *internal validity* (see Table 1 for a summary of definitions). Internal validity is the extent to which bias and confounding variables that may unintentionally affect the results are kept to a minimum in the conduct of a trial. Efficacy trials emphasise internal validity, and answer the question: 'Does this intervention work under optimal conditions?'

Effectiveness trials, by contrast, answer a different question: 'Does this intervention work under "real world" conditions?'[7]

Efficacy and effectiveness exist on a continuum. Taking part in research often involves procedures and commitments that are different from routine practice. It may not be possible for an intervention delivered under carefully controlled research conditions to be replicated under routine conditions. This presents a challenge to evaluating the impact of large-scale public health programmes.[8] Limitations associated with how study participants are selected, participant characteristics and trial management may also affect the relevance and feasibility of interventions based on RCT research. For these reasons, there is debate about the use and relevance of RCTs, especially in non-medical fields.[9]

**Table 1:** Definitions

| | |
|---|---|
| **Control group** | The group of individuals who do not receive the treatment condition, against which the outcomes of the intervention can be compared. |
| **Effectiveness** | The extent to which a specific intervention, when used under ordinary circumstances, does what it is intended to do. |
| **Efficacy** | The extent to which an intervention produces a beneficial result under ideal conditions. |
| **External validity** | The extent to which the results can be generalised to populations beyond the trial. Are the results valid for populations in which the intervention was not originally tested? |
| **Internal validity** | This gives researchers the confidence to conclude that what they did in the study caused what they observed to happen, i.e., that the outcome is the result of the treatment. A research study with high internal validity lets you choose one explanation over another with a lot of confidence, because it avoids (many possible) confounds. |
| **Intervention group** | A group of participants allocated a particular treatment. |
| **Selection bias** | A systematic distortion of evidence that arises because people with certain important characteristics are disproportionately more likely to wind up in one condition. Although random assignment theoretically eliminates selection biases, a bias can still occur. Another common problem is bias in selection to the trial at all – not only to which arm of the trial. |

### Generalisability

Related to issues of efficacy and effectiveness, another important question is whether the intervention will work with a different group of people. If a parenting programme was tested in Soweto with Setswana speakers, will it also work with isiZulu speakers in Ixopo, and Afrikaans speakers in Eldorado Park? This question – one of *external validity*, or *generalisability* – is crucial if policymakers wish to roll the programme out widely (see Box 1). If it was established as effective in one place, will it remain effective when taken to other places?

Efficacy and effectiveness are linked to the concept of generalisability. When a trial is conducted in an ideal setting with all factors and variables being controlled (as far as is possible) by the researcher, it may lack a measure of generalisability. Characteristics of those enrolled in a study (e.g. sex, age, severity of the disease, racial groups) are primary factors in generalisability.[10] For example, a study of a counselling intervention targeted at women may not necessarily generalise to men or children.

Geographic settings (urban versus rural) and health care systems can also be significant factors,[11] particularly when something more complex than a drug (e.g. screening programmes, behavioural therapy) is being tested. Multiple factors determine the external validity (i.e. generalisability or applicability) of studies, including of RCTs: characteristics of those taking part in the programme and in the study, the problem under investigation, costs, compliance, co-morbidities and concomitant interventions. Also, certain aspects of study design – eligibility criteria, study duration, mode of intervention, outcomes, adverse events assessment, or type of statistical analysis – greatly influence the degree of generalisability.[12]

## Phases of scientific discovery

For scientific evidence to be useful to policymakers, they need to distinguish which research and types of evidence will be most useful to them, which means understanding how new interventions are developed and taken to scale. Thornicroft and colleagues[13] propose a five-phase schema to understand research terminology and the discovery, development, dissemination and implementation of new interventions.

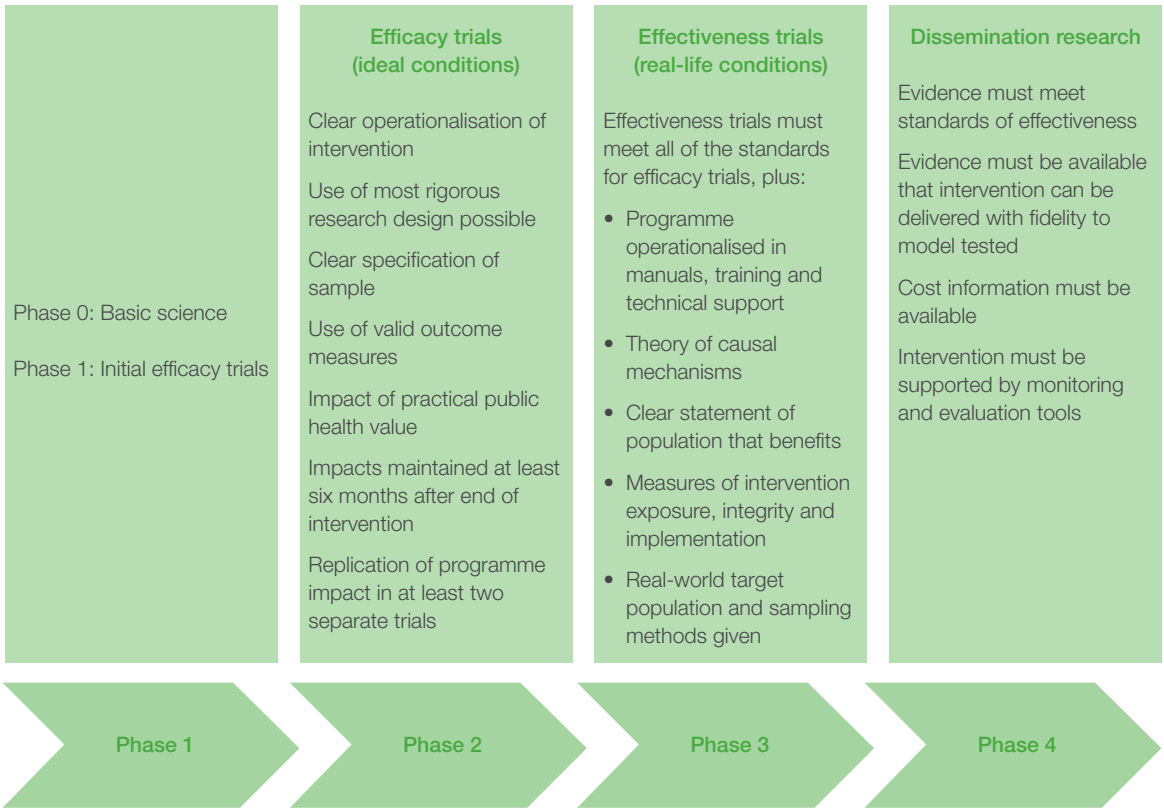
The starting point for any scientific discovery (**Phase 0**) is exploring relevant theories, generating hypotheses about how interventions might work, and conducting fundamental epidemiological research to understand factors driving the problem. These understandings can then be transferred to develop interventions. **Phase 1** includes early studies that aim to identify key components of an intervention. In **Phase 2**, investigators include efficacy studies (usually an RCT) that assess whether the intervention is effective under ideal conditions.[14] After efficacy of the intervention has been established, investigators shift the focus to studies in routine clinical conditions, to investigate intervention effectiveness in the real world (**Phase 3**). These studies may be implemented in target populations over a longer time period to identify other effects. Scaling up interventions that are scientifically proven and applicable to the everyday procedures of violence prevention practice can be challenging, and form **Phase 4**.

These five phases work together with standards set by the Institute of Medicine,[15] the Society for Prevention Research and other communities of researchers[16] to provide a framework for understanding what is good and sufficient evidence for establishing that an intervention should be implemented as a matter of policy. According to these standards, scale-up or countrywide implementation would be dependent on the completion (for each intervention) of (a) two high-quality efficacy RCTs, (b) two high-quality effectiveness RCTs, followed by (c) dissemination research that has established that the intervention can be delivered with fidelity to the model, and (d) information about the intervention's costs (see Figure 1 for a summary of these stages).

In addition, policymakers need to make decisions about how to weigh the evidence when considering implementation.[18] Victora and colleagues have proposed three levels of evidence to guide decisions:[19]

- Adequacy evidence – was the intervention implemented and found to be successful?
- Plausibility evidence – were the changes found in adequacy evidence shown to not be due to other influences?
- Probability evidence – were the changes observed not due to chance? For probability evidence, RCTs are needed.

**Figure 1:** Phases of scientific discovery and research standards




## Why randomised controlled trials, and where do they fit in?

RCTs are most successful in achieving high levels of internal validity and are thus considered the standard method for efficacy and effectiveness trials.[20] RCTs have a simple intention: to compare what would have happened in one group if the intervention was not received, with what happens when the intervention is received in another, otherwise equivalent, group. At the start (before the intervention is provided) those two groups must be equal in terms of their experience of the problem and characteristics that affect their experience. For instance, if the problem being addressed is child aggression, at the start both groups of children must be equal on a measure of child aggression, and have the same spread of age and gender of children (since older children and boys tend to be more aggressive, one must have equal numbers of older and younger children, and of boys and girls, in both groups). A defining characteristic of the RCT is that research participants who receive the intervention and the participants who make up the *control* group (i.e. those who do not receive the intervention) are *randomly assigned* to those groups (hence *randomised controlled* trial). With a sufficiently large sample, randomisation ensures fair distribution of the problem and related characteristics across the two groups. This capacity of RCTs to ensure a fair comparison between intervention and control groups is a particular strength, as it allows the most accurate possible estimate of what would have happened if the intervention group had not received the intervention.[21] Given an adequate sample size, the RCT typically surpasses all other designs in terms of its statistical power to detect the predicted effect of the intervention.[22]

However, randomisation may face opposition from policymakers and practitioners, who may believe in the value of an intervention for certain individuals or groups, often regardless of its actual evidence base, and therefore oppose random allocation.[23] For instance, in one trial – testing a substance abuse intervention in a community health centre, with the hope that it would reduce substance-related

aggression as well as substance misuse and HIV risk behaviours – nurses in the health centre tried to refer patients to the intervention group in the belief that the intervention would help them, regardless of the fact that the intervention had yet to be tested. However, only after the intervention has been tested in a high-quality evaluation can we have any certainty that it is effective. It is entirely possible that the intervention could have very little effect (as was in fact the case for that substance abuse intervention)[24] or even do harm. Famously, a substance abuse intervention that was rolled out widely in US high schools cost an enormous amount and made no difference to those receiving the programme: they were just as likely to use drugs and alcohol as those who did not.[25] Even more concerning, a common-sense delinquency prevention programme – taking youth at risk into prisons so that convicted offenders could scare them away from their lives of delinquency – turned out to increase offending in the young people, rather than deterring them.[26] In the long run, therefore, randomly assigning people to groups – knowing that people in need may end up in the control group and receive nothing – is more ethical than not using either random assignment or a control group,[27] providing of course that implementers truly do not know what the outcome of the intervention will be.

In the case of difficulties with, or objections to, individual randomisation, one possible solution is to use a cluster RCT, with the group (cluster) rather than the individual as the unit of randomisation. Members of a cluster (e.g., village, clinic, community) who might naturally influence one another or be affected as a group by prevailing conditions are clustered together and then randomised.[28]

RCTs are one of the most reliable methods of determining the effects of a treatment, because they are high in internal validity. However, they – like other trial designs that are used under very particular conditions – are not necessarily high in external validity. For instance, RCTs are often conducted with specific types of people under highly controlled conditions, and making inferences to the wider population may be difficult.[29] Recruitment often employs stringent eligibility criteria to minimise adverse events and potential non-responders. Some trials screen up to 68 people for each person enrolled.[30] In many settings, RCTs emphasise standardised interventions that might be too rigid when they need to be tailored for local population needs or other settings.[31] There are also concerns about the extent to which trials conducted in high-income settings apply to low- and middle-income countries (LMIC).[32] It cannot be assumed that there will be a universal response to an intervention across contexts, since a delivery system (such as a health system) in one context may have particular capacity for training, contact between health workers, supervision and population differences that will determine the effect of an intervention and to what extent it can be successfully implemented,[33] while delivery systems in other contexts may have different characteristics.

Other limitations of RCTs are that they are time- and energy-intensive as well as expensive, and may not be feasible for all interventions or settings. These threats to external validity limit the potential generalisation of the research results, an important consideration given the increasing emphasis on the translation of research results into practice.[34]

One common response to this is to try to have tests of programmes explicitly examine 'what works for whom, in what circumstances, in what respects and how', an approach called 'realist evaluation'.[35] This makes sure that the mechanisms that actually produced the change are clearly specified and consistent with the best available scientific theory and evidence, providing policymakers with the very detailed and practical understanding of a programme that is necessary before deciding whether that programme may be suitable for their context or not.[36]

**Case study: Box 1**

**Cognitive therapy-based intervention using community health workers (Pakistan)**

Rahman and colleagues implemented a cognitive behavioural intervention in which local health workers, known as Lady Health Workers, delivered a mental health intervention component.[37] One of the difficulties with implementing health interventions is the lack of adequately trained professionals in most

LMIC, especially in the case of mental health interventions where, in some countries, the treatment gap approaches 90%.[38] In Pakistan, Lady Health Workers are women who have completed secondary school and are trained to deliver preventive maternal, neonatal and child health care and education in the community. Lady Health Workers provide services to about 80% of the rural population of Pakistan. A cluster RCT was conducted with depressed women in their third trimester of pregnancy. Lady Health Workers were trained to deliver the intervention, while in control clusters Lady Health Workers who had not been trained in mental health made an equal number of visits to depressed women. The intervention halved the rate of prenatal depression in the intervention group. In addition, women receiving the intervention had better overall functioning and less disability up to a year later. Other health benefits included fewer episodes of diarrhoea and higher levels of immunisation in the intervention group. The intervention is a pivotal one because it is not dependent on a new or separate mental health workforce for its delivery. Rahman and colleagues argue that evidence of this sort is crucial in order to convince LMIC policymakers of the importance of integrating interventions such as these into the existing health system. This study is frequently used as evidence for how mental health interventions can be delivered by community health workers and how they can feasibly be delivered at scale – and this is undoubtedly true. There are a number of potential problems, however, with using evidence such as this in countries other than Pakistan. One is the lack of similar existing cadres of functioning community health workers such as the Lady Health Workers. Most LMIC do not have such an extensive workforce, and when they do there are significant problems with management, care delivery and supervision.[39] In addition, it is likely that the prevailing cultural and contextual conditions in this region of Pakistan (such as maternal seclusion after birth, and not being permitted visitors unless they are family) may limit the external validity of these data.

## Alternatives to the RCT

Aside from external validity, there are many other reasons why an RCT might not be the best method to assess intervention effectiveness. Reasons might include the following: when the impact is likely to be large, making randomisation potentially unethical; when the timing of the impact is likely to be long, making follow-up and assessment too expensive; or in a situation where a national roll-out of an intervention (such as in the Integrated Management of Childhood Illness [IMCI]) has already occurred, because a policy (or ideological) decision has been made about implementing a particular intervention.[40] In these cases, random allocation may not be possible. But there are alternatives, for instance:

- In consultation with policymakers, it might be possible to use a 'stepped-wedge' design, where implementing the intervention in certain areas is delayed – here the order of receiving the intervention is randomised.
- In some cases, there may be a clear cut-off that defines who gets the programme and who does not. For instance, the government may decide that only those whose household income is below a certain level will get the programme. Bonell and colleagues argue that in cases such as this a 'regression-discontinuity' analysis can be used, which examines the association between the outcome of the intervention and the measure of need.[41] Under certain conditions (such as a very large sample size), regression discontinuity designs can be just as powerful as RCTs. This approach was used to evaluate pre-kindergarten (the equivalent of Grade R) in Tulsa, Oklahoma.[42] All children had to attend pre-kindergarten, and so randomisation was impossible – but the regression discontinuity design used in the evaluation provided convincing evidence that the city's investment in pre-kindergarten led to worthwhile outcomes for children.[43]
- Another alternative design is what is known as non-random quantitative assignment of treatment.[44] In this design, participants are assigned to a treatment group based on need or merit, rather than random assignment. A good example of this is the school lunch programme in

the United States (US) where household income (below the poverty line) is used to assign children to receiving school lunches. Statistical analysis then models the functional relationship between the quantitative assignment variable (household income level) with the known outcome variables (such as health, concentration at school and academic achievement).[45]

- A similar design is a non-randomised cohort study where two groups are followed over time with baseline assessments, intervention is delivered to one group and not the other, and follow-up interviews are conducted to assess outcome. In this case two neighbourhoods can be chosen and matched as closely as possible. Without randomisation, ascribing change solely to the intervention is difficult, but if changes are in the hypothesised direction, policymakers might have sufficient evidence of effectiveness to implement.
- A final option is a repeated cross-sectional survey (or interrupted time series), which permits the evaluation of secular trends.[46] These are, however, expensive and prone to selection bias, although if routinely collected administrative data is of sufficient quality they can be very helpful and are relatively cheap, since they are gathered routinely and not just for the purposes of the evaluation. For instance, crime statistics collected by the US Federal Bureau of Investigation (FBI) were combined with data collected by the television industry to explore whether the introduction of television had increased violent crime in the US. A time series design was used to clearly demonstrate that violent crime had not increased, but that theft had increased as television was introduced.[47]

The point is that programmes that are to be rolled out widely (and where people cannot be randomised) must still be evaluated, using the best possible research design.

## Scale-up and 'when is there enough evidence'

Attempts have been made to rank the levels of evidence in order to assist policymakers in making decisions about evidence-based policy and practice. Within this framework the design and conduct of the research is categorised in terms of strength of evidence. In one of the most widely-used frameworks, there are five levels of evidence.[48] These range from level 5, the lowest level of support (expert opinion), to level 1, the highest – a meta-analysis of randomised controlled trials addressing the same problem, which can provide clarity on both whether the proposed intervention works and under what conditions. If many studies carried out in different settings together result in the conclusion that the programme generally has an effect, then one can have greater confidence that it will work in a new setting. Each of the other 'levels' of evidence (levels 2, 3, 4) of experimental design can then be seen as increasing the potential for the outcomes to be confounded by factors that are external to the experiment, or an inherent part of it, and are therefore weaker and less useful for making policy decisions.[49] Olds has argued that if policy and practice recommendations (in his case, for parenting interventions) are based on RCTs that meet the most stringent RCT requirements, they will have the greatest chance of being efficacious when disseminated and implemented at scale.[50]

Weaker evaluations mean that there is less chance that programmes will be effective when implemented widely and under real-world conditions. In addition, even implementing an established programme with a strong evidence base in a new setting runs the risk of changing some of the fundamental characteristics that led to programme success in the earlier settings (see Box 2). For this reason, every programme, no matter how strong its evidence base, should be evaluated when it is moved to a new setting.[51] For instance, when Strengthening Families (a substance misuse prevention programme shown to be effective in one setting) was implemented in a different setting (in the US – the same country in which it was originally tested) it was much appreciated by the families receiving it, but made no difference to their children's behaviour.[52] In cases where a programme is moved, but a full evaluation is not possible, some basic monitoring (for instance, comparing children's behaviour at the start and at the end of the programme) should be carried out.

**Case study: Box 2**

**Nutrition and psychosocial stimulation and mental development of stunted children (Jamaica)**

Grantham-McGregor and colleagues implemented an intervention study of nutritional supplementation and psychosocial stimulation of stunted children.[53] A total of 129 children were randomly assigned to four groups: nutritional supplementation only; psychosocial stimulation only; nutritional plus psychosocial stimulation; and a control group. There was also a group of matched non-stunted children. Community health aides delivered the intervention. The results of the study were compelling and showed how nutritional supplementation had a beneficial effect on stunted children's mental development. Importantly though, the treatment effects were additive, with the combined intervention (nutritional plus psychosocial stimulation) being significantly more effective than either of the stand-alone interventions.[54] This study is one of the most frequently cited papers in the child development literature and has had a significant impact on the design of interventions in many LMIC.[55] A recent 20-year follow-up on the same sample found that the earnings of the stimulation group were 25% higher than those of the control group and had caught up to the earnings of a non-stunted comparison group.[56] This study is unquestionably an important and seminal one. There are, however, two particular issues that should be borne in mind when using this data to inform scale-up or interventions in other countries. The first is the small sample size – only 32 children received the supplementation and psychosocial intervention. The second has to do with the relevance of this data (particularly the long-term economic finding) to most other LMIC. Jamaica has a very high rate of pre-school attendance, unlike most LMIC. The early impact of the supplementation and psychosocial stimulation is an important and compelling finding, but it is possible that part of the explanation for the long-term benefit of the early intervention is the additive booster benefit of a high enrolment in pre-school. It is possible that in countries where enrolment in crèches or pre-school is very low, the benefits of the early intervention may disappear over time. This is of course an empirical question and should be tested, but the issue is testament to the limitations of RCTs and how longitudinal assessment in many countries is vital in order to make meaningful policy decisions.

## Conclusions

Where does this leave policymakers? There are several principles to apply. Firstly, if a meta-analysis finds that a programme is effective, it is likely to be a good investment. At that point, experts should be commissioned to ensure cultural acceptability in the new setting, and to evaluate it – preferably using an RCT, to ensure good estimation of effect. Secondly, if there is no meta-analysis, one might commission experts to conduct one if enough RCTs testing the programme have been carried out. Thirdly, if a programme has shown promise in one RCT or in other forms of evaluation, conduct at least two RCTs before considering rolling the programme out. Programmes that are grounded in strong theory and have clear manuals to guide them are more likely to be effective than those that do not meet these criteria.[57] If programmes must be taken to scale immediately, there is no reason not to phase them in carefully in a cluster RCT. For instance, the government of the Democratic Republic of the Congo has invested in a programme aimed at improving children's numeracy, literacy and socio-emotional well-being. Schools were clustered together in clusters of three to six schools, with clusters randomly assigned, either to receive the new programme immediately or to be allocated to the control group, which will receive the programme at a later date. This allows for the programme to be tested in a thorough cluster RCT, at a level approaching scale, achieving two goals for policymakers: (1) making a potentially effective programme available to many children, while (2) ensuring that it is rigorously tested under real-world conditions, before scale is completely reached.[58]

In this article we have argued that policymakers should consider evaluation of programmes an

essential investment, as part of their responsibility to taxpayers, to ensure that public funds are wisely invested. We have discussed how RCTs are very powerful designs but may not always be possible, and have a number of limitations. Given this, we have suggested a number of alternative designs and approaches to evaluation that can help policymakers decide on which programmes might work best, and how to assess them in new settings. That policymakers should draw on the strongest possible evidence, and that programmes should be monitored and evaluated, are, however, beyond question.

To comment on this article visit
http://www.issafrica.org/sacq.php

## Notes

1 C Pope, Resisting evidence: the study of evidence-based medicine as a contemporary social movement, *Health,* 7:3, 2003, 267–282.

2 T Greenhalgh et al, Evidence based medicine: a movement in crisis?, *BMJ,* 348, 2014, g3725.

3 DT Campbell and JJ Russo, *Social experimentation,* Thousand Oaks: Sage, 1999.

4 WR Shadish, TD Cook and DT Campbell, *Experimental and quasi-experimental designs for generalized causal inference,* Boston: Houghton Mifflin, 2002.

5 CP Bonell et al, Alternatives to randomisation in the evaluation of public health interventions: design challenges and solutions, *J Epidemiol Community Health,* 65:7, 2011, 582–7.

6 See the World Health Organization's series of briefings on the evidence for effective violence prevention interventions: World Health Organization (WHO), *Violence prevention: the evidence,* Geneva: WHO, 2009, http://www.who.int/violence_injury_prevention/violence/4th_milestones_meeting/publications/en/.

7 M Godwin et al, Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity, *BMC Medical Research Methodology,* 3:1, 2003, 28.

8 CG Victora, J-P Habicht and J Bryce, Evidence-based public health: moving beyond randomized trials, *Journal Information,* 94:3, 2004.

9 L Rosen et al, In defense of the randomized controlled trial for health promotion research, *American Journal of Public Health,* 96:7, 2006, 1181; S Tilford, Evidence-based health promotion, *Health Education Research,* 15:6, 2000, 659–663; SG West et al, Alternatives to the randomized controlled trial, *American Journal of Public Health,* 98, 2008, 1359–1366; M Slade and S Priebe, Are randomised controlled trials the only gold that glitters?, *The British Journal of Psychiatry,* 179:4, 2001, 286–287.

10 G Gartlehner et al, *Criteria for distinguishing effectiveness from efficacy trials in systematic reviews,* Rockville: Agency for Healthcare Research and Quality, 2006.

11 PM Rothwell, External validity of randomised controlled trials: 'to whom do the results of this trial apply?', *The Lancet,* 365:9453, 2005, 82–93.

12 G Gartlehner et al, *Criteria for distinguishing effectiveness from efficacy trials in systematic reviews,* Rockville: Agency for Healthcare Research and Quality, 2006.

13 G Thornicroft, H Lempp and M Tansella, The place of implementation science in the translational medicine continuum, *Psychological Medicine,* 41:10, 2011, 2015–2021.

14 NC Campbell et al, Designing and evaluating complex interventions to improve health care, *BMJ,* 334:7591, 2007, 455–459.

15 PJ Mrazek and RJ Haggerty, *Reducing risks for mental disorders: frontiers for preventive intervention research*, Washington DC: Committee on Prevention of Mental Disorders, Institute of Medicine, 1994.

16 BR Flay et al, Standards of evidence: criteria for efficacy, effectiveness and dissemination, *Prev Sci,* 6:3, 2005, 151–75; G Thornicroft, H Lempp and M Tansella, The place of implementation science in the translational medicine continuum, *Psychol Med,* 41:10, 2011, 2015–21; DL Olds, L Sadler and H Kitzman, Programs for parents of infants and toddlers: recent evidence from randomized trials, *J Child Psychol Psychiatry,* 48:3–4, 2007, 355–91.

17 DL Olds, L Sadler and H Kitzman, Programs for parents of infants and toddlers: recent evidence from randomized trials, *J Child Psychol Psychiatry,* 48:3–4, 2007, 355–91; BR Flay et al, Standards of evidence: criteria for efficacy, effectiveness and dissemination, *Prev Sci,* 6:3, 2005, 151–75; M Tomlinson et al, Scaling up Health: where is the evidence?, *PLoS Med,* 10:2, 2013, e1001382; G Thornicroft, H Lempp and M Tansella, The place of implementation science in the translational medicine continuum, *Psychol Med,* 41:10, 2011, 2015–21.

18 DA Ross et al, The weight of evidence: a method for assessing the strength of evidence on the effectiveness of HIV prevention interventions among young people, in DA Ross, B Dick and J Ferguson (eds.), *Preventing HIV/AIDS in young people: a systematic review of the evidence from developing countries,* Geneva: WHO, 2006, 79–102.

19 CG Victora, JP Habicht and J Bryce, Evidence-based public health: moving beyond randomized trials, *Am J Public Health,* 2004, 400–5; JP Habicht, CG Victora and JP Vaughan, Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact, *Int J Epidemiol,* 28:1, 1999, 10–8.

20 CG Victora, J-P Habicht and J Bryce, Evidence-based public health: moving beyond randomized trials, *Journal Information,* 94:3, 2004.

21 CP Bonell et al, Alternatives to randomisation in the evaluation of public health interventions: design challenges and solutions, *J Epidemiol Community Health,* 65:7, 2011, 582–7.

22 Ibid.

23 Ibid.

24 JR Mertens et al, Effectiveness of nurse-practitioner-delivered brief motivational intervention for young adult alcohol and drug use in primary care in South Africa: a randomized clinical trial, *Alcohol and Alcoholism,* 49:4, 2014, 430–38.

25 S Birkeland, E Murphy-Graham and CH Weiss, Good reasons for ignoring good evaluation: the case of the drug abuse resistance education (D.A.R.E.) program, *Evaluation and Program Planning*, 28, 2005, 247–56.

26 A Petrosino, C Turpin-Petrosino and J Buehler, Scared straight and other juvenile awareness programs for preventing juvenile delinquency: a systematic review of the randomized experimental evidence, *Annals of the American Academy of Political and Social Science,* 589, 2003, 41–62.

27 L Rosen et al, In defense of the randomized controlled trial for health promotion research, *American Journal of Public Health,* 96:7, 2006, 1181.

28 Ibid.

29 Ibid.

30 CP Gross et al, Reporting the recruitment process in clinical trials: who are these patients and how did they get there?, *Annals of Internal Medicine,* 137:1, 2002, 10–16.

31 D Nutbeam, Evaluating health promotion – progress, problems and solutions, *Health Promotion International,* 13:1, 1998, 27–44.

32 PM Rothwell, External validity of randomised controlled trials: 'To whom do the results of this trial apply?', *The Lancet,* 365:9453, 2005, 82–93.

33 CG Victora, J-P Habicht and J Bryce, Evidence-based public health: moving beyond randomized trials, *Journal Information,* 94:3, 2004.

34 SG West et al, Alternatives to the randomized controlled trial, *American Journal of Public Health,* 98, 2008, 1359–1366.

35 R Pawson et al, Realist review – a new method of systematic review designed for complex policy interventions, *J Health Serv Res Policy,* 10:1, 2005, 21–34.

36 Ibid.

37 A Rahman et al, Cognitive behaviour therapy-based intervention by community health workers for mothers with depression and their infants in rural Pakistan: a cluster-randomised controlled trial, *The Lancet,* 372:9642, 2008, 902–9.

38 C Lund et al, PRIME: a programme to reduce the treatment gap for mental disorders in five low- and middle-income countries, *PLoS Med,* 9:12, 2012, e1001359.

39 K Daniels et al, Research supervision of community peer counsellors for infant feeding in South Africa: an exploratory qualitative study, *Human Resources for Health,* 8:6, 2010.

40 CP Bonell et al, Alternatives to randomisation in the evaluation of public health interventions: design challenges and solutions, *J Epidemiol Community Health,* 65:7, 2011, 582–7.

41 Ibid.

42 WT Gormley Jr et al, The effects of universal pre-K on cognitive development, *Dev Psychol,* 41:6, 2005, 872–84.

43 Ibid.

44 SG West et al, Alternatives to the randomized controlled trial, *Am J Public Health,* 98:8, 2008, 1359–66.

45 Ibid.

46 CP Bonell et al, Alternatives to randomisation in the evaluation of public health interventions: design challenges and solutions, *J Epidemiol Community Health,* 65:7, 2011, 582–7.

47 KM Hennigan et al, Impact of the introduction of television on crime in the United States: empirical findings and theoretical implications, *Journal of Personality and Social Psychology,* 42:3, 1982, 461–477.

48 PB Burns, RJ Rohrich and KC Chung, The levels of evidence and their role in evidence-based medicine, *Plastic and Reconstructive Surgery,* 128:1, 2011, 305.

49 Oxford Centre for Evidence-Based Medicine (OCEBM) Levels of Evidence Working Group, The Oxford 2011 levels of evidence, OCEBM, 2011, http://www.cebm.net/index.aspx?o=5653.

50 DL Olds, L Sadler and H Kitzman, Programs for parents of infants and toddlers: recent evidence from randomized trials, *J Child Psychol Psychiatry,* 48:3–4, 2007, 355–91.

51 FG Castro, M Barrera and LKH Steiker, Issues and challenges in the design of culturally adapted evidence-based interventions, *Annual Review of Clinical Psychology,* 6, 2010, 213–39.

52 D Gottfredson et al, The Strengthening Washington DC Families Project: a randomized effectiveness trial of family-based prevention, *Prevention Science,* 7:1, 2006, 57–74.

53 SM Grantham-McGregor et al, Nutritional supplementation, psychosocial stimulation, and mental development of stunted children: the Jamaican study, *Lancet,* 338:8758, 1991, 1–5.

54 Ibid.

55 S Grantham-McGregor et al, Developmental potential in the first 5 years for children in developing countries, *The Lancet,* 369:9555, 2007, 60–70.

56 P Gertler et al, Labor market returns to an early childhood stimulation intervention in Jamaica, *Science,* 344:6187, 2014, 998–1001.

57 BR Flay et al, Standards of evidence: criteria for efficacy, effectiveness and dissemination, *Prev Sci,* 6:3, 2005, 151–75.

58 JL Aber et al, Cluster randomized trial of a large-scale education initiative in the Democratic Republic of the Congo: baseline findings and lessons, paper presented at Society for Research on Educational Effectiveness Spring 2012 Conference, Washington DC, 9 March 2012.