How do we find out what works?

Mike Hough questions the thinking behind current evaluation of crime reduction schemes and calls for a broader, more inclusive approach.

I was educated in the traditions of psychological empiricism, and used to think that randomised controlled trials (RCTs) were the 'gold standard' of evaluation. I no longer think this to be true in relation to criminal policy, and I shall try to explain why. The catalyst to my change of heart has been my experience in evaluating several of the projects in the Home Office's Crime Reduction Programme. The relatively low return on the enormous investment in evaluating this programme has persuaded Home Office researcher managers to place a higher premium on RCTs and other robust evaluation techniques. I have been led to different conclusions.

RCTs – where people or places are randomly allocated to experimental or control groups – are undoubtedly a powerful way of identifying cause and effect. Provided that an RCT is done properly, one can be confident to a measurable degree that any differences observed between experimental and control group can be attributed to the experimental intervention. This is because the allocation process ensures, within limits, that the two groups are genuinely comparable on all factors that are likely to affect outcomes – except exposure to the experimental treatment. I genuinely take some comfort from the fact that the medicines that my doctor prescribes will have been tested through RCTs (though I have some residual concerns about small samples and the sources of funding of such research.) Some social interventions have also been successfully evaluated through RCTs – and indeed this issue of CJM reports the positive results of RCTs assessing early intervention (see Olds p.4).

However, RCTs are very hard to mount properly in criminal justice settings. The problems are partly ethical – in that judges, for example, may properly take exception to RCTs usurping their role as sentencer. Whilst there is no ethical problem in randomly allocating an additional *benefit* to an experimental group, there are serious (but sometimes surmountable) problems in randomly allocating *disbenefits* such as punishments.

There are equally pressing practical problems, in that it is hard to persuade staff in the experimental condition to implement the experiment as planned, and just as hard to persuade those in the control condition not to vary their practice. (The most consistent finding from the Crime Reduction Programme was that agencies failed to implement properly the projects to which they had committed themselves.) Just as important, it has typically proved very hard to assemble large enough samples for experimental and control groups. Rarely have criminal justice RCTs reached the scale of those on early intervention reported in this issue.

This means that those RCTs which *have* been carried out within criminal justice settings often have limited 'statistical power'. Most research literate people are all familiar with the concept of statistical significance. Tests of statistical significance allow us to quantify our confidence that a difference between groups hasn't arisen simply by chance. The research and policy worlds are still much less comfortable with the concept of statistical power – that is, having a research design that is sensitive enough to identify differences between groups as genuine experimental effects. Those RCTs that have been carried out within criminal justice tend – precisely because of the difficulties in assembling viable samples – to lack statistical power. This means that they have an inbuilt tendency to favour the 'null hypothesis' and to conclude wrongly that there is no experimental effect.

There is no reason to think that schemes shown to be effective in one town at one time will prove effective in other cultural and social contexts at other times.

These are all points that have been well rehearsed in methodological debate. But I am increasingly persuaded that the real limiting factor on the value of RCTs has been the conceptual crudity in the way that experimental evaluation has treated crime reduction. The social complexity of crime reduction work has been ignored.

This can be seen most clearly in evaluations of programmes for offenders. Over the last two decades, 'What works' evaluations have lost sight of the fact that work with offenders is a human art as well as – or as much as – a technology. The 1990s saw a 'programme fetishism' reflected in a preoccupation with identifying the activities that had most impact on offending – as if the training of offenders in anger management, for example, or in literacy skills or in relapse prevention, was the critical feature of

successful practice. Remarkably little research effort was put to questions about 'who works, in what settings?' and what it was about the most effective practitioners that made them more effective than their colleagues. The 'Who works?' question raises many issues beyond the obvious one about the personal – and immutable – qualities of individuals. Just as there are transmissible skills in staff management, there is a craft in working with offenders – even if the components of this craft have received little policy attention, and are rarely articulated in such terms.

This is not to say that the programmes developed in prison and probation settings in the 1990s were valueless or irrelevant. It probably makes sense to think of them as providing an important vehicle through which the processes of moral persuasion can be undertaken. What it implies, however, is a programme that is right for one institution at one time with one group of staff and offenders may not 'work' across all times and settings. On the contrary, the shelf-life of evaluative findings may actually be quite short. My own view is that cognitive behavioural programmes that have been proven to 'work' will over time lose their cultural resonance and with it their effectiveness.

It might be argued that crime reduction efforts aimed at 'people changing' are at the complex end of a long continuum of complexity, making them harder than average to evaluate. There is something to this, of course. However, even apparently straightforward crime prevention techniques such as CCTV surveillance or property marking or alley-gating achieve their impact in complex ways that are socially mediated. CCTV schemes, for example, are likely to have differential impact according to what people know and believe about them, how they are used, whether or not they command public support and so on. There is no reason to think that schemes shown to be effective in one town at one time will prove effective in other cultural and social contexts at other times.

What are the implications for those who fund or commission policy research into crime reduction? They should aim to lower some of their ambitions for research, and raise others. They need to temper the pursuit of methodological rigour. There is no point in spending large amounts of time and money trying to find out what does and doesn't 'work' when the chances of a clear answer are low, and where even clear answers are heavily context-dependent. In the jargon of research methodology, there is no point in over-investing in the pursuit of internal validity if external validity is likely to be limited. Comparison must remain central to the evaluation enterprise, however, but the idea should be ditched that anything short of an RCT is second best. My own view is that evaluations would do better to compare experimental outcomes with outcomes in a *wide range of broadly comparable* settings, than to rely on a *single comparison* with a closely matched control group.

Where should funders raise their game? Much more systematic thought needs to be given to the building and testing of middle-range theories about securing compliance with the law. The preoccupation within the Home Office and Ministry of Justice about testing and accrediting 'tools' to go into the crime control 'tool-box' has distracted attention from strategic questions about the best ways of getting people to behave well, and over-focused thought on less important tactical issues. Empirical criminal policy research has tended to shun the big issues about how best to secure normative compliance with the law from those who live at the social and economic margins of society.

We have enjoyed a decade of falling crime. My guess is that the trend will not continue. Globalisation and increasing competition, for example from the South Asian and Chinese economies, will drive down the wages of the less affluent sectors of industrialised countries. Our economy could be destabilised by climate change or by viral pandemics. The precise impact of these threats is unknown, but they are all likely to increase income disparities, and thus to intensify problems of crime and order maintenance.

There are two broad policy responses to the crime problems created by socially marginalised groups: deploying strategies to secure instrumental compliance (or strategies of repression, as our European neighbours would say) or inclusionary strategies designed to increase commitment to the law and to mitigate the impact of social and economic inequalities. My concern about current trends in government funding of research in this field is that as a general rule, inclusionary strategies are less amenable to tight experimental evaluation than those that focus on instrumental compliance. We badly need more policy research which sheds light on normative compliance and on effective ways of bolstering the legitimacy of the criminal justice system in the eyes of the socially marginal. My fear is that this much needed programme of research will be displaced by methodologically rigorous – but overly narrow – evaluations of strategies designed to fine-tune repressive strategies.

■

Professor Mike Hough *is Director of the Institute for Criminal Policy Research, King's College London.*