

Asking the right questions in criminal justice evaluations

Nick Tilley describes the limitations of experimental methods.

There is something seductive about the idea of well-designed 'experimental' evaluations in criminal justice. They draw their authority from medicine. They offer an unbiased means of estimating what would have happened anyway (the 'counterfactual'). They enable effect size measurements, which can inform cost-benefit analysis. Randomly selected experimental and control groups from different studies can supposedly be aggregated to find statistically significant small effects, which may be absent in constituent small sample studies. Furthermore, experimental studies employ pleasingly simple, elegant, and common-sense methods, which all can understand and recognise. They are widely deemed to provide the 'gold standard' against which all others should be judged.

Ray Pawson and I have raised some fundamental objections to the experimental orthodoxy in the course of our discussions of 'realistic evaluation' (see Pawson and Tilley, 1997). Committed experimentalists have rejected our arguments. Many sympathisers with realistic evaluation believe there continues to be a major role for experimentalism. At a distance of fifteen years since we first floated the arguments, however, I find myself largely unrepentant. I am also distinctly unimpressed with further developments in experimentalism in the interim, many under the auspices of the Campbell Collaboration.

Let me rehearse and reinforce the arguments against experimentalism, using a recent Campbell Collaboration example. Petrosino *et al* (2002) looked at programmes intended to deter potential offenders from crime through prison visits. The most notable and durable of these is 'Scared Straight'.

Evaluations of prison visit programmes avoid many of the specific practical difficulties that face area-based evaluations. The interventions are applied to discrete individuals, who can be assigned to different treatments on a randomised basis. Blinding subjects and researchers to interventions is more straightforward. If randomised control trials produce valid and useful findings at all, they should do so in this case. Let us see how they do.

Petrosino *et al* began by surveying the literature to find studies that used random or quasi-random procedures (alternate assignment) to allocate delinquents or 'pre-delinquents' (including at least some aged 17 or less), to treatment or no-treatment groups, with or without blinding, and with some measurement of before and after offending. Prison visits had to be included, and most evidently included a presentation by inmates, although this could be more, or less, 'graphic' or 'educational'. Of 487 items of literature, 30 evaluation studies were identified of which 11 were potential randomised trials. Of the 11, two had to be eliminated either because data could not be obtained or because there was not random assignment. The nine remaining were all from the United States, and one is excluded from some of the analysis due to violations in random assignment protocols.

The stark conclusion of the review is that: "The (randomised trials) provide empirical evidence – under

experimental conditions – that these programmes likely increase the odds that children exposed to them will commit offences in future."

So, what's wrong with this? From a realist perspective, there are some critical but unanswered questions. Consider the following, as they relate to Scared Straight.

- Was it because the children were not scared?
- Or that they liked being scared, and so were unintentionally encouraged to be criminal?
- Or, were they risk-hungry, so inclined to undertake activities that had a revealed significant risk?
- Or, did different subsets of children react differently, some being put off but others encouraged by the prison visits, but with numbers balanced?
- Or, was the dosage too low to have an effect?
- Or, was the dosage so high that it had a negative effect?
- Or, for those already inclined to commit crime, was prison demystified and hence less frightening?
- Or, did the children soon return to other more salient and more immediate influences?
- Or, did the intervention confirm or bestow deviant identities on some participants?
- Or, did the visits bring (pre)delinquent children together and cement their distinctiveness and separation from relatively law-abiding youngsters?
- Or, did the children discount the messages from the offenders as cynically and strategically produced to achieve their own ends within the prison?
- Or were the prisons visited and the offenders talking to the children so alien to them, they seemed to be visiting a zoo, looking at an interesting spectacle irrelevant to their daily lives?
- Or...?

These are 'realist' questions about *how* subjects interacted with the programmes. In contrast to medical trials that try to exclude the perceptions and reasoning of those treated, in social interventions these are the very guts of how programmes operate. Why do these questions matter at all, however, given that the programmes consistently failed? The problem is that without answers to these questions we do not know what failed: the idea(s), the implementation, or the target population. We also do not know whether there was a mix of success or failure within the programme: whether there were counterbalancing impacts amongst different sub-groups.

The ultimate paucity of the pay-off is revealed where the authors explain the implications of their experimentalist review. They say, "...the onus is on every jurisdiction to show how their current or proposed (prison visit) program is different than the ones studied here. Given that, they should then put in place rigorous evaluation to ensure that no harm is caused by the intervention."

The authors cannot, and do not, conclude that prison visit programmes cannot deter young delinquents or pre-delinquents, only that the specific set failed. They can provide no guidance on what kinds of prison visit programme is liable to be experienced in what ways by children of differing kinds. They can draw no conclusions about implications findings might have for efforts to scare children in a different setting. There is and can be no indication of what would count as a similar programme or a replication. It is not clear what would be gained by another experimental study, whether it showed a net benefit, no net effect, or a net negative effect. They can conclude at best that these programmes with these children in these places at these times were associated with a small net negative effect. Any of the remaining 479 items of literature that might throw light on the unanswered questions is cast aside on technical grounds. Much might be gleaned from them. Finckenauer *et al's* book on Scared Straight makes some use of experimental findings but refers to much else besides. Many of the ingredients -- programme change and variation, participant variation, sub-set variations in response etc -- are there to build a richer picture of what can be learned from the programme, including a specific explanation of why any panacea such as that promised by the programme is unlikely to be appropriate, given the diversity of offenders and those at risk (Finckenauer *et al*, 1999).

(context), and 'Considerations' (including, for example, side-effects, costs, implementation problems, and some 'for whom' notes).

These strike me as much more useful. The contents of these guides relate specifically to issues facing the police. Their general features might profitably be emulated in other fields in criminal justice also. With a strong emphasis on a proper initial analysis of the specifics of presenting problems, they also avoid the risks of the sorts of panacea attacked by Finckenauer *et al*.

To conclude, nothing since Ray Pawson and I published *Realistic Evaluation* in 1997 has lessened my scepticism about experimental studies, or dimmed my view that a realist approach is preferable to experimentalism in terms both of validity and usefulness. Indeed, developments since 1997 seem to me to have reinforced the arguments in that book. None of this makes me believe that *Realistic Evaluation* is without flaws, though the basic argument still stands. The flaws are in the details and in the implication that conducting realist evaluations is a breeze. It isn't, but I continue to believe it is generally the best way to go.

Lastly, attentive readers will notice that I used scare quotes round 'experimental' at the start of this article. The reason is that there are real experiments in criminal justice that do not

For realists, the key question is not 'What works?' but 'What works for whom in what circumstances?'

For realists, the key question is not 'What works?' but 'What works for whom in what circumstances?' This focuses on mechanisms (how are the effects brought about through the way subjects interact with the measure(s) introduced?) and 'context' (what conditions are needed for the mechanisms to be activated?) The aim is to produce portable, middle-range 'context-mechanism-outcome-pattern configurations'. *Realistic Evaluation* may have made the production of these seem too simple, too easily read from programmes, and too easily cranked out with standard techniques, rather as experimental evaluations can be cranked out by anyone who has grasped the basic logic. Done well, they require theoretical sophistication and the ability to design studies appropriate to the theory, with no bar to the kinds of data and method used. Ray Pawson has recently argued strongly, but counter-intuitively, that even weak research may fruitfully be drawn on (Pawson, forthcoming). I have recently argued that for most purposes good practice guidance needs to be informed by realist readings across wide ranges of literature, rather than depending on findings of 'experiments' (Tilley 2006).

In contrast to the 'experimental' approach, I turn now to some recent work tacitly incorporating a realist method, and published by the US Department of Justice, Office of Community Oriented Policing Services (available at: www.cops.usdoj.gov). Over 30 guides have been produced since 2001, drawing together a wide range of research findings to suggest what works for whom in what context in relation to specific problems (for example street prostitution, rave parties, and acquaintance rape of college students). Having gone through the research findings (which do not exclude experimental studies), an instructive summary table is drawn up in each guide with headings including 'Response' (what's done), 'How it works' (mechanism), 'Works best if...'

involve RCTs or their close equivalents, and they can be very informative. Experiments, properly speaking, do not require RCTs or RCT look-alikes. Think about your school physics and chemistry! And in applied work, think about engineering!

Nick Tilley is Professor of Sociology and Head of the Policy-orientated Social Sciences Research Group at Nottingham Trent University.

References

- Finckenauer, J., Gavin, P., with Hovland, A. and Storvoll, E. (1999) *Scared Straight*. Long Grove Illinois: Waveland Press.
- Pawson, R. (forthcoming) *Evidence-based Policy: A Realist Perspective*. London: Sage.
- Pawson, R. and Tilley, N. (1997) *Realistic Evaluation*. London: Sage.
- Petrosino, A., Turpin-Petrosino, C. and Buehler, J. (2002) "'Scared Straight" and Other Juvenile Awareness Programs for Preventing Juvenile Delinquency' in *The Campbell Collaboration Reviews of Intervention and Policy Evaluations (C2-RIPE)*. Philadelphia, Pennsylvania: Campbell Collaboration.
- Tilley, N. (2006) 'Knowing and Doing: Guidance and Good Practice in Crime Prevention' in J. Knuttson and R. Clarke (eds.) *Crime Prevention Studies*, Vol 20. Cullompton, Devon and Monsey, New York: Willan and Criminal Justice Press.