Being more certain about Random Assignment in Social Policy Evaluations

Bruce Stafford

Director of the Centre for Research in Social Policy, Loughborough University
Email: b.r.stafford@lboro.ac.uk

Social experiments have been widely utilised in evaluations of social programmes in the US to identify 'what works', whilst in the UK their use is more controversial. This paper explores the paradigmatic, technical and practical issues evaluators confront in using randomised experiments to evaluate social policies. Possible remedies to some of these problems are outlined. It is argued that although no evaluation methodology is problem-free, policy makers and researchers should be more confident about the merits of using random assignment, provided it is used in conjunction with other methodologies more suited to understanding why and how interventions work.

Introduction

In the UK the New Labour Government's rhetoric on policy-making highlights a concern with 'what works'. The implication is that the policy knowledge base is more important than, or at least on par with, ideology in developing policies. It follows that evidence of a high quality is preferable to data that are less valid, reliable or comprehensive. Where research evidence is used in policy-making the quality of the information used will largely reflect the robustness of methodology adopted. Social policy analysts have a number of approaches they can use to assess the impacts of social programmes. One of these methodologies is social experiments, or randomised control trials (RCTs), which involve randomly assigning people to groups that do and do not receive a programme or intervention. The estimate of a programme's impact is obtained by comparing the outcomes (for example, employment, earnings or number of GCSEs attained) for those participating in the programme with those not in receipt of the intervention. As such social experiments can address the question: *'What works better and in what sense? For whom? And at what cost?'* (Boruch, 1997: 11).

Social experiments have been used in over 200 evaluations of social programmes in the US covering, for example, welfare-to-work, education and training initiatives, low income housing assistance and negative income taxes (Greenberg and Schroder, 1997; Greenberg *et al.*, 1999). They are less widely used in Europe. In Britain, for instance, the effects of labour market policies are typically measured using quasi-experimental approaches. Social experiments have been used in only relatively small-scale evaluations of labour market policies, such as Restart Interviews, 13 week reviews and Jobplan as well as a few New Deal initiatives (Stafford *et al.*, 2002).

Like all research methodologies random assignment has its pros and cons. This paper

discusses the use of random assignment and argues that, in conjunction with other relevant methodologies, it should be more widely used in the UK. The next section outlines why social experiments are used and their strengths. This is followed by a discussion of the paradigmatic, technical and practical obstacles to using random assignment in evaluating social policies.

Social experiments

Social experiments are designed to measure the net impact or additionality of a programme, for example, the average number of additional people in work as a result of an employment initiative. In measuring net impacts the key evaluation question is what is the difference between what happened and what would have happened in the absence of the programme. The latter, known as the counterfactual, is for those participating in the programme often unobservable, and evaluators have devised a number of methodologies, including social experiments, to overcome this problem.

In a social experiment people (or, occasionally, areas – see, for instance, Riccio, J., 2000) are randomly assigned to one or more action groups who have access to the programme and one or more control groups, who during the evaluation period are denied the programme. The control group is the counterfactual, and as such solves the evaluation problem by providing a group whose outcomes can be observed and compared to programme invitees. It is the randomisation process that makes the methodology so potentially powerful, and why experimental designs are regarded by some, but not all, evaluators as the 'gold standard' in evaluation. For if conducted properly, random assignment produces well-matched samples and consequently avoids the selection bias inherent in other impact methodologies. As the only differences between the action and control groups should be attributable to the programme (and random variation), any observed or unobserved systematic differences (for example, in motivation, gender or class) that might affect the variation in outcomes between people in the action and control groups can be discounted in assessing net impacts (provided the sample is of a sufficient size). Such factors though will be important in understanding why and how a programme works, and will need to be explored using non-RCT methods.

As well as producing unbiased estimates of mean net impacts, social experiments deliver estimates that are internally valid and researchers can state their degree of statistical confidence that the estimate is the 'true' measure of the impact of the intervention (Boruch, 1997; Orr, 1999).

However, the use of randomised experiments is controversial in evaluating social programmes. There are paradigmatic (at the epistemological and ontological levels), technical and practical objections to the use of random assignment and these are briefly considered in turn below.

Paradigmatic objections to social experiments

There are social policy academics and analysts with fundamental objections to the experimental approach (Pawson and Tilley, 1997; Guba and Lincoln, 1989). It is an approach that can be seen as falling within the 'positivistic' tradition, although this term is not necessarily a useful label because it can be used in a derogatory way. Whilst the

precise meaning of positivism is not unambiguous (Halfpenny, 1982), it does emphasise that social researchers should adopt the methodology of the natural sciences and that there are causal laws to be discovered that will predict and explain social phenomena. This is seen as requiring the empirical testing of (probabilistic) theories through the rigorous application of deductive reasoning and objective measurement of events (Marsh, 1982). Other social science traditions, which fall under a variety of headings (including humanism, interpretative social science, phenomenology, symbolic interactionism, ethnography and realism) correctly point out that social realities are socially constructed, focus on understanding the meaning of social action (agency and structure), and highlight that research is itself a social process with interactions between researcher and subject (Bauman, 1978; Berger and Luckman, 1966; Blumer, 1969; Giddens, 1976). Such approaches have criticised social experimenters for ignoring the complexities of social situations. Critically the power relations and nuances of social interaction that can be vital in influencing the implementation and outcomes of a programme are not explored in RCTs. Moreover, the theory of causation underlying the experimental approach is said to be faulty because it does not take account of the causal powers and liabilities of objects which depending upon the prevailing conditions will have differential effects (Harré, 1972; Pawson and Tilley, 1997; Sayer, 1984). That is, that the simple (or successionist) cause → effect model inherent in experiments is inadequate for understanding social situations, instead it is a starting point for explaining social phenomena (see Harré, 1972). As a consequence the experimental approach can reveal what works but not why or how a programme works.

The thrust of this criticism is to some extent acknowledged by proponents of randomised experiments, and is often referred to as the 'black box' problem (Riccio and Bloom, 2001). Social experiments provide an estimate of the net impact of an intervention in its entirety and cannot explain why programme participants achieved their observed outcomes. In practice, experimenters attempt to address this by conducting process studies, multi-group designs and a variety of quasi-experimental statistical analyses. Process studies use a variety of methodologies (including qualitative techniques) to explore programme awareness and participation, the implementation and delivery of interventions and the attitudes and experiences of stakeholders (Purdon *et al.*, 2001). Multi-group designs allow different treatments to be compared with one another and this helps to reveal what aspects of an intervention are effective (Riccio and Bloom, 2001). The statistical methods used to unpack the 'black box' include pooling data in order to compare similar programmes across different locations and analyses of outcomes and impacts for sub-groups of participants (Greenberg *et al.*, 2001). These non-experimental analyses exploit the random samples provided by RCTs to investigate the differential impacts of a programme on groups of participants and the effect of organisational, managerial and community/environmental factors on their outcomes/impacts (Bloom *et al.*, 2001; Riccio and Bloom, 2001).

Notwithstanding the interpretative paradigm's critique of social experiments, it remains the case that experiments are preferable, indeed for some superior, to non-experimental methods when measuring the effectiveness of programmes in terms of outcomes and impacts (Betsey *et al.*, 1985; Boruch, 1997; Fraker and Maynard, 1987; LaLonde 1986). Their use, therefore, in part depends upon the aims and purposes of an evaluation: in other words is a measure of additionality required. Social experiments provide statistical confidence and precision in estimates of programme effects, but at the

cost of requiring a high degree of control over interventions as departures from the random assignment will undermine study findings. More specifically there are technical issues that social analysts must address.

Technical issues in conducting social experiments

All evaluation methodologies have technical limitations on their applicability and implementation. Social experiments should not be used where it is impossible to distinguish between the consumption of an intervention by individuals. For example, with a community-wide programme like a Health Action Zone it is not viable to randomly assign people living in the same area to action and control groups (Purdon *et al.*, 2001).

In addition, social experiments provide an unbiased estimate of the average net impact and not of the distribution of impacts. The estimates do not distinguish between situations where most people received the average amount and those where a few benefit greatly and the majority make small or no gains.

The external validity of a social experiment also needs to be assessed as the social, political and economic context to the study and the representativeness of the participant samples may have changed since the experiment commenced.

Moreover, there are a number of potential biases that can undermine RCT estimates of programme effects, these include (Burtless and Orr, 1986; Björklund and Regnér, 1996; Cook *et al.*, 1997; Friedlander *et al.*, 1997; Heckman and Smith, 1996; Orr, 1999; Soloman and Draine, 1995):

- Queuing bias – which can arise because in a social experiment only a (small) percentage of the eligible population receives the intervention and this may give them a comparative advantage (or, depending upon the programme, a relative disadvantage) to the control group that they would not possess if the programme was fully implemented. This serves to exaggerate the impact of the programme. On the other hand, as only a small proportion of the in-scope population receives the intervention the evaluation may only estimate its partial impacts compared to when the programme is fully implemented (this is known as the partial equilibrium effect).
- Substitution bias – arises when controls access a similar (or, indeed, the actual) treatment to that received by the action group, but had the evaluation not taken place they would not have received it. This reduces the 'service differential' between the action and control groups and consequently means that the 'true' net impact will be underestimated.
- Randomisation bias – occurs if the random assignment itself influences the behaviour of people. For example, if some people refuse to take part in a programme because of the use of random assignment this could conceivably limit the generaliseablity of the study's findings. Whilst others might alter their behaviour simply because they are participating in a study, rather than because of the programme itself.
- Disruption bias – arises if the RCT influences the behaviour of programme staff, and, say, they manipulate the outcome of the random assignment so affecting who receives the intervention.

- Response and attrition bias – is due to attrition from samples and differential survey response rates between action and control groups.

The existence of these biases is well know and designers of social experiments can seek to minimise them through the careful design and execution of studies and application of statistical techniques (such as the 'no-show' correction factor (Orr, 1999) and the weighting of data). Whilst these are serious potential sources of bias they do not invalidate the use of social experiments.

Potential operational problems when using random assignment

The implementation of experimental designs is not straightforward. The principal practical difficulties in the use of random assignment can include:

- opposition to random assignment by service provider staff, volunteers/clients, and local and national groups;
- the extra 'burden' of administering the random assignment;
- low take-up of the programme;
- difficulties encountered in managing some individuals assigned to the control group; and
- threats to maintaining the integrity of the random assignment design.

Evaluators of social programmes in the US have had to address these issues and arguably social policy analysts elsewhere can learn from their experience and so minimise the potential risks.

Opposition to the use of random assignment

American experience shows that resources, time and effort have to be devoted to gaining acceptance for randomised experiments (Doolittle and Traeger, 1990; Gueron, 1999; Orr, 1999). Opposition may stem from programme staff, (potential) clients and local and national organisations. Staff concerns in particular should be taken seriously, as their cooperation is a prerequisite for a successful evaluation (Björklund and Regnér, 1996).

There can be extensive and deeply held ethical objections to use of random assignment. Random assignment can be seen as unfair when, say, the intervention is regarded as beneficial and it is denied to members of the control group, even if access to the programme is only deferred for a set period. The ethics of denying access to controls needs to be considered in two different experimental contexts: demonstration projects and on-going policies (Boruch, 1997; Gueron, 1999; Orr, 1999). Ethically, it is easier to justify using random assignment in evaluations of demonstration projects. This is because, first, the denial of services to controls leaves them in the same situation as if there was no evaluation – existing services remain available to them. Secondly, if the programme turns out to cause harm or is ineffective, then it is arguably better that it is tested on a small scale than on the wider population.

For an on-going programme it is harder to justify the denial of an intervention which would otherwise be available to individuals. However, this has to be weighted against gains in knowledge about 'what works'. Moreover, social experiments may be reason-

able where resources are scarce and there is an excess demand for a programme, as random assignment may be a fairer way of allocating resources than some other methods administrators might adopt, such as first-come, first-served. Or it might be acceptable if control groups are compensated in some way for their loss. The difficulty with compensating control groups is that their response can be uncertain. For example, relatively high financial recompense might mean control groups can purchase a similar intervention to that denied, whilst a smaller sum may have no tangible affect on their behaviour. This may lead to an under- or over-estimate of the programme's effects.

Notwithstanding the experimental context, there is an onus on the evaluators and project funders to minimise the risk of personal loss to the client group. In seeking to alleviate people's concerns US evaluators usually point out that the findings from the evaluation will be of benefit to the client group (as well as to taxpayers) as it will highlight what works (Boruch, 1997; Gueron, 1999; Newman and Brown, 1996; Orr, 1999). It can be argued that it would be unethical to allow potential programme clients to receive an ineffective and poorly evaluated service. Indeed, an outcome will be better social policy.

There is also the assurance that informed consent is required before random assignment is conducted. For vulnerable groups evaluators can request that parents/guardians and/or staff from referral agencies be present during the initial meeting when informed consent to participate is sought.

If ethical objections are keenly held then evaluators can modify the experimental design, for example, exempting certain sub-groups from the random assignment, allowing staff a limited number of discretionary exemptions for clients they judge in dire need of the intervention, shortening the embargo period, and providing control group members with a list of non-programme services. The latter may limit more pro-active advice giving by staff, and avoid staff turning people away empty handed. However, the extent to which such lists lead members of control groups to access services that they would not otherwise have used biases any impact estimates. Although not ideal, issuing lists of alternative services/treatments is a widespread practice.

In addition, staff delivering the programme may feel that the experiment has nothing to offer the service provider. Nevertheless, the programme being evaluated must be administered in the 'normal way'; both non-cooperative and over-enthusiastic staff can undermine an evaluation. The US literature on social experiments includes suggestions for engendering the support and cooperation of provider organisations, such as, agreeing to provide it with interim and summary results, and showing that random assignment has worked elsewhere (Stafford *et al.*, 2002). Providers may also be keen to cooperate because evidence from the evaluation may help them to secure funding for the same (or a similar) programme after the evaluation.

Extra burden of administering random assignment

Administering random assignments will place an additional burden on service providers. However, experimenters can often explore measures to reduce the load on organisations, such as, integrating the random assignment procedures with the service provider's own proposed application process and conducting the random assignment procedure early on in the application process (Orr, 1999).

Low take-up

There is a risk that take-up of programmes involving random assignment will be lower than anticipated. This could lead to insufficient sample sizes, which in turn reduces the statistical power of the evaluation (Doolittle and Treager, 1990; Heckman and Smith, 1996). Or, programmes may admit applicants who they would not 'ordinarily' take on, so diminishing the performance of the programme. Failure to estimate correctly the size of the target population and participation rates has been a feature of US employment and training programmes (Boruch, 1997).

Nevertheless there are a variety of recruitment methods to attract potential clients to a programme. Methods used in the US include:

- conducting a feasibility study to investigate the likely number of applicants;
- targeted mailings;
- adjusting temporarily or permanently the treatment group/control group ratio to allow more people to receive the intervention;
- providing staff with specialist training in marketing;
- extending the time period over which recruitment takes place;
- offering financial and other inducements (e.g. gifts, entry to a prize draw) for taking part. (Any incentive must be made available to all members of the control and intervention groups); and
- providing feedback to organisations on the progress of any clients they refer to the programme.

Managing controls

Staff may be concerned about how to deal with complaints about the use of random assignment (especially from those assigned to a control group). Failure to have procedures to deal with applicants' complaints can lead to staff referring controls to other sources of information or services they would not otherwise have received. As already mentioned, such contamination of the control group will bias the estimates of the impact of the programme. Minimising control group contamination requires staff to understand and accept that there are limits to the information they can give to controls.

Possible mechanisms for minimising complaints include, first, holding a pre-random assignment orientation session where it is explained that receipt of the intervention is a lottery and that random assignment is a fair method of allocation (Stafford *et al.*, 2002). Secondly, ensuring that front-line staff have the skills and expertise to establish rapport with clients, especially vulnerable applicants (Boruch, 1997). This may involve agreeing beforehand what the staff is going to tell aggrieved applicants and providing them with scripts, and/or videos and other guidance.

Maintaining the integrity of the random assignment

The integrity of the random assignment must be maintained for the duration of the project (Gueron, 1999). However, it is almost inevitable that some members of the control group receive a similar or the same service as those in the intervention group (Friedlander *et al.*, 1997). In addition, substitution bias (see above) can occur if the very

existence of the programme alters the services that are available to members of the control group. This would occur, for example, if the intervention leads to more places in conventional services being available to the control group because members of the action group are no longer using them.

Development of random assignment procedures involves evaluators agreeing with providers the protocols and procedures for quality assurance. These may encompass monitoring and policing the delivery of the intervention, piloting procedures, observing the random assignment, checking the baseline characteristics of the action and control groups and monitoring the build-up of the sample on a weekly basis (Stafford *et al.*, 2002).

There are practical problems to be overcome in using social experiments. However, the evidence from the US is that they can be minimised with careful planning and execution of RCTs (Stafford *et al.*, 2002).

Conclusion

This paper has considered some of the objections to random assignment and has attempted to highlight how they might be addressed. Nevertheless, the use of social experimentation in the UK is controversial. For example, the current phase of the New Deal for Disabled People was originally to be evaluated by a study that incorporated a randomised experiment. Much to the delight of critics of social experiments the Government announced the abandonment of this aspect of the evaluation on the 10th December 2001 (Prasad, 2002).

The challenges evaluators confront in applying randomised experimental designs are formidable. However, the paradigmatic, technical and practical concerns about social experiments are not fatal, provided the evaluation strategy includes other methods that explore key why and how research questions (Macdonald, 1999; Cronbach, 1982; Rossi and Freeman, 1985). Such a pluralistic approach to social policy evaluation reflects the complexity of social phenomenon and allows the impacts of schemes to be measured and understood. Social experiments are a powerful methodology which social policy analysts should use when appropriate, not for their own sake, but because RCTs that are properly designed and executed are capable of letting us know what works, with confidence.

Acknowledgements

References

Bauman, Z. (1978), *Hermeneutics and Social Science: Approaches to Understanding*, London: Hutchinson.

Berger, P. and Luckman, T. (1966), *The Social Construction of Reality: A treatise in the sociology of knowledge*, Harmondsworth, Middlesex: Penguin.

Betsey, C., Hollister, R. and Papageorgiou, M. (1985), *Youth Employment and Training Programs: The YEDPA Years*, Committee on Youth Employment Programs, Commission on Behavioral and Social Sciences and Education, National Research Council, Washington, DC: National Academy Press.

Bjökland, A. and Regnér, H. (1996), 'Experimental evaluation of European labour market policy', in G. Schmid, J. O'Reilly and K. Schömann (eds) International Handbook of Labour Market Policy and Evaluation, Cheltenham: Edward Elgar, 89–114.

Bloom, H., Hill, C. and Riccio, J. (2001), 'Modelling the performance of welfare-to-work programs: the effects of program management and services, economic environment, and client characteristics', MDRC Working Paper on Research Methodology, MDRC, New York.

Blumer, H. (1969), *Symbolic Interactionism: Perspective and Method*, New Jersey: Prentice-Hall.

Boruch, R. (1997), *Randomized Experiments for Planning and Evaluation: A Practical Guide*, Thousand Oaks, CA: Sage.

Burtless, G. and Orr, L. (1986), 'Are classical experiments needed for manpower policy?', *Journal of Human Resources*, 21, 4, 606–639.

Cook, T., Cook, F. and Mark, M. (1997), 'Randomised and quasi-experimental designs in evaluation research: an introduction', in L. Rutman (ed.), *Evaluation Research Methods*, Beverley Hills: Sage.

Cronbach, L. (1982), *Designing Evaluations of Educational and Social Programs*, San Francisco: Jossey-Bass.

Doolittle, F. and Traeger, L. (1990), *Implementing the National JTPA Study*, New York: MDRC.

Fraker, T. and Maynard, R. (1987), 'The adequacy of comparison group designs for evaluations of employment-related programs', *Journal of Human Resources*, 22, 194–227.

Friedlander, D. Greenberg, D. and Robins, P. (1997), 'Evaluation government training programs for the economically disadvantaged', *Journal of Economic Literature*, 35, 1809–1855.

Giddens, A. (1976), *New Rules of Sociological Method: A Positive Critique of Interpretative Sociologies*, London: Hutchins.

Greenberg, D. and Shroder, M. (1997), *The Digest of Social Experiments*, Second edition, Washington, DC: Urban Institute Press.

Greenberg, D., Meyer, R., Michalopoulos, C., and Wiseman, M. (2001), 'Explaining variation in the effects of welfare-to-work programs', Institute for Research on Poverty, Discussion Paper No. 1225-01, *www.ssc.wisc.edu/irp*.

Greenberg, D., Shroder, M., Onstott, M. (1999), 'The social experiment market', *Journal of Economic Perspectives*, 13, 3, 157–172.

Guba, E. and Lincoln, Y. (1989), *Fourth Generation Evaluation*, Newbury Park: Sage.

Gueron, J. (1999), *The Politics of Random Assignment: Implementing Studies and Impacting Policy*, New York: MDRC.

Halfpenny, P. (1982), *Positivism and Sociology: Explaining Social Life*: London: George Allen & Unwin.

Harré, R. (1972), *The Philosophies of Science*, Oxford: Oxford University Press.

Heckman, J. and Smith, J. (1996), 'Experimental and nonexperimental evaluation', in Schmid, G., O'Reilly, J. and Schömann, K. (eds.), *International Handbook of Labour Market Policy and Evaluation*, Cheltenham: Edward Elgar, 37–88.

LaLonde, R. (1986), 'Evaluation the econometric evaluations of training programs with experimental data', *American Economic Review*, 76, 604–620.

Macdonald, G. (1999), 'Social work and its evaluation: A methodological dilemma?', in F. Williams, J. Popay and A. Oakley (eds.), *Welfare Research: A critical review*, London: UCL Press, 89–103.

Marsh, C. (1982), *The Survey Method The contribution of surveys to sociological explanation*, London: George Allen & Unwin.

Newman, D. and Brown, R. (1996), *Applied Ethics for Program Evaluation*, Thousand Oaks, CA: Sage.

Orr, L. (1999), *Social Experiments: Evaluating Public Programs with Experimental Methods*, Thousand Oaks, CA: Sage.

Pawson, R. and Tilley, N. (1997), *Realistic Evaluation*, London: Sage.

Prasad, R. (2002), 'U-turn on disability work plans', *The Guardian*, Guardian Society, 9 January, 5.

Purdon, S., Lessof, C., Woodfield, K. and Bryson, C. (2001), 'Research methods for policy evaluation', Department for Work and Pensions Research Working Paper No. 2, DWP, London.

Riccio, J. (2000), 'Extending the reach of randomized social experiments: new directions in evaluations of American welfare-to-work and employment initiatives', Paper prepared for meeting of the Royal Statistical Society, 4 July, Draft.

Riccio, J. and Bloom, H. (2001), 'Extending the reach of randomized social experiments: new directions in evaluations of american welfare-to-work and employment initiatives', MDRC Working Paper on Research Methodology, MDRC, New York.

Rossi, R. and Freeman, H. (1985), *Evaluation: A Systematic Approach*, Beverley Hills: Sage.

Sayer, A. (1984), *Method in Social Science: A Realist Approach*, London: Hutchinson.

Soloman, P. and Draine, D. (1995), 'One year outcomes of a randomised trial of case management with seriously mentally ill clients leaving jail', *Evaluation Review*, 19, 3, 256–273.

Stafford, B., Greenberg, D. and Davis, A. (2002), 'A literature review of the use of random assignment methodology in evaluations of us social policy programmes', DWP In-house Report No. 94, DWP, London.