Criminology & Criminal Justice

www.sagepublications.com
ISSN 1748–8958; Vol: 8(1): 89–106
DOI: 10.1177/1748895807085871

Evaluating offending behaviour programmes:
Does only randomization glister?

CLIVE R. HOLLIN
University of Leicester, UK

Abstract

Despite considerable investment there has been a marked reluctance by the Home Office to publish the evaluations of the various Pathfinder Programmes. Arguably, this reluctance stems from the 'official' view that the commissioned researchers conducted the wrong type of research, specifically in not using randomized control trials (RCTs). The utility of RCTs is considered here with particular reference to the evaluation of the Offending Behaviour Pathfinder Programmes. It is argued that the Home Office 'Reconviction Scale', favouring RCTs, is seriously flawed and is used to present a misleading view of the extant research. An overview of the wider literature shows that RCTs are not uniformly agreed to be the single design of choice in evaluating complex interventions such as offending behaviour programmes. The trend in disciplines such as the clinical sciences, with a history steeped in RCTs, is to utilize a range of research designs, both quantitative and qualitative, to evaluate complex interventions.

Key Words

evaluation • offending behaviour programmes • Pathfinder Programmes • randomized control trials • research design

Introduction

Following dissemination of the 'What Works?' research (McGuire, 1995, 2002), offending behaviour programmes have been adopted in England and Wales by both the prison and probation services. In particular, the advent of Pathfinder Programmes within the probation service represented a significant public investment in evidence-based practice. As Raynor (2004) notes, evaluation of the four Pathfinder Programmes—offending behaviour programmes, basic skills, enhanced Community Service and resettlement projects for short-term prisoners—was commissioned by the Home Office from independent researchers. The main purpose of evaluation within the context of evidence-based practice is to inform and refine the effectiveness of practice. Thus, the collective efforts of the various Pathfinder research groups have the potential greatly to inform the field, as Raynor again notes:

> The group of evaluative studies carried out under the aegis of the probation Pathfinders represents, at least proportionately, a massive increase in our research-based knowledge of what happens to offenders who are being supervised in the community or prepared for release from prison.
>
> (2004: 320)

Writing as an evaluator of one of the Pathfinders, Raynor (2004) comments that the evaluations were heavily stage managed by the Research, Development and Statistics Directorate (RDS) within the Home Office. As the research programmes progressed so various interim reports on implementation were produced (e.g. Hollin et al., 2002; Lewis et al., 2003) and some preliminary reconviction studies appeared (e.g. Hollin et al., 2004; Stewart-Ong et al., 2004). However, as Maguire and Raynor state with respect to the Resettlement Pathfinder (although the point is generic), the Home Office has shown a distinct 'Lack of interest in disseminating the findings of the second phase of the Pathfinder study' (2006: 32). Thus, paradoxically, the increase in knowledge from the Pathfinder research, particularly with regard to reconviction, has been mainly limited to those conducting the evaluations, to conference presentations (e.g. Hollin, 2005; Palmer, 2005) and to publications independent of the Home Office (Clancy et al., 2006; Lewis et al., 2007; Palmer et al., 2007).

There may be several explanations for the Home Office's reluctance to publish fully the Pathfinder research. For example, the findings are inevitably complex and do not produce clean messages by which to hail a policy success: indeed, some commentators have suggested that when the evidence does not suit the policy, the pressures of policy outweigh the integrity of independent research so that data and findings are managed accordingly (Hope, 2004).

None the less, the absence of official Home Office publication of the substance of the Pathfinder studies requires an explanation. Raynor takes the view that the explanation apparently favoured by RDS is that the Pathfinder research should have used randomized control trials (RCTs) and

hence the evaluators 'Did the wrong kind of research' (2004: 319). A comment by a Home Office researcher reinforces this point:

> There is a need to develop randomised control trials in the correctional services, so that our knowledge of what works is truly improved and the existing equivocal evidence is replaced with greater certainty and ultimately, greater confidence for the correctional services that they are delivering effective interventions with offenders.
>
> (Chitty, 2005: 80)

The point that RCTs are the research 'gold standard' is driven home elsewhere in the same government publication, written and edited by Home Office researchers (Harper and Chitty, 2005). Thus, Elliott-Marshall et al. argue that while there is a substantial body of evidence to suggest that interventions can reduce reoffending, 'There is limited evidence to demonstrate what impact these interventions have in practice. There is also evidence of research failure … the design of most studies looking at outcomes is significantly below the gold standard' (2005: 68). The same orchestrated theme reappears elsewhere in this publication:

> Current evidence in the UK is predominately based on quasi-experimental or non-experimental evaluation studies, which makes it difficult to attribute the outcomes to the effects of the treatment or intervention … Outcome studies therefore should be based on more effective research designs.
>
> (Debidin and Lovbakke, 2005: 51)

In truth, the Home Office and random allocation studies have a long history (Nuttall, 2003), however the use of language in the Harper and Chitty (2005) publication is interesting on two fronts. First, there is reference to 'the existing *equivocal* evidence', the '*limited* evidence' and that it is '*difficult*' to interpret any evidence from a quasi-experimental study. These comments might easily lead to the conclusion that there is very little trustworthy evidence to help answer the question 'What Works?' Second, the statement is made repeatedly that RCTs are the gold standard in evaluating offending behaviour programmes: thus, there is a 'need to develop randomized control trials' to tackle the 'research failure' of studies that fall below the gold standard, and that 'more effective research designs' are absolutely necessary. Indeed, it might be thought that any empirical evidence other than from an RCT is scientifically unsafe and that, as Raynor suggests, the blame for the absence of publication of the Pathfinder evaluations has been directed at the independent researchers and their failure to use RCTs.

Ironically, much of the evidence that informed 'What Works?' (and so precipitated the development of offending behaviour programmes and the Pathfinders) came from studies using quasi-experimental, rather than randomized, designs. It is self-evident that RCTs would be a welcome addition to any body of knowledge, including the evaluation of offending behaviour programmes. However, would RCTs really offer 'greater certainty' and 'greater confidence' over and above the evidence that is currently available?

The facile answer to this question is that they would because the Home Office has a scale that says so.

Scientific Methods Scale

In commenting on evaluation within the criminal justice system, Gondolf observes that:

> [Evaluation] is a difficult and complex task that complicates the interpretation of the evaluation results. As has been the case in other fields, such as alcohol, sex offence, and depression treatment, different program conceptions, outcome measures, research designs, and statistical analyses can produce contrary results … Program evaluations that specifically address these sorts of issues are likely to further their validity, and those that at least acknowledge them will help clarify interpretation of the results.
>
> (2004: 607–8)

The point Gondolf makes with respect to *validity* is important. For applied research reliably to inform practice it must strive for high levels of validity. Cook and Campbell (1979) describe four types of validity: these are, construct validity, external validity, internal validity and statistical conclusion validity. In a quantitative evaluation of offending behaviour programmes all four types of validity are important, but they are neither independent of each other nor even necessarily sympathetic. For example, an evaluation may maximize its internal validity by, say, using strict sampling criteria and having absolutely rigorous control over the running of the offending behaviour programme. However, achieving high internal validity may be at the expense of external validity: rigorous control over the intervention may create such artificial circumstances that the findings from the evaluation are meaningless in the real world where such tight controls are not feasible. Clearly, there is a balance to be struck between maximizing internal validity while maintaining external validity.

Not all research methodologies are of equal utility with respect to validity: in the criminological literature, the Scientific Methods Scale (SMS), devised for the Maryland Report (Sherman et al., 1997), is a system for ranking research designs that is widely disseminated and applied (Farrington et al., 2002; Wilson et al., 2005). As summarized in Table 1, the SMS ranges from a basic correlational design at one extreme to a fully randomized control trial (RCT) at the other. As Farrington et al. (2002) note, the Scientific Methods Scale has its single focus on *internal* validity; it makes no reference to the other three types of validity described by Cook and Campbell.

Of the five research designs included in the SMS it is the latter three that are generally taken as producing evidence of acceptable scientific quality. Following Wilson et al. (2005), Level 3 equates to a low-quality quasi-experimental design, with the threat to internal validity resulting from

Table 1. The Scientific Methods Scale (after Sherman et al., 1997)

1. A simple correlation between a crime prevention programme and some measure of crime.
2. A temporal sequence between the crime prevention programme and the measure of crime clearly observed; or the use of a comparison group but without demonstrating comparability between the comparison and treatment groups.
3. A comparison between two or more groups, one participating in the programme, the other not.
4. A group comparison, with and without the programme, in which there is control of relevant factors or a non-equivalent comparison group with only minor differences from the treatment group.
5. Random assignment to groups with analysis of comparable units for programme and comparison groups.

a selection bias due to uncontrolled differences between the treatment and comparison groups. Level 4 is a high-quality quasi-experimental design, in which the threat to internal validity through the absence of randomization to condition (i.e. treatment and control conditions) is countered by either methodological or statistical control of group differences. Finally, Level 5 is an experimental design in the proper sense, with randomization of allocation to condition to attain high levels of internal validity.

The SMS has been modified by Home Office researchers to assess reconviction studies (Friendship et al., 2005), as shown in Table 2, and its relationship with the SMS is clearly evident.

Now, as Farrington et al. stress, the Scientific Methods Scale 'Focuses only on internal validity' (2002: 17): thus the Home Office adaptation of the SMS loses the original intent of the scale in claiming to be an unqualified measure of research quality. Given the absence of any scientific justification, or even reasoned argument for the shift in what the adapted scale claims to measure, the validity of the scale must be considered doubtful at best. None the less, Debidin and Lovbakke (2005) used this 'reconviction scale' to give ratings to a string of studies investigating the effects of offending behaviour programmes. On this dubious basis the findings from 'higher quality' and 'lower quality' are discussed as if the scale had proven utility. Chitty seriously compounds the error: 'To help fellow researchers and correctional stakeholders to understand the quality (and hence value) of the research evidence and following from the work first done by Sherman, this report has proposed a hierarchy of research standards for reconviction studies' (2005: 80). Without a trace of irony, given the scientific validity of the 'reconviction scale', Chitty (2005) then condemns the extant research as 'sub-optimal'.

Leaving the Home Office scale to one side, there are two issues to consider that are genuinely germane to the evaluation of offending behaviour programmes. First, would RCTs really produce definitive evidence with regard to the evaluation of offending behaviour programmes? Second, can

Table 2. The SMS as adapted by the Home Office for reconviction studies (after Friendship et al., 2005)

1.	Reconviction measured for intervention group only.
2.	Comparison of actual and predicted reconviction for intervention group only.
3.	A comparison of the reconviction rates from treatment and unmatched controls.
4.	A comparison of the reconviction rates from treatment and controls matched on theoretically relevant factors.
5.	A comparison of the reconviction rates from treatment and control groups with randomization to group.

we say whether the evidence from RCTs would be significantly different to that produced by quasi-experimental studies in the context of offending behaviour programmes?

Research design and strength of evidence

Farrington et al. made the point that: 'While randomized experiments in principle have the highest internal validity, in practice they are uncommon in criminology and also often have implementation problems' (2002: 17). Farrington and Welsh's (2005) review similarly notes that randomized experiments are relatively infrequent in criminology, particularly so outside the United States. Aside from technical and statistical issues relevant to RCTs (Boruch, 2007), there are practical difficulties with regard to the implementation of an RCT with offending behaviour programmes. These practical issues include passing sentences that would randomly assign offenders to programmes, or giving researchers powers of 'sentence-override' in order to allocate offenders randomly to condition. There are also myriad problems concerning withholding treatment, as would be necessary with an RCT. If treatment is withheld and the offender commits further crimes then the question will invariably be asked whether these offences could have been prevented. Allocation to a non-treatment condition may be detrimental to the individual prisoner, in that not participating in treatment might influence decisions about security classification, release from security and so on. It is not difficult to envisage legal challenges, with associated political ramifications, to any decision to deny treatment to an individual offender.

The implementation of an RCT can pose its own practical problems. Gondolf (2001) notes that in practice randomization can introduce a bias as offenders who might well have gone in different directions are pooled for the sake of the experimental design. Gondolf (2001) also notes the dropout problem in that within an RCT dropouts continue to be part of the treatment condition. However, once offenders drop out of community treatment (which may be a substantial proportion of those commencing) the position is unlikely to remain neutral, with an increased likelihood for dropouts of consequences such as going to prison. As Gondolf states, 'The real world usually does not work the way an experiment does' (2001: 85).

Farrington and Jolliffe (2002) reviewed the conditions necessary for an RCT outcome study of a prison-based therapeutic unit. Farrington and Jolliffe highlighted the logistical and practical issues in running an RCT in prisons to the required standard. For example, prisoners would have to be assessed consistently in the eight prisons feeding into the programme, with several hundred prisoners necessary to conduct the RCT. Further, in a clear example of the tension between internal and external validity, Farrington and Jolliffe suggested that the research would benefit if the length of the intervention were shortened. When an RCT is not possible, Farrington and Jolliffe recommended the use of high-quality quasi-experimental designs: 'The treatment should be evaluated by using matched treated and control groups, by comparing before and after outcomes in each group, or by statistical adjustment (e.g. in a regression equation) for pre-existing differences between groups' (2002: 4).

Alongside difficulties with implementation and practice there are other problems and limitations with RCTs. Gondolf (2004) notes that the introduction of an RCT may disrupt practice, thereby changing the intervention and its evaluation. Hedderman states that 'RCTs do not answer other important questions such as *why* an intervention works or *which parts* have the most effect' (2004: 187, emphases in original). Hedderman makes the further point that well-designed quasi-experimental studies, introducing control over appropriate factors, can reduce significantly the likelihood of a bias through selection effects. Hedderman's comments chime with Gondolf's (2004) view that as more precise questions are asked about explaining treatment so experimental studies shift towards quasi-experimental designs.

However, the critical point with regard to the evaluation of offending behaviour programmes lies in the comparative nature of the evidence produced by quasi-experimental and fully experimental designs.

Quasi-experimental designs

In the absence of randomization, quasi-experimental designs rely on assembling a non-treatment control (or comparison) group using various strategies to control potential group differences. The control and experimental groups may be matched on a case-for-case basis according to key variables related to outcome. The problem with this methodology is that the greater the number of variables to be matched, the more difficult it becomes to find exact matches (for a directly relevant example of this problem see Friendship et al., 2003). Another approach is to introduce control by forming broadly similar treatment and comparison groups at the onset of the evaluation. It is also possible to introduce control over key variables using statistical methods (for an example see Hollin et al., 2004). In all cases, there is the disadvantage that despite attempts to introduce control, procedurally or statistically, the absence of randomization allows the probability of some systematic variation between groups. It follows that any between-group difference in outcome might be a consequence of

any bias introduced by this (unidentified) variation rather than by the intervention.

The designs used by Friendship et al. (2003), Hollin et al. (2004) and Palmer et al. (2007) are referred to by Wilson et al. (2005) as 'high-quality quasi-experimental', in contrast to 'low-quality quasi-experimental', designs. In low-quality designs there are threats to the internal validity of the study by using, for example, non-equivalent treatment and comparison groups, or comparing programme completers with programme dropouts. The problem of comparing programme completers with programme dropouts highlights the issue of bias. The use of programme dropouts as a control group may well introduce a systemic bias as it is possible that offenders who complete a programme have different characteristics to non-completers (Wormith and Olver, 2002). Clearly, any systematic group differences are a potential threat to internal validity when comparing group outcomes. However, as Gondolf (2004) notes, it is entirely realistic to study naturally occurring treatment subgroups. Research that considers what happens in the real world of practice by achieving good *external* validity may well contribute to a greater understanding of programmes and hence meaningfully inform practice.

Thus, there are quandaries associated with the use of quasi-experimental designs to evaluate offending behaviour programmes. Would the evidence produced by RCTs provide a definitive answer, allowing such doubts to be put to one side? This question of design is not peculiar to the evaluation of offending behaviour programmes. In the wider behaviour change and medical literatures, in which RCTs are well practised, there is a great deal of information from which to begin to formulate an answer to this question.

What type of evidence do RCTs produce?

The basic premise underpinning offending behaviour programmes is that they seek to change the dynamic risk factors associated with criminal behaviour thereby reducing the likelihood of offending. In practice, offending behaviour programmes are based on the same principles and methods of behaviour change that would be found in mainstream practice with non-criminal populations. Thus, many of the issues relating to the evaluation of interventions, including offending behaviour programmes, within the criminal justice system have been rehearsed in the parallel clinical literature (e.g. Seligman and Levant, 1998; Clark, 2004; Levant, 2004). However, there is remarkably little cross-referencing across literatures so that when this comparative exercise is undertaken some interesting points arise.

In their overview of clinical trials, Everitt and Wessely make the comment that:

> The simpler the intervention, the easier the trial. The RCT methodology was developed principally for drug interventions, in which both intervention and control can be easily controlled and described. Later, the methodology was adapted for psychological interventions, the principal differences included

> the impossibility of ensuring double blindness, and the difficulties in ensuring treatment fidelity.
>
> (2004: 64)

As Everitt and Wessely note, many psychological interventions—specifically including cognitive-behavioural treatments, typical of many offending behaviour programmes—can be classified as 'complex interventions'. With reference to Medical Research Council (MRC) guidelines, Everitt and Wessely suggest that the development and evaluation of complex interventions should pass through various stages as shown in Table 3. The sophistication of the research planning as seen in the MRC guidelines is singularly absent from the offending behaviour programme literature.

The process of conducting a randomized trial is not straightforward: following Hotopf (2002), Everitt and Wessely (2004) contrast the textbook design of an RCT in clinical practice with 'what happens in the real world'. For example, in an ideal design individuals are randomly allocated to treatment but in reality allocation is often by negotiation; or ideally researchers are blind to allocation but in reality everyone is aware of who sits where.

RCTs are often referred to as if they were a single, uniform design. However, the methodological quality of RCTs can vary considerably (Juni et al., 2001), which is not surprising given the numerous problems with bias and methodology inherent to this design (Lewis and Warlow, 2004). Indeed, the availability of the Jadad scale to rate the quality of reporting of RCTs (Jadad et al., 2001) belies the notion that RCTs are an invariable commodity. Drawing on the healthcare literature, Everitt and Wessely (2004) make the distinction between *exploratory* and *pragmatic* trials. An exploratory trial measures the direct effect of the intervention to test whether, under controlled conditions, the intervention has the intended effect on the target group. A pragmatic trial tests what happens when the treatment is introduced into routine clinical practice. The exploratory trial is clearly a critical part of a treatment evaluation and a necessary antecedent to a pragmatic trial.

Given these methodological and procedural points, what type of evidence is produced by an RCT? The basis of an RCT is that those individuals that are

Table 3. Stages in the development and assessment of complex interventions (after Everitt & Wessely, 2004)

1. *Theory*. The development of the theoretical basis of the intervention.
2. *Modelling*. The development of and understanding of the intervention and its effect using small-scale surveys, focus groups, and observational studies.
3. *Exploratory trial*. Preliminary evidence is gathered in support of the intervention.
4. *Definitive RCT*. A randomized study is conducted.
5. *Long-term implementation*. Can the intervention's effects be replicated over time and in different settings?

randomly allocated either to a condition that receives a treatment (usually referred to as the Experimental or Treatment Group), or to a condition where there is no treatment, or a placebo, or 'treatment as usual' (the Control Group). The outcome for both groups is measured against a common variable, typically reconviction in criminal justice studies. Although there are some complexities with respect to data presentation and statistical analysis (Everitt and Wessely, 2004), the outcome from an RCT is a comparison of the outcome for those initially allocated to the Experimental and Control Groups. It is important to note that the Experimental Group is composed of all those individuals who were *allocated* to treatment, regardless of whether or not they actually receive the treatment.

With an RCT one approach to analysis is called *Intention to Treat* (ITT), 'In which analysis is based on original treatment assignment rather than the treatment actually received' (Everitt and Wessely, 2004: 90). Alternatively, *Treatment Received* (TR) analysis considers what happens 'According to the treatment ultimately received' (Everitt and Wessely, 2004: 90) by those individuals who can be shown to have participated satisfactorily in the treatment. As Sherman notes, these alternatives to analysis are the point at which 'Experimentalists often divide their own ranks' (2003: 11). From a research perspective, ITT is considered the cleanest form of analysis: the formation of subgroups within the randomized conditions violates the principle of randomization and so negates the integrity of the RCT. However, from a practice perspective, employing an ITT allows little to be learned about the effectiveness of a treatment when substantial numbers of the Experimental Group do not comply with the treatment. A poor outcome for those who do not comply with treatment, particularly if this is a significant proportion of those allocated, may nullify positive outcomes for those who do comply. Hollis and Campbell (1999) reported a survey of publications in medical journals where an ITT analysis was used. Noting that ITT is better applied to pragmatic trials, Hollis and Campbell concluded that: 'The intention to treat approach is often inadequately described and inadequately applied' (1999: 674).

In fact, ITT and TR analyses offer answers to different questions. Sherman makes the comment that: 'The ITT principle holds that an RCT can test the effects of trying to get someone to take a treatment and, thus, provides a valid inference about the effect of the attempt, as distinct from the actual treatment received' (2003: 12). Similarly, Gondolf notes that: 'The comparison of an experimental group versus control group, therefore, may tell less about treatment effectiveness and more about the procedures of referring to and retaining men in a certain program' (2004: 610). On the other hand, a TR analysis provides an estimate, albeit with the risk of bias, of the actual effects of the treatment when delivered and received as intended. As Sherman suggests, an ITT analysis offers a test of a policy of offering something; a TR analysis shows what happens when that offer is accepted.

Given that ITT analysis was originally devised for biomedical trials, its use in the evaluation of complex psychological treatments has been referred

to as the 'drug metaphor' (Shapiro et al., 1994). In this light, it is not surprising that Goetghebeur and Loeys (2002) discuss the need to move beyond ITT. Writing from a medical perspective, Goetghebeur and Loeys suggest that: 'The more we seek to tailor possibly dynamic treatments to individual characteristics, encouraged by genetic discoveries, the more imperative it becomes to acknowledge treatment received as an important source of variation in treatment effect' (2002: 89). Goetghebeur and Loeys' (2002) comments are part of a growing trend to question a sole reliance on RCTs in clinical research: as Munro states, 'There is an increasing realisation that the issues that can be dealt with by the RCTs are limited' (2005: 381). This is not to say that RCTs do not provide important evidence, but Munro makes the further point that 'There are important activities in clinical research that are neither randomized nor systematic' (2005: 381). Similarly, Victora et al. (2004), discussing public health research, argue that there are complexities in evaluation which mean that evidence must be gathered using a range of research designs, including but not exclusively limited to RCTs. Gilbody and Whitty (2002) have made essentially the same point with respect to evaluation of mental health services. Overall, the trend in research evaluating complex interventions is to move beyond an unqualified reliance on RCTs, applying a wide range of research designs.

Do RCTs and quasi-experimental studies give different findings?

Whether there is a relationship between methodology and outcome is an empirical question: are the findings from RCTs substantially different from the findings produced by quasi-experimental studies? Heinsman and Shadish compared the findings from randomized and non-randomized experiments reported in four meta-analyses of areas of psychological research: they concluded that if 'Randomized and nonrandomized experiments were equally well designed and executed, they would yield roughly the same effect size' (1996: 162). However, not all experiments, randomized or otherwise, are carried out with identical rigour. Heinsman and Shadish note several features of quasi-experimental studies that increase the probability of a reliable effect size. First, a high level of control is desirable over the extent to which participants are able to self-select into and out of conditions. Second, as might be expected, large pre-treatment differences between groups on important variables can produce large effects at posttest. It follows that the control of important variables related to outcome, either statistically or via group matching, is important. Third, once the study is underway effective strategies to minimize attrition are necessary.

With respect to research in the criminal justice system, Weisburd et al. (2001) considered the effect of research design in studies of crime prevention—ranging across communities, corrections, families, labour markets, policing and schools—taken from the Maryland Report (Sherman

et al., 1997). Using the Scientific Methods Scale, Weisburd et al. coded the 308 studies (of which 15 per cent had a fully randomized design) and compared their outcomes according to experimental design. In contrast to Heinsman and Shadish (1996), Weisburd et al. concluded, 'There is a moderate inverse relationship between the quality of a research design, defined in terms of internal validity, and the outcomes reported in a study' (2001: 64). Weisburd et al. suggest, with due caution, that non-randomized designs may introduce a bias in favour of treatment, although they make the further point that 'Randomized studies may not allow investigators the freedom to carefully explore how treatments or programs influenced their intended subjects' (2001: 66). However, Lum and Yang (2005) have noted that the wider criminological literature contains a mixture of findings regarding the relative effect sizes of experimental and non-experimental studies.

Narrowing the focus to offending behaviour programmes, several reviews have included an empirical analysis of the impact of experimental design on outcome. Lipsey (1992) examined the effect of methodological variables in a meta-analysis of offender treatment studies: the main factor to emerge was the pre-treatment equivalence of the treatment and control groups, such that substantial initial differences between groups were strongly associated with greater differences in outcome following the intervention. However, Lipsey makes the comment that:

> More surprising was the finding that the nature of the subject assignment to groups (random versus nonrandom), often viewed as synonymous with design quality, had little relationship to effect size. What mattered far more was the presence or absence of specific areas of non-equivalence—for example sex differences—whether they occurred in a randomized design or not. (1992: 120)

Lipsey et al. (2001) reported a systematic review of the outcomes of 14 studies of cognitive-behavioural interventions with offenders. Eight studies used randomized allocation to condition, and six studies employed a quasi-experimental design. The six quasi-experimental studies did not use groups that fell out in practice, such as treatment completers and dropouts, and the treatment and control groups were initially equivalent and matched on key variables. The comparison of the randomized and non-randomized studies showed that the non-randomized studies gave a marginally larger treatment effect, but there was no statistical difference in outcome according to design.

Babcock et al. (2004) reported a meta-analysis, including studies using quasi-experimental and experimental designs, of treatment outcome for men who had committed domestic violence. Babcock et al. reported that both designs showed a significant treatment effect, with no difference in effect size according to type of design. A quantitative review of offender treatment programmes reported by Wilson et al. (2005) compared findings from random allocation studies and high-quality studies that used statistical methods to control group differences. Wilson et al. reported that the both

types of high-quality studies, using either random allocation or statistical control, produced broadly similar findings. Lösel and Schmucker (2005) reported a meta-analysis of treatment effectiveness for sex offenders, coding study methodology using the Maryland Scientific Methods Scale. The results of studies using randomized designs did not differ significantly from those using quasi-experimental designs.

Are RCTs always the gold standard?

The issues associated with the use of RCTs to evaluate complex interventions are increasingly being recognized, even in medical research where RCTs are highly prized. Thus, a series of articles in the *Annals of Internal Medicine* took the view that systemic reviews should include high-quality studies, both randomized and non-randomized (Hartling et al., 2005; Norris and Atkins, 2005; Reed et al., 2005). It was advocated that rather than dismissing evidence from quasi-experimental studies, use should be made of guidelines for conducting high-quality quasi-experimental studies (see Des Jarlais et al., 2004). In keeping with this view, several commentators have also made recommendations for the design of quasi-experimental outcome studies of offending behaviour programmes (Lipsey, 1992; Heinsman and Shadish, 1996).

Slade and Priebe (2001) have discussed the importation of RCTs from medicine into mental health service evaluation: in particular, they compare the drug-based therapy of medical interventions with the social and psychological focus typical of mental health interventions. Slade and Priebe take the position that, 'Regarding RCTs as the gold standard in mental health care research results in evidence-based recommendations that are skewed, both in the available evidence and the weight assigned to evidence' (2001: 287). This statement would be true for the evaluation of offending behaviour programmes and stands in direct contrast to the narrow approach advocated within Harper and Chitty (2005) in which RCTs are portrayed as the gold standard and the extant research is dismissed as 'suboptimal' and 'equivocal'. Raynor makes an interesting appraisal of the situation: 'Criminal justice research in Britain has suffered as a result of their [RCTs] rarity, but it would be unwise to put all our heuristic eggs in this one basket' (2004: 319).

Conclusion

In conclusion, there has been considerable investment in offending behaviour programmes within the prison and probation services. The importance of evaluation with regard to this investment cannot be understated in terms of increasing clarity around 'What Works?', improving practice and informing decisions regarding continued investment in offender rehabilitation.

It is therefore essential that evaluation is seen to be conducted with integrity and impartiality and the evidence judged on its merits: spurious diversions about 'gold standard' optimum research designs simply detract from the main issues. The conclusion made by Slade and Priebe with reference to mental health research applies equally well to the evaluation of offending behaviour programmes:

> Mental health research needs to span both the natural and social sciences. Evidence based on RCTs has an important place, but to adopt concepts from only one body of knowledge is to neglect the contribution that other, well-established methodologies can make … RCTs can give better evidence about some contentious research questions, but it is an illusion that the development of increasingly rigorous and sophisticated RCTs will ultimately provide a complete evidence base.
>
> (2001: 287)

If we are serious about producing a strong 'What Works?' evidence base in relation to offending behaviour programmes, then we should learn from the evidence that is currently available rather than relying on illusions of a grand design that will deliver the ultimate truth.

Acknowledgements

I am grateful for the constructive and helpful comments from two anonymous reviewers on an earlier version of this article. I am also indebted to my fellow Pathfinder researchers, Charlotte Bilby, Ruth Hatcher, James McGuire and Emma Palmer, for, among many things, long discussions about the Home Office's views on research design. It was our collective ire that led me to research and write this article, all responsibility for which is mine alone.

References

Babcock, J.C., C.E. Green and C. Robie (2004) 'Does Batterers' Treatment Work? A Meta-Analytic Review of Domestic Violence Treatment', *Clinical Psychology Review* 23(8): 1023–53.

Boruch, R. (2007) 'The Null Hypothesis Is Not Called That for Nothing: Statistical Tests in Randomized Trials', *Journal of Experimental Criminology* 3(1): 1–20.

Chitty, C. (2005) 'The Impact of Corrections on Re-offending: Conclusions and the Way Forward', in G. Harper and C. Chitty (eds) *The Impact of Corrections on Re-offending: A Review of 'What Works'*, 2nd edn, pp. 75–82. Home Office Research Study 291. London: Home Office.

Clancy, A., L. Hudson, M. Maguire, R. Peake, P. Raynor, M. Vanstone and J. Kynch (2006) *Getting Out and Staying Out*. Bristol: Policy Press.

Clark, D.M. (2004) 'Developing New Treatments: On the Interplay between Theories, Experimental Science and Clinical Innovation', *Behaviour Research and Therapy* 42(9): 1089–104.

Cook, T.D. and D.T. Campbell (1979) *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston, MA: Houghton Mifflin Company.

Debidin, M. and J. Lovbakke (2005) 'Offending Behaviour Programmes in Prison and Probation', in G. Harper and C. Chitty (eds) *The Impact of Corrections on Re-offending: A Review of 'What Works'*, 2nd edn, pp. 31–55. Home Office Research Study 291. London: Home Office.

Des Jarlais, D.C., C. Lyles and N. Crepaz (2004) 'Improving the Quality of Nonrandomized Evaluations of Behavioural and Public Health Interventions: The TREND Statement', *American Journal of Public Health* 94(3): 361–6.

Elliott-Marshall, R., M. Ramsey and D. Stewart (2005) 'Alternative Approaches to Integrating Offenders into the Community', in G. Harper and C. Chitty (eds) *The Impact of Corrections on Re-offending: A Review of 'What Works'*, pp. 57–74. Home Office Research Study 291, 2nd edn. London: Home Office.

Everitt, B.S. and S. Wessely (2004) *Clinical Trials in Psychiatry*. Oxford: Oxford University Press.

Farrington, D.P. and D. Jolliffe (2002) *A Feasibility Study into Using a Randomised Controlled Trial to Evaluate Treatment Pilots at HMP Whitemoor*. Home Office Online Report 14/02. London: Home Office.

Farrington, D.P. and B.C. Welsh (2005) 'Randomized Experiments in Criminology: What Have We Learned in the Last Two Decades?', *Journal of Experimental Criminology* 1(1): 9–38.

Farrington, D.P., D.C. Gottfredson, L.W. Sherman and B.C. Welsh (2002) 'The Maryland Scientific Methods Scale', in L.W. Sherman, D.P. Farrington, B.C. Welsh and D.L. MacKenzie (eds) *Evidence-Based Crime Prevention*, pp. 3–21. London: Routledge.

Friendship, C., R. Street, J. Cann and G. Harper (2005) 'Introduction: The Policy Context and Assessing the Evidence', in G. Harper and C. Chitty (eds) *The Impact of Corrections on Re-offending: A Review of 'What Works'*, 2nd edn, pp. 1–16. Home Office Research Study 291. London: Home Office.

Friendship, C., L. Blud, M. Erikson, L. Travers and D.M. Thornton (2003) 'Cognitive-Behavioural Treatment for Imprisoned Offenders: An Evaluation of HM Prison Service's Cognitive Skills Programmes', *Legal and Criminological Psychology* 8(1): 103–14.

Gilbody, S. and P. Whitty (2002) 'Improving the Delivery and Organisation of Mental Health Services: Beyond the Conventional Randomised Controlled Trial', *British Journal of Psychiatry* 180: 13–18.

Goetghebeur, E. and T. Loeys (2002) 'Beyond Intention to Treat', *Epidemiologic Reviews* 24(1): 85–90.

Gondolf, E.W. (2001) 'Limitations of Experimental Evaluation of Batterer Programs', *Trauma, Violence, and Abuse* 2(1): 79–88.

Gondolf, E.W. (2004) 'Evaluating Batterer Counselling Programs: A Difficult Task Showing Some Effects and Implications', *Aggression and Violent Behaviour* 9(6): 605–31.

Harper, G. and C. Chitty (eds) (2005) *The Impact of Corrections on Re-offending: A Review of 'What Works'*, 2nd edn. Home Office Research Study 291. London: Home Office.

Hartling, L., F.A. McAlister, B.H. Rowe, J. Ezekowitz, C. Friesen and T.P. Klassen (2005) 'Challenges in Systematic Reviews of Therapeutic Devices and Procedures', *Annals of Internal Medicine* 142(12 Pt 2): 1100–11.

Hedderman, C. (2004) 'Testing Times: How the Policy and Practice Environment Shaped the Creation of "What Works" Evidence-Base', *Vista* 8(3): 182–8.

Heinsman, D.T. and W.R. Shadish (1996) 'Assignment Methods in Experimentation: When Do Nonrandomized Experiments Approximate Answers from Randomized Experiments?', *Psychological Methods* 1(2): 154–69.

Hollin, C. (2005) 'Offending Behaviour Programmes II: Programme Outcome', paper presented at the XVth European Conference for Psychology and Law, Mykolas Romeris University, Vilnius, Lithuania, June.

Hollin, C., J. McGuire, E. Palmer, C. Bilby, R. Hatcher and A. Holmes (2002) *Introducing Pathfinder Programmes into the Probation Service*. Home Office Findings 177. London: Home Office.

Hollin, C.R., E.J. Palmer, J. McGuire, J. Hounsome, R. Hatcher, C. Bilby and C. Clark (2004) *Pathfinder Programmes in the Probation Service: A Retrospective Analysis*. Home Office Online Report 66/04. London: Home Office.

Hollis, S. and F. Campbell (1999) 'What Is Meant by Intention to Treat Analysis? Survey of Published Randomised Control Trials', *British Medical Journal* 319(7211): 670–4.

Hope, T. (2004) 'Pretend It Works: Evidence and Governance in the Evaluation of the Reducing Burglary Initiative', *Criminal Justice* 4(3): 287–308.

Hotopf, M. (2002) 'The Pragmatic Randomised Control Trial', *Advances in Psychiatric Treatment* 8(5): 326–33.

Jadad, A.R., R.A. Moore, D. Carroll, C. Jenkinson, J.M. Reynolds, D.J. Gavaghon and D.M. McQuay (2001) 'Assessing the Quality of Reports of Randomized Clinical Trials: Is Blinding Necessary?', *Controlled Clinical Trials* 17(1): 1–12.

Juni, P., D.G. Altman and M. Egger (2001) 'Assessing the Quality of Controlled Clinical Trials', *British Medical Journal* 323(7303): 42–6.

Levant, R.F. (2004) 'The Empirically Validated Treatments Movement: A Practitioner/Educator Perspective', *Clinical Psychology: Science and Practice* 11(2): 219–26.

Lewis, S.C. and C.P. Warlow (2004) 'How to Spot Bias and Other Potential Problems in Randomised Controlled Trials', *Journal of Neurology, Neurosurgery and Psychiatry* 75(2): 181–7.

Lewis, S., M. Maguire, P. Raynor, M. Vanstone and J. Vennard (2007) 'What Works in Resettlement? Findings from Seven Pathfinders in England and Wales', *Criminology & Criminal Justice* 7(1): 33–53.

Lewis, S., J. Vennard, M. Maguire, P. Raynor, M. Vanstone, S. Raybould and A. Rix (2003) *The Resettlement of Short-Term Prisoners: An Evaluation of Seven Pathfinder Programmes*. RDS Occasional paper 83. London: Home Office.

Lipsey, M.W. (1992) 'Juvenile Delinquency Treatment: A Meta-Analytic Inquiry into the Variability of Effects', in T.D. Cook, H. Cooper, D.S. Cordray, H. Hartmann, L.V. Hedges, R.J. Light, T.A. Louis and F. Mosteller (eds)

Meta-Analysis for Explanation: A Casebook, pp. 83–127. New York: Russell Sage Foundation.

Lipsey, M.W., G.L. Chapman and N.A. Landenberger (2001) 'Cognitive-Behavioral Programs for Offenders', *Annals of the American Academy of Political and Social Science* 578(1): 144–57.

Lösel, F. and M. Schmucker (2005) 'The Effectiveness of Treatment for Sexual Offenders: A Comprehensive Meta-Analysis', *Journal of Experimental Criminology* 1(1): 117–46.

Lum, C. and S.-M. Yang (2005) 'Why Do Evaluation Researchers in Crime and Justice Choose Non-Experimental Methods?', *Journal of Experimental Criminology* 1(2): 191–213.

McGuire, J. (ed.) (1995) *What Works: Reducing Reoffending*. Chichester: John Wiley & Sons.

McGuire, J. (ed.) (2002) *Offender Rehabilitation and Treatment: Effective Programmes and Policies to Reduce Re-offending*. Chichester: John Wiley & Sons.

Maguire, M. and P. Raynor (2006) 'How the Resettlement of Prisoners Promotes Desistance from Crime: Or Does It?', *Criminology & Criminal Justice* 6(1): 19–38.

Munro, A.J. (2005) 'The Conventional Wisdom and Activities of the Middle Range', *British Journal of Radiology* 78(929): 381–3.

Norris, S.L. and D. Atkins (2005) 'Challenges in Using Nonrandomized Studies in Systematic Reviews of Treatment Interventions', *Annals of Internal Medicine* 142(12 Pt 2): 1112–19.

Nuttall, C. (2003) 'The Home Office and Random Allocation Experiments', *Evaluation Review* 27(3): 267–89.

Palmer, E.J. (2005) 'Offending Behaviour Programmes I: Issues in Evaluation', paper presented at the XVth European Conference for Psychology and Law, Mykolas Romeris University, Vilnius, Lithuania, June.

Palmer, E.J., J. McGuire, J.C. Hounsome, R.M. Hatcher, C.A. Bilby and C.R. Hollin (2007) 'Offending Behaviour Programmes in the Community: The Effects on Reconviction of Three Programmes with Adult Male Offenders', *Legal and Criminological Psychology* 12(2): 251–64.

Raynor, P. (2004) 'The Probation Service "Pathfinders": Finding the Path and Losing the Way?', *Criminal Justice* 4(3): 309–25.

Reed, D., E.G. Price, D.M. Windish, S.M. Wright, A. Gozu, E.B. Hsu, M.C. Beach, D. Kern and E.B. Bass (2005) 'Challenges in Systematic Reviews of Educational Intervention Studies', *Annals of Internal Medicine* 142(12 Pt 2): 1080–9.

Seligman, M.E.P. and R.F. Levant (1998) 'Managed Care Policies Rely on Inadequate Science', *Professional Psychology: Research and Practice* 29(3): 211–12.

Shapiro, D.A., H. Harper, M. Startup, S. Reynolds, D. Bird and A. Suokas (1994) 'The High Water Mark of the Drug Metaphor: A Meta-Analytic Critique of Process-Outcome Research', in R.L. Russell (ed.) *Reassessing Psychotherapy Research*, pp. 1–35. New York: Guilford Press.

Sherman, L.W. (2003) 'Misleading Evidence and Evidence-Led Policy: Making Social Science More Experimental', *Annals of the American Academy of Political and Social Science* 589: 6–19.

Sherman, L.W., D.C. Gottfredson, D.L. MacKenzie, J.E. Eck, P. Reuter and S.D. Bushway (1997) *Preventing Crime: What Works, What Doesn't, What's Promising*. Washington, DC: Department of Justice, National Institute of Justice.

Slade, M. and S. Priebe (2001) 'Are Randomised Controlled Trials the Only Gold That Glitters?', *British Journal of Psychiatry* 179: 286–7.

Stewart-Ong, G., L. Harsent, C. Roberts, R. Burnett and Z. Al-Attar (2004) *What Works: Think First Prospective Research Study: Effectiveness and Reducing Attrition*. London: Home Office.

Victora, C.G., J.-P. Habicht and J. Bryce (2004) 'Evidence-Based Public Health: Moving beyond Randomized Trials', *American Journal of Public Health* 94(3): 400–5.

Weisburd, D., C.M. Lum and A. Petrosino (2001) 'Does Research Design Affect Study Outcomes in Criminal Justice?', *Annals of the American Academy of Political and Social Science* 578(November): 50–70.

Wilson, D.B., L.A. Bouffard and D.L. Mackenzie (2005) 'A Quantitative Review of Structured, Group-Orientated, Cognitive-Behavioural Programs for Offenders', *Criminal Justice and Behavior* 32(2): 172–204.

Wormith, J.S. and M.E. Olver (2002) 'Offender Treatment Attrition and Its Relationship with Risk, Responsivity, and Recidivism', *Criminal Justice and Behavior* 29(4): 447–71.

CLIVE R. HOLLIN is Professor of Criminological Psychology at the University of Leicester. He has published widely on the interplay between psychology and criminology, including the best-selling text *Psychology & Crime: An Introduction to Criminological Psychology*, and he is a co-editor of the journal *Psychology, Crime, and Law*.