Parable of two agencies, one of which randomizes.

Dominic Pearson †
David Torgerson *
Cynthia McDougall *
Roger Bowles *

† County Durham Probation Area
* University of York

Abstract

This article examines the design of evaluations in settings where there is a choice as to how an intervention is to be introduced and evaluated. It uses data from a rehabilitation programme for offenders on probation in the UK (Bruce and Hollin In press) that had been indicated by a pilot evaluation in one probation area to merit wider-scale implementation and evaluation. For the remaining two probation areas in the region, a randomized controlled allocation of participants to conditions was recommended. One of the areas adopted a stepped-wedge design, in which probation offices were randomly allocated sequentially to the programme. The second area opted to launch the programme across the whole area simultaneously, with a retrospective sample as control group. The paper compares the results of implementation in each probation area and seeks to draw wider inferences about the management of programme implementation and the randomized controlled designs appropriate for similar field studies.

Keywords · Programme implementation · Randomized controlled trial · Field study · Community corrections · Offender

Introduction

It is recognised as good practice when planning the implementation of an intervention or programme to give consideration at the beginning of the process to the method of evaluation to be used. Frequently constraints may be imposed on the evaluation by the type of programme to be implemented, the environment in which it is to be

implemented, the types of participant taking part in the study, and the planned outcome measures. Of prime concern in any organisation are the practicalities of applying the evaluation methodology, the additional costs, and the ethical issues associated with the methodology. In some cases, where belief in the implicit value of a programme is strong, the prime concern of an organisation may be to implement the programme as quickly and as cost-effectively as possible, with less importance being placed on the evaluation. These considerations have been frequently raised and have proved to be a barrier to using experiments in the field of criminal justice in the United Kingdom (UK). In an academic forum, the problem was highlighted by Farrington (2003a) who noted that, since the objections raised by Clarke and Cornish (1972), the implementation of experimental designs in the UK has generally been viewed as impractical in criminal justice settings and therefore not feasible to implement. Challenges to experimental designs have also been levelled on grounds that the controlled conditions impair the context and therefore the external validity of an evaluation (e.g. Pawson and Tilley 1998). What is not always recognised is that the method of evaluation may impose a structure on the implementation which might actually be beneficial and improve the effectiveness of the programme. The purpose of this paper is to describe the adoption of such a structured implementation process and to compare the impact of two different designs in the evaluation of the same programme as implemented in two different geographical areas.

Applying research designs to practice

There seems little doubt that the choice of research design selected for an evaluation can have an impact on the results obtained from the evaluation. A review of evaluations of offending behaviour programmes using different designs found that effect size estimates were different depending on the design selected, with weaker designs more likely to find an effect of an intervention (Weisburd, Lum and Petrosino 2001). The 'What Works' evaluation evidence has been helpful in aggregating diverse research studies quantitatively to give a summary of the size of the overall effect of programmes in reducing re-offending (e.g. Andrews et al. 1990; Antonowicz and Ross 1994; Lipsey 1995; Lipsey, Wilson and Cothern 1998; Redondo, Sanchez-Meca, and Garrido 1999). Although these have generally been studies with quasi-experimental designs, the meta-analyses have assisted with hypothesis formation for replication in better controlled designs. Experimental designs however are rarely implemented in routine practice. In the UK criminal justice system, the location of the

current study, Farrington (2003a; McDougall, Perry, and Farrington 2006) identified just fourteen Randomized Controlled Trials (RCTs) published or underway since 1960. A further RCT in British prisons has been completed by McDougall, Perry et al. (2009), the first in recent times. The limited extent to which RCTs have been applied is lamentable not least because such designs may represent an opportunity for best practice in terms of encouraging integration between research and practice in the initial stages of implementation and evaluation design. Furthermore, thinking through the design of an experiment encourages evaluators to address other sources of bias, such as ascertainment bias, by introducing safeguards such as blinded assessment of outcome.

A review of the findings from the British RCTs by Farrington (2003a) showed how experimental research can challenge certainty amongst practitioners and policy-makers about intervention effectiveness by unsettling preconceptions formed on the basis of weaker research. For example after random allocation to one of three conditions Williams (1975) found, contrary to predictions, that the most disturbed subjects did better than the least disturbed in the 'traditional' treatment. Similarly Cornish and Clarke (1975) were surprised to obtain a null finding from their RCT that compared reconviction rates in young offenders exposed to a therapeutic community with those in cases exposed to a traditional institutional regime. Since RCTs equalise as far as possible the chance of systematic differences between groups, both on measured and on unmeasured variables, this offers the best possible opportunity for isolating the effects of the intervention on the key outcome(s). The importance of this may not be understood by most practitioner staff; Farrington (2003b) attributed practitioner opposition to RCTs to the limited extent to which practitioners in criminal justice are trained in research standards. Instead the randomisation procedure is seen as interfering with practitioner decision-making. The practical implications of this as well as ethical concerns about withholding treatment in the absence of an alternative are the chief concerns identified in RCT feasibility reviews (e.g. Campbell 2003; Farrington and Jolliffe, 2002).

On account of the difficulties in implementing a traditional RCT, more common practice has involved analysis of groups 'naturally occurring' in the field setting and have addressed variations between the groups by using matched samples or statistical adjustment. Since in such incidental designs the assignment of cases is not random, groups are systematically different. The effect of this upon the results of outcome evaluations was apparent across the three large-scale evaluations of UK

prison-based offending behaviour programmes (Friendship et al. 2003; Falshaw et al. 2004; Cann et al. 2003). Friendship et al. (2003) found a significant reduction in reconviction rates of 11-14 percentage points associated with the programme participants, relative to matched participants in a control group, by comparing offenders in both groups that had been assessed at a 'medium' risk of reoffending. Although the groups were matched on a number of relevant factors they may have been systematically different on their motivation to address their offending behaviour. Using a similar methodology the replication studies did not find statistically different reconviction rates in the two groups (Cann et al. 2003; Falshaw, et al. 2004). Results became statistically significant however, when the non-completer cases were excluded from the programme referrals group (Cann et al. 2003). The problem with this is that it violates an intention to treat principle and potentially introduces greater selection bias. The importance of systematic factors related to motivation to change, such as programme completion status, on between groups differences illustrates a key limitation of the quasi-experimental design.

A US study by van Voorhis et al. (2004) amplifies this in undertaking an experimental design and a quasi-experimental design on the same data. In the experimental condition parolees had a 60% rate of programme completion. When experimental and control groups were compared on a number of outcome measures, no significant effect was discerned. However when the design was altered to compare three groups (completers, non-completers and controls) a significant programme effect was then found when controlling for variation in risk factors. This points up the importance of randomisation by illustrating the powerful effect of differences on unmeasured variables, such as motivation to change.

Implementation failure due to organisational factors may have been equally responsible for outcome differences within the programme condition in the above quasi-experimental studies rather than simply due to selection bias because of non-random allocation. This was impossible to discount since the research design was not able to establish what would have happened to the non-completers in the absence of the programme. This was the conclusion of the UK Ministry of Justice's own researchers (Debidin and Lovbakke 2005). Debidin and Lovbakke (2005) cited qualitative research (Clarke, Simmonds, and Wydall 2004) highlighting problems in institutional support for programmes, as well as issues relating to offender motivation due to long waiting lists and the timing of programmes within sentences. In the van Voorhis et al. (2004) study, the completion rates varied considerably across the 16

parole districts, from a low of 42% to a high of 80%, suggesting variable organisational performance. Similar variability in completion rates across districts has been reported within a UK probation area (Briggs and Turner 2003). Van Voorhis et al. (2004) described the non-completers as being younger on average, statistically more likely than the completers to have a violent previous conviction and statistically less likely to have completed high school. As acknowledged by van Voorhis et al., young age, low education and aggression history are all known to predispose for poor organisational outcomes. It is not clear what was the occurrence of such outcomes in similar offenders randomized to the control group.

Inability to rule out the adverse impact of implementation failure and/or selection effects clearly limits the conclusions that can be drawn from quasi-experimental designs. However given the disruption that can be caused by experimental designs there is a question over whether the results of such evaluations are generalisable to routine practice where the allocation of participants to conditions is no longer controlled in the same way. One view holds that practical interventions evaluated by non-equivalent designs may have higher external validity and are potentially more informative concerning the application of findings in everyday practice (e.g. Lipsey 1999a; Pawson and Tilley 1998). This presents a challenge to research and practice as to finding the means to test hypotheses experimentally but in a way in which the results can be generalised. Since RCTs require greater researcher involvement in terms of monitoring practitioners' adherence to random assignment of cases, they may be associated with better implementation. In the 'What Works' evidence, strong implementation where all cases receive the intended treatment, has been associated with reductions in re-offending compared to results based on incomplete implementation (e.g. Lipsey 1999b).

Overcoming barriers to implementation in experimental designs

As alluded to above, practitioners and managers in the criminal justice agencies represent the key barriers to the use of experimental designs in routine practice. McDougall, Clarbour et al. (2009) reviewed problems previously associated with RCTs and sought to address them within their experimental design. They reported that these generally fall into three categories: ethical, practical and statistical. Ethical concerns surround the presumed negative effect on an individual by withholding (or giving) treatment intervention on the basis of random allocation. This view favours quasi-experimental designs since they do not involve randomisation. McDougall,

Clarbour et al. (2009) addressed this by employing a waiting list control group. This had the advantage that all individuals eventually received the intervention. The disadvantage of this was that outcomes for comparison were limited by the waiting time of the controls and that long-term follow up was not possible as part of the RCT since all of the groups would be treated. In the absence of evidence that the programme works, a valid but less popular ethical view is that treatment and non-treatment groups should be equally and randomly distributed. This is especially at issue in offender groups where there may be inherent risks that the programme may have a negative impact on outcomes (e.g. Sherman et al., 1997). Practical issues reviewed by McDougall, Clarbour et al. (2009) concern the occasional operational need to prioritise certain offenders for intervention due to factors such as an imminent release from custody. This was addressed by allowing such practice to continue but allocating the case to a 'cohort' group to be analysed separately. Statistical issues concern the sample size required to detect an effect of the intervention. Farrington and Jolliffe (2002) identified that the required size of the control group decreases dramatically as the size of the intervention group increased. McDougall, Clarbour et al. (2009) addressed this by ensuring that the establishments selected had a sufficient number of intervention groups in operation. It was recognised that problems in cooperation with randomization of participants, e.g. where the allocated offenders are not deemed suitable for immediate intervention (Farrington and Jolliffe 2002), can represent a fatal barrier to the feasibility of an experimental design due to its potential to diminish statistical power.

Implementing a novel probation supervision programme

The current study examines the implementation of a supervision programme in each of three probation areas in North-East England (one of which had already implemented the programme). Programme implementation involved training of probation officers and support staff, delivery of the programme with offenders, and instituting the monitoring arrangements to ensure programme integrity. The strategy for implementation and evaluation selected in each area was different, owing to specific local constraints. Area A was the probation area in which the programme had been developed. A pilot evaluation was conducted before the programme was fully implemented, and, based on these results, the programme was then launched area-wide prior to the commissioning of a large-scale evaluation. This meant that Area A was unable to opt for a randomized experimental design, as all offenders were already being offered the intervention. The position in the other two probation

areas was different however; they were yet to take on the programme and this offered the opportunity of a choice as to the optimum fit between research design and implementation strategy.

Both remaining areas (Area B and Area C), initially had the same reservations as previously encountered regarding the professional ethics of withholding an intervention from an offender for the sake of a research study. The programme was designed to be delivered as part of a court imposed 'supervision' requirement and provided a framework to intervene with offenders on those 'need' areas that were seen as responsible for the offending behaviour. Denying this facility to some offenders but not others was seen as unjust by a number of practitioner staff. A very reasonable point was also made as to how random allocation of the programme would be managed in the courts process.

In addition to the ethical reservations we encountered, a number of objections to randomisation were raised on the basis of operational management. Random assignment would mean that offender managers (those supervising offenders) would be required to deliver supervision in different ways to different offenders depending on their assignment to the intervention or to the control group. This would present logisitical difficulties in terms of managing cross-over in practice between cases assigned to different groups. In addition to being a threat to construct validity, fidelity of delivery - ensuring that the intervention offered is the intervention received - is seen as a key mediator of successful interventions (Andrews et al., 1990). This is an issue for programme integrity managers tasked with ensuring that cases receive the intended treatment (Hollin 1995). Monitoring by managers is made easier if the manager is clear that one system prevails in the office. The operation of different systems simultaneously also presents a dilemma for staff training: would practitioners be expected to employ new skills and awareness with some offenders and not with others, or alternatively should the randomization be at the practitioner level so that all individual practitioners would be following a single system? This would then require a degree of non-random assignment since, when allocating work, managers often need to match offenders to staff according to diversity (e.g. gender) or geographical factors (e.g. offender home location).

Solutions adopted

The difficulties discussed above persuaded us that it was not going to be manageable to randomize individual offenders within a single probation team. We concluded that the choice for the two areas was between a pre/post quasi-experimental design involving a single area-wide launch, and an experimental design which randomized to clusters of offenders at the level of probation office (a 'stepped wedge' design, described in the method section below). This meant that the experimental design would require a series of staggered launches, whereas the pre/post design would require no more than one large launch process.

The option of a pre/post design carried a number of practical advantages in terms of consistency in training with a brief/intensive training schedule, consistency in monitoring and consistency in delivery. It was possible that a single launch would have greater impact than a staggered launch and might therefore be better in terms of senior leadership and momentum. For this reason the pre/post design was affectionately referred to as the "big bang" approach. From an evaluation perspective, however, such an approach would not allow proper control for temporal changes and consequently any differences before and after the programme introduction could be confounded.

The staggered approach conversely would require a series of smaller training events and the concept of gradual expansion. This might be advantageous in terms of learning from experience within the area. It would also mean that the units for integrity monitoring of programme delivery were smaller. This was important to prevent contamination between units – a risk inherent to the staggered implementation approach. All cases assigned to the intervention group would initially be analysed within that group ('intention to treat' analysis). Given that programme implementation in the community is not expected to be 100%, Complier Average Causal Effect (CACE) analysis might then be required (Hewitt, Torgerson, and Miles 2006). This uses the randomisation as an instrumental variable to assess the impact of the intervention among those who received it.

These different advantages meant that from a programme management perspective one was unsure as to which approach would be more successful. We were however unequivocally clear about our preferred approach from a research perspective, and recommended the staggered implementation approach as this would allow us to control for temporal changes that the previous approach would not.

The senior management teams of the two probation areas were working within different sets of operational challenges at the time a decision was required. Area B's internal monitoring had uncovered disparities between offices in performance. Area C meanwhile was confronting various issues relating to resourcing and staff workloads. There was a perception by Area C that a single launch event ("big bang") would be easier and less costly to launch than a series of smaller events. Area B was however more persuaded by the benefits of a phased roll-out so that the offices that were least affected by performance difficulties would not be influenced by the other units. Hence Area B opted to randomize while Area C did not.

Ahead of programme implementation, three main questions emerged. First, which approach would prove to be more successful? Success here would be defined by the extent to which supervising offender managers used the programme with offender cases starting supervision. Second, what would prove to be the cost implications of the two approaches? When considered alongside the effectiveness in implementation (question one) this would identify the relative cost-benefits of each implementation approach. Finally, it would illustrate to what extent the key benefits of the more successful approach were inextricably linked to that approach, and to what extent they could be incorporated in the alternative methodology.

Method

Description of the programme implemented

The Citizenship programme (Citizenship) is based on the assessment of crime-related need driving structured cognitive-behavioural intervention with individual offenders. Citizenship was designed in Area A by a working group of practitioners under the supervision of an academic consultant (Bruce and Hollin In press). The resulting programme was designed to be consistent with the principles of effective practice and accessible to a wider range of offenders than are 'accredited' programmes. As such Citizenship is targeted at all medium and high risk offenders and the programme is able to respond to a wide range of crime-related needs. The modular nature of the programme and its links to supporting external agencies, is illustrated in Figure 1.

Figure 1 about here

A pilot evaluation of Citizenship was conducted in 2006 in Area A. This examined the first 100 cases starting Citizenship and compared these cases to a matched sample of 100 finishing supervision before Citizenship was introduced. The pilot study showed encouraging results in terms of reduced reconviction and improved contact with agencies compared to the prior practice sample. However, the reduction in reconviction was not statistically significant, which may have been due to the small sample size (a Type II error) or alternatively, no real difference actually existed. Nevertheless, this pilot study formed a key plank in a bid for larger scale evaluation of Citizenship. Funding was granted, contingent on implementation and evaluation of Citizenship in all three areas of the North-East Region. Strategies (research designs) selected by each area for implementation are presented below.

Participants

Participants in the current study were adult offenders starting community supervision in three probation areas in North-East England. Community supervision is a requirement of post-release licences and some community penalties. A variety of offenders are therefore subject to community supervision, ranging from public disorder and theft offences, to offences of robbery and serious violence. Minor offenders who did not have an official requirement to report for community supervision were not targeted for Citizenship and were therefore not included in the research.

A requirement of supervision in the community is generally only court-ordered in cases where the risk of reconviction is deemed 'medium' or higher ('Tier' 2-4). Area A and Area C both chose to provide Citizenship supervision to all supervision cases. Area B, however, opted to target Citizenship only at higher risk offenders ('Tier' 3-4). This meant that the target group in Area B was a sub-set of that in the other two probation areas.

Design/Procedure – Area A

Area A was the operational area in which the programme was developed. The programme was implemented area-wide prior to the large-scale evaluation. The fact

that programme delivery staff had all already been trained in the programme meant that it was not possible to conduct a randomisation of the intervention (due to possible contamination). The method selected in Area A was therefore a quasi-experimental design in which the effect of the intervention was examined against outcomes in cases that had received traditional treatment prior to implementation (prior practice). In Area A, as in the other two Areas, data on implementation were collected as part of programme evaluation. Cases were allocated to the intervention group and were subject to Citizenship implementation, if their supervision commenced between 1st August 2005 and 1st August 2007 in Area A.

Design/Procedure – Area B

Area B opted for the staggered launch, i.e., a randomized 'stepped wedge' research design. In a stepped wedge design an intervention is rolled-out in sequence to the participants, either as individuals or as clusters of individuals (see Brown and Lilford 2006). This happens over a number of time periods. The order in which the intervention is rolled-out to the different individuals or clusters is determined at random. By the end of the randomization all units will have been allocated to receive the intervention at some point in the study period. Due to the difficulties described above in terms of possible contamination between intervention and control participants where the experimental intervention is delivered at the same site as the control intervention, Area B opted to randomize to participants in clusters. Each probation office was a cluster in the stepped wedge.

In a stepped wedge design, data is collected from all clusters before and after the point where a new cluster receives the intervention. Figure 2 illustrates a stepped wedge design with six steps. Data analysis to determine the overall effectiveness of the intervention subsequently involves comparison of the data points in the control section of the wedge with those in the intervention section.

Figure 2 about here

For operational reasons Area B required that offices were paired so that wherever they occurred in the randomisation they were accompanied by their pair office. Therefore randomisation to each step in the wedge was completed with a coin toss

performed 3 times to allocate the 6 steps in the stepped wedge. This was performed in a management group meeting consisting of the evaluation team and the key stakeholders to avoid any issue of allocation compromise.

Table 1 shows the time periods for implementation in each probation office in Area B. There was a minimum of a 2 month interlude between launches in each office, to allow time for new cases in the intervention step(s) to undergo Citizenship.

Table 1 about here

For individual offenders beginning sentences during the given time-period, it meant that the availability of the Citizenship programme depended on whether his/her office had been randomly selected at the time of commencement of their supervision. However, all offices were scheduled to receive Citizenship meaning that some offenders would receive the programme after an interlude. The drawback to this is that because all cases should receive the programme by the end of the roll-out, it is not possible to compare outcomes with a comparison group once step 6 has started nor is it possible to examine the longer-term effects of the programme on offenders.

Our sample size was constrained by the number of offices in Area B. Ideally, we would have preferred to have a greater number of offices to randomize. However, we felt that using a stepped wedge approach maximised our statistical power for the number of offices available. Implementation data were collected as part of the evaluation and collated by the evaluation team (not staff employed by Area B).

Design/Procedure – Area C

Area C selected the "big bang", or single area-wide launch process, as this was seen to be more efficient than a series of incremental launches. This necessarily implied the need for a retrospective comparison sample, and a quasi-experimental design. All cases receiving supervision in Area C during the implementation period, 1st April 2007 to 1st April 2008, were allocated to the Citizenship intervention group. Cases commencing community supervision in the previous year, were allocated to the control group. This is illustrated in Figure 3, for comparison with Figure 2 (the

stepped wedge design). Implementation figures were provided by Area C's performance team.

Figure 3 about here

Results

Results of programme implementation in each Area are presented below under three headings: i) take-up rates; ii) implementation costs; and iii) relationship between key factors and the methodology selected.

Take-up rates

The overall use of the programme post-implementation in all three areas of the region is shown in Table 2 below. Area A, the area responsible for designing Citizenship, used a "big bang" or blanket launch as their strategy for implementation. Table 2 shows a good level of use of the programme by offender managers in Area A. The programme is targeted at 'tiers' 2, 3 and 4; in this group of offenders the programme was implemented in Area A with approximately 75% of cases. Since Area A was not presented with a choice over implementation design the focus of comparison is on Areas B and C.

Table 2 about here

Area B was the probation area that chose to implement the programme to offices in sequence according to a random assignment in which offices were randomly allocated to steps in the 'stepped wedge' (see Figure 2). Table 2 gives the overall results of implementation in terms of the use of the programme by officers post-enrolment into the experimental section of the wedge. In Area B the programme was targeted at offenders of 'tier' 3 and 4. Such offenders are a sub-set of those targeted by Areas A and C, but they represent the more troublesome group as by definition they have a higher level of crime-related need than lower tier offenders. Table 2 shows that the programme was used by practitioners in approximately 44% of these cases.

The extent to which this take-up varied by 'step' (office) is shown in Table 3. Area B conducted file inspections of a sample of cases in each office at the end of the

introduction of each office into the experimental part of the stepped wedge. These spot-checks found higher rates of take-up, ranging from 63% to 79%. Indeed in many offices there was evidence of under-use of the contact codes for recording evidence of the programme session, suggesting that the figures are if anything an under-representation of the true rate of programme implementation. This may also have been true for the take-up rates shown for the other areas.

Table 3 about here

Area C was the second area with a choice as to how to implement to meet the needs of both practice and evaluation. Area C chose to implement at once, that is in a similar approach to that used by Area A, with a single "big bang" inauguration process (see Figure 3). Here the programme was targeted at offenders at tiers 2, 3, and 4, again as in Area A. The programme was implemented with just over one-quarter of these offenders (27.6%). Area C's internal monitoring led the area to believe that some teams were not engaging fully with the programme. Issues were identified relating to workload / resourcing, as well as related issues in understanding how and with whom the programme should be applied. The Area therefore took the decision to re-launch the programme – a second "big bang".

Implementation costs

One cost consideration is the frequency of training events. A total of eighteen training events were run in Area A to ensure that new staff as well as those that may have missed earlier training, were all fully equipped to run the programme with their cases. In Area A training events were scheduled to take place when enough new participants had gathered to warrant a new session. This method was therefore responsive but not timely for all participants.

Unlike Area A, only nine training events were required by Area B. These doubled as mini- launch events and allowed further opportunities for those offender managers that were required to use the programme but had missed the training for their own office. In order that individual offices could learn directly from the designing area, a practitioner was selected from each office to be trained as a trainer by Area A practitioners. Training delivery was always by practitioners from the office to be trained. This was therefore a model whereby each office had a trained 'champion'.

Area B found this to be an efficient model for the transfer of learning from Area A in how to deliver the intervention. In terms of programme training this therefore seemed to be a convenient model that kept costs to a minimum. Area B also undertook monitoring, including a 'dip-sample' inspection two months after each office had been trained in the programme. A larger scale audit was also done following the area-wide roll-out of the programme.

Area C trained all of their staff also in nine training events all between February and March 2007. This is the same number of training events as required by Area B. This intensive process followed Area C's desire for efficiency in programme implementation (attempting to train all staff in a short period of time). All staff were reported as having attended the sessions provided. This of course carries opportunity costs as well as capital costs due to the inevitable impact of withholding such a large proportion of operational staff during the training period. Since then Area C have conducted internal monitoring on a monthly basis, focussing on the extent to which the programme was delivered by the various teams, and also on the integrity of delivery of the various components of the programme. This led to an area-wide relaunch in April 2008, mid-way through the evaluation window, with consequences for costs as well as for the evaluation. The relaunch involved raising the programme on teams' agenda's through meetings as well as a coordinated leafleting campaign.

Relationship between key factors and selected methodology

Area A's results can be more attributed to commitment to programme implementation at senior management level, including a number of file inspections, rather than the implementation method adopted. An area-wide launch process was also used in Area C as they saw this as the most efficient implementation design. An attractive feature of this approach for their senior management was the fact that, in theory, there was only need to attend a minimum number of briefing sessions and this therefore represented least cost. Also from a leadership perspective dilution of messages about the importance of the programme is kept to a minimum. Neither of these assumptions is necessarily correct since there are inevitably a number of staff that are unable to attend the first launch events, meaning that there may be need for 'after-shocks' or further launches to ensure all practitioners are on-board. This appeared to be the case in Area C where an entire re-launch was required due to

misunderstandings about the programme. This was also seen previously in Area A as evident from the number of training events that were scheduled.

A planned sequence of training events was a key benefit for Area B and avoided them having to attempt to train all staff at one time. The use of a phased method of staff training is not restricted to the stepped wedge implementation design. Staggered implementation was however one of the key features of the methodology selected by Area B and required the area to take this approach to training. This facilitated programme championing at site level, in a way not naturally facilitated by a "big bang" approach to training.

Area B was also able to implement a schedule of monitoring to correspond to the staggered roll-out of the programme. This allowed the performance in different delivery units to be compared. This can be seen as an advantage specifically associated with the selected methodology. It also allows corrections, where necessary, to be made to practice during the relevant step in the wedge thereby enhancing delivery and evaluation.

Discussion

The implementation of offending behaviour programmes in the community is known to be problematic and to affect the results of effectiveness evaluations (Andrews et al. 1990; Dowden and Andrews, 2004; Lipsey 1999b). At the same time there is concern that weaker evaluation designs, for example those not using a randomized experimental approach to the assignment of cases, may produce different results or find inflated effects compared to randomized experimental designs (e.g. Farrington 2003b). Consequently there is much interest in the application of randomized experimental designs in actual delivery settings. The current paper aimed to report on one such application in the North-East of England, where it occurred in a context of financial austerity. Two probation areas were faced with a choice of implementation designs. Area A did not have a choice to make; they had already implemented the programme at the time a large-scale evaluation was being considered. Their implementation results have been provided as a context. The main interest of this paper was in the results of implementation pertaining to the other

two Areas, ‘B’ and ‘C’, where the option of a randomized experiment was considered against blanket implementation to all cases.

One of the main concerns of Area B related to maintaining its standards on a number of measures used by the UK Ministry of Justice to manage national performance. Area B was aware of performance differences between its offices and was therefore persuaded by the suggestion of a randomized ‘stepped wedge’ experimental design. This was thought to be of benefit on two fronts. First, it meant that the organisation, across the board, would not be impaired by a widescale implementation regime at the same time that it was attempting to maintain its performance. A phased introduction would be somewhat easier to absorb. Second, Area B had listened to and accepted the arguments in favour of a randomized experimental approach in terms of better quality of the eventual evaluation evidence. This highlights the importance of close working between staff with research skills and those operational staff whose focus is on day-to-day practice.

The same information provided to Area B to help them make a decision was also given to Area C. Against best recommendations however Area C decided upon blanket implementation. This decision was taken in the interests of resources and workloads; area-wide implementation was seen to be easier and less costly than the staggered approach that was suggested as an alternative. The current paper therefore sought to answer whether the implementation approaches selected produced the expected results for the organisations. Did Area B find that a phased approach to implementation was successful in expanding the use of the programme across different offices? And did Area C indeed find that a “big bang” approach to implementation was effective in terms of take-up in practice across its geographical area? Was this as thought by Area C, easier and less costly than staggering programme implementation?

The results provided a salutory lesson to operational managers and policy-makers responsible for the implementation of a new programme where the outcomes are un-tested or in need of replication. While Area B targeted their intervention at more troublesome cases, their use of a stepped implementation design produced take-up rates better than those seen in Area C where the cases targeted also included offenders considered lower risk and easier-to-reach. Area B achieved rates of approximately 44% while Area C only implemented with approximately 28% of cases. As a result of regular internal monitoring Area C took the decision to re-launch the

implementation of their programme. This undermined the idea that a single "big bang" or blanket implementation effort can carry sufficient momentum to reach all operational staff. This was clearly not the case as staff were unsure about how the programme should be applied. Area C's performance and information team should be commended for detecting the problem through their regime of regular internal monitoring.

Workloads and resources were important considerations to both areas. In a previous period Area A, where the programme was developed, undertook a number of thorough and costly internal inspections to follow up the many training events that had been held to ensure that all staff were fully equipped to run the programme. The good implementation results for Area A (75% uptake) reflect the level of investment made, as well as the fact that the take-up rates reflect a longer amount of time elapsed to allow them to routinize delivery of the programme. Area B however achieved reasonable take-up rates with the harder-to-reach offenders on the back of just 9 small scale training sessions and a similar number of inspections over the course of a year. One of the biggest advantages seen by Area B that was absent in Area C during the first year of implementation, was product championing at office level to enable the programme to embed properly. This was made necessary by the type of implementation design adopted. The stepped wedge design meant that Area B was obliged to train each office independently to avoid contamination between experimental and control sections of the wedge. This method of gradually mainstreaming practice was seen as highly beneficial by Area B. Growing confidence in Area B was off-set against mounting confusion amongst practitioners in Area C where the implementation strategy was immediate rather than gradual. Ironically, given that Area C was targeting a greater number of offenders than was Area B, it may have particularly benefited from the use of a staggered implementation strategy.

The results of programme implementation in the two areas have a number of implications. Area B will want to continue to embed the programme within routine practice and will need to continue in monitoring and auditing how the programme is being delivered. Area B will also await the results of the evaluation comparing delivery of the intervention with the old standard practice, within and between randomly selected offices. Area C meanwhile will need to review the success of their re-launch and re-consider the options for evaluation in their area. Since programme implementation is often cited as an indicator of effectiveness (e.g. Andrews et al.

1990), it may be better to change the evaluation window in order to compare more distinctly the intervention programme with the old standard practice.

From a cost-benefit perspective the findings are perhaps best thought of by considering the various possible experimental designs, and identifying the strengths and weaknesses of each. For the sake of argument let us define a fourth Area D where a fully randomized design might have been used. For policy purposes in the future, the critical issue is what can be inferred about the relative merits of: a stepped wedge cluster design (B), a blanket pre/post approach (A or C) and an individually randomized blanket approach (D).

There are clearly strengths and weaknesses associated with all three designs. Since the terms of the trade-off between them are likely to vary with the setting, a set of criteria are needed by which choices might be made. From the findings from the experience in North-East England outlined above, the key criteria emerging can be summarised as:

Statistical integrity: the reliability of estimates of effect size, which will in turn depend on completion rates, selection issues, sample size and extent of randomisation.

Evaluation duration, cost and risk of delay: the time likely to elapse from the beginning to the end of the evaluation phase, the risks of delay associated with the various options, and the training, launch and familiarisation costs of each option.

Managerial issues: variation in the degree of managerial commitment across sites, in the degree to which randomisation is feasible with given staffing levels and in the degree to which the experimental outcome might influence further decisions about provision.

The superior performance of option B in the event resulted from the superior statistical properties of the design (relative to design C) and the managerial impracticality of design D. This serves to establish approaches of design type B as well worth exploring. More generally, however, the implication is that the design of such policy experiments and evaluations need take account of managerial concerns and may not turn exclusively on issues of statistical reliability.

The implications of the present paper extend to a variety of other arenas and their field settings. For one it contributes evidence on the feasibility of randomized controlled trials. The benefits to hypothesis testing are often considered by operational managers to be outweighed by the attendant practical and ethical difficulties (see Farrington and Jolliffe 2002). Not only was the approach taken in Area B seen to be ethical, its practical benefits in terms of mainstreaming a new intervention while causing minimal disruption to general performance impressed Area B and its neighbours. The 'stepped wedge' design may therefore offer a best means of upholding internal validity while implementing a programme in a way that the results can be generalised rather than seen as a relic of the controlled procedure whose conditions are quite unlike those that characterise routine delivery (e.g. Lipsey 1999a).

We look forward to reporting the results of the three individual outcome evaluations, but in the mean-time we hope that when faced with a choice as to the best design for implementation and evaluation, policy makers, operational managers, researchers and practitioners everywhere remember the parable of the two agencies where the one that randomized did so with minimum disruption and were ready to evaluate on time, while the one that did not randomize had to start their implementation again.

References

Andrews, Don A., Ivan Zinger, Robert D. Hoge, James Bonta, Paul Gendreau, and Francis T. Cullen. 1990. Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology 28,* 3: 369-404.

Antonowicz, Daniel H., and Robert R. Ross. 1994. Essential components of successful rehabilitation programmes for offenders. *International Journal of Offender Therapy and Comparative Criminology* 38:97-104.

Briggs, Sarah and Russell Turner. 2003. *Barriers to starting programmes: Second phase report.* Research report, National Probation Service, West Yorkshire.

Brown, Celia A. and Richard J. Lilford. 2006. The stepped wedge trial design: A systematic review. *BMC Medical Research Methodology* 6:54-62.

Bruce, Russell, & Clive R. Hollin. In press. Developing Citizenship. Submitted to *Vista*, February 2009.

Campbell, Siobhan. 2003. The feasibility of conducting an RCT at HMP Grendon. In Home Office Online Report. [cited 27 May 2009]. Available from http://rds.homeoffice.gov.uk/rds/pdfs2/rdsolr0303.pdf.

Cann, Jenny, Louise Falshaw, Francis Nugent, and Caroline Friendship. 2003. Understanding what works: Accredited cognitive skills programmes for adult men and young offenders. *Home Office Research Findings 226*. London: Home Office.

Clarke, Ronald V.G. and Derek B. Cornish. 1972. The controlled trial in institutional research: Paradigm or pitfall for penal evaluators? *Home Office Research Study 16*. London: Home Office.

Clarke, Alan, Rosemary Simmonds, and Sarah Wydall. 2004. Delivering cognitive skills programmes in prison: A qualitative study. *Home Office Research Findings, 242.* London: Home Office.

Cornish, Derek B., and Ronald V.G. Clark. 1975. Residential treatment and its effects on delinquency. *Home Office Research Study* 32. London: Home Office.

Debidin, Mia and Jorgen Lovbakke. 2005. Offending behaviour programmes in prison and probation. In *The impact of corrections on reoffending: A review of 'what works'*, edited by G. Harper and C. Chitty, *291* (2): 31-54. Home Office Research Study. London: Home Office.

Dowden, Craig and Don A. Andrews. 2004. The importance of staff practice in delivering effective correctional treatment: A meta-analytic review of core correctional practice. *International Journal of Offender Therapy and Comparative Criminology* 48:180-187.

Falshaw, Louise, Caroline Friendship, Rosie Travers and Francis Nugent. 2004. Searching for "what works": HM Prison Service accredited cognitive skills programmes. *British Journal of Forensic Practice* 6:3-13.

Farrington, David P. 2003a. British randomised experiments on crime and justice. *ANNALS AAPSS* 589:150-167.

Farrington, David P. 2003b. A short history of randomised experiments in criminology: A meagre feast. *Evaluation Review 27* (3): 218-227.

Farrington, David P. and Darrick Jolliffe. 2002. A feasibility study into using a randomised. London: Home Office Online Report 14/02. [cited 28 May 2009]. Available from www.homeoffice.gov.uk/rds/pdfs2/rdsolr1402.pdf.

Friendship, Caroline, Linda Blud, Matthew Erikson, Rosie Travers and David Thornton. 2003. Cognitive-behavioural treatment for imprisoned offenders: An evaluation of HM Prison Service's cognitive skills programmes. *Legal and Criminological Psychology* 8:103-114.

Hewitt, Catherine J., David J. Torgerson, and Jeremy N. V. Miles. 2006. Taking account of non-compliance in randomised trials. *Canadian Medical Association Journal* 175:347-348.

Hollin, Clive R. 1995. The meaning and implications of programme integrity. In J. McGuire (Ed.) *What works: Reducing reoffending: Guidelines from research and practice*. Chichester: Wiley.

Lipsey, Mark W. 1995. What do we learn from 400 research studies on the effectiveness of treatment with juvenile delinquents? In *What works: Reducing reoffending: Guidelines from research and practice,* edited by James McGuire, 63-78. Chichester: Wiley.

Lipsey, Mark W. 1999a. Can rehabilitative programs reduce the recidivism of juvenile offenders? An enquiry into the effectiveness of practical programs. *Virginia Journal of Social Policy & the Law 6* (3): 610-641.

Lipsey, Mark W. 1999b. Can intervention rehabilitate serious delinquents? *ANNALS AAPSS* 564:142-166.

Lipsey, Mark W., David B. Wilson and Lynn Cothern. 1998. Effective intervention for serious juvenile offenders: A synthesis of research. In *Serious and Violent Juvenile Offenders: Risk Factors and Successful Interventions,* edited by Ralph Loeber and David P. Farrington. Thousand Oaks, CA.: Sage Publications. In National Institute of Justice [database online]. [cited 28 May 2009]. Available from http:/www.ncjrs.gov/pdffiles1/ojjdp/181201.pdf

McDougall, Cynthia, Jane Clarbour, Amanda E. Perry, Roger Bowles and Gillian Worthy. 2009. Evaluation of HM Prison Service Enhanced Thinking Skills Programme: Report on the implementation of a randomised controlled trial. *Ministry of Justice Research Series 4/09.*

McDougall, Cynthia, Amanda E. Perry and David P. Farrington. 2006. Reducing crime. In *Reducing crime: The effectiveness of criminal justice interventions,* edited by Amanda E. Perry, Cynthia McDougall and David P. Farrington, 1-11. Chichester: Wiley.

McDougall, Cynthia, Amanda E. Perry, Jane Clarbour, Roger Bowles and Gillian Worthy. 2009. Evaluation of HM Prison Service Enhanced Thinking Skills Programme: Report on the outcomes from a randomised controlled design. *Ministry of Justice Research Series 3/09.*

Pawson, Ray and Nick Tilley. 1998. Caring communities, paradigm polemics, design debates. *Evaluation* 4:73-90.

Redondo, Santiago, Julio Sanchez-Meca, and Vicente Garrido. 1999. The influence of treatment programmes on the recidivism of juvenile and adult offenders: A European meta-analytic review. *Psychology, Crime and Law* 5:251-278.

Sherman, Lawrence W., Denise Gottfredson, Doris MacKenzie, John Eck, Peter Reuler and Shawn Bushay. 1997. *Preventing crime: What works, what doesn't, what's promising: A report to the United States Congress.* In National Institute of Justice [database online]. [cited 28 May 2009]. Available from http://www.ncjrs.gov/pdffiles/171676.pdf.

van Voorhis, Patricia, Lisa M. Spruance, P. Neal Ritchey, Shelley J. Listwan and Renita Seabrook. 2004. The Georgia cognitive skills experiment: A replication of Reasoning and Rehabilitation. *Criminal Justice and Behavior* 31:282-305.

Weisburd, David, Cynthia M. Lum, and Anthony Petrosino. 2001. Does research design affect study outcomes in criminal justice? *ANNALS AAPSS* 578:50-70.

Williams, Mark. 1975. Aspects of the psychology of imprisonment. In *The use of imprisonment,* edited by S. McConville, 32-42. London: Routledge and Kegan Paul.

TABLES AND FIGURES

Figure 1: Citizenship programme

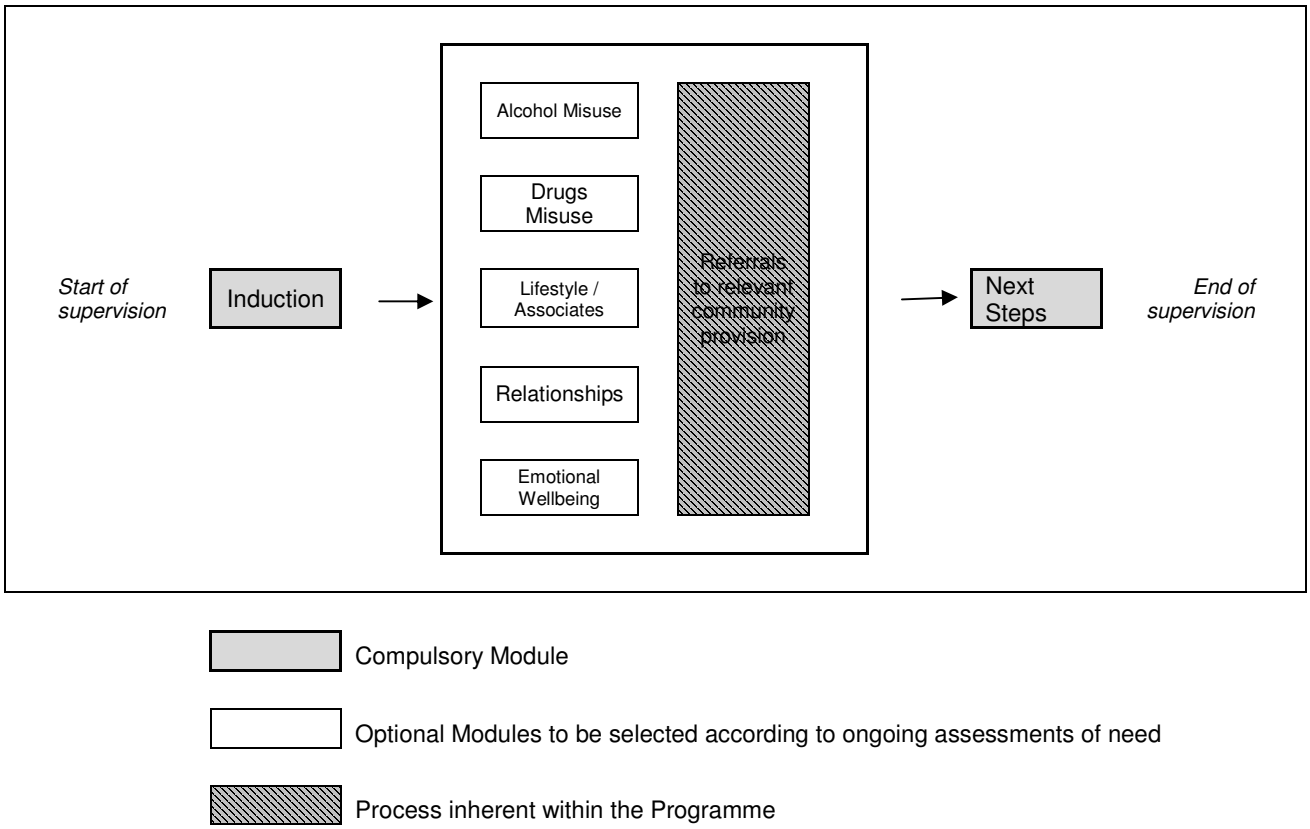
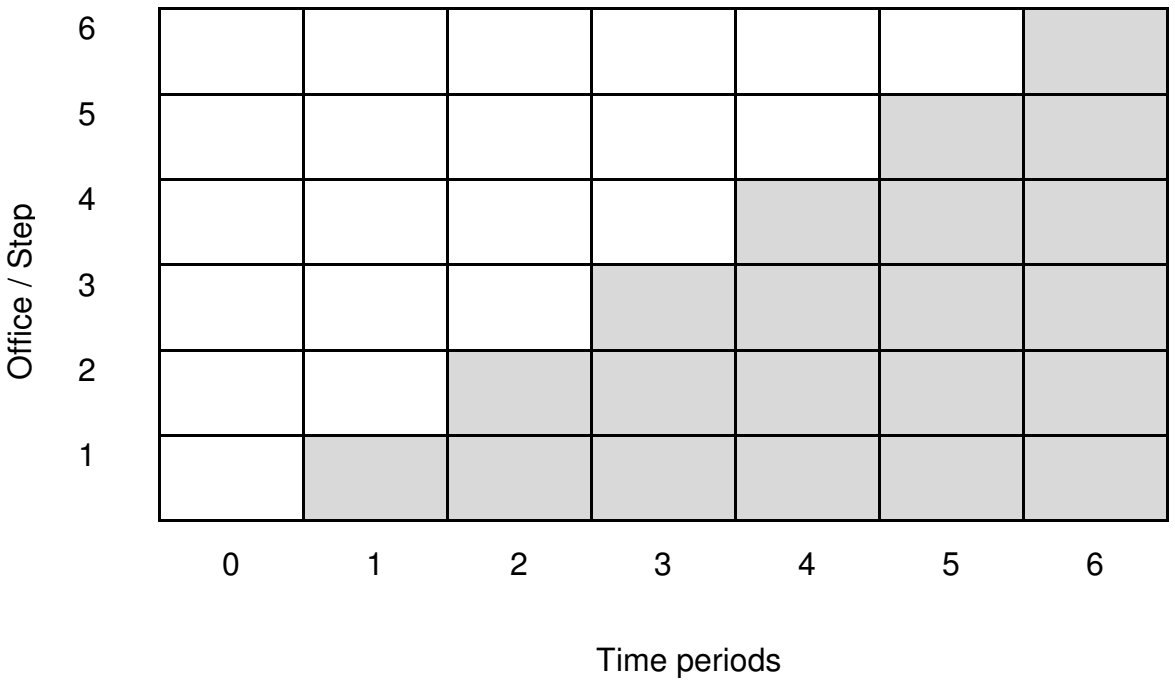


TABLE 1
AREA B IMPLEMENTATION PLAN

Office / Step	Roll-out date
1	April 2007
2	June 2007
3	August 2007
4	October 2007
5	December 2007
6	February 2008

Figure 2: Programme roll-out in Area B



= Citizenship implementation

= Comparison group

Figure 3: Programme roll-out in Area C

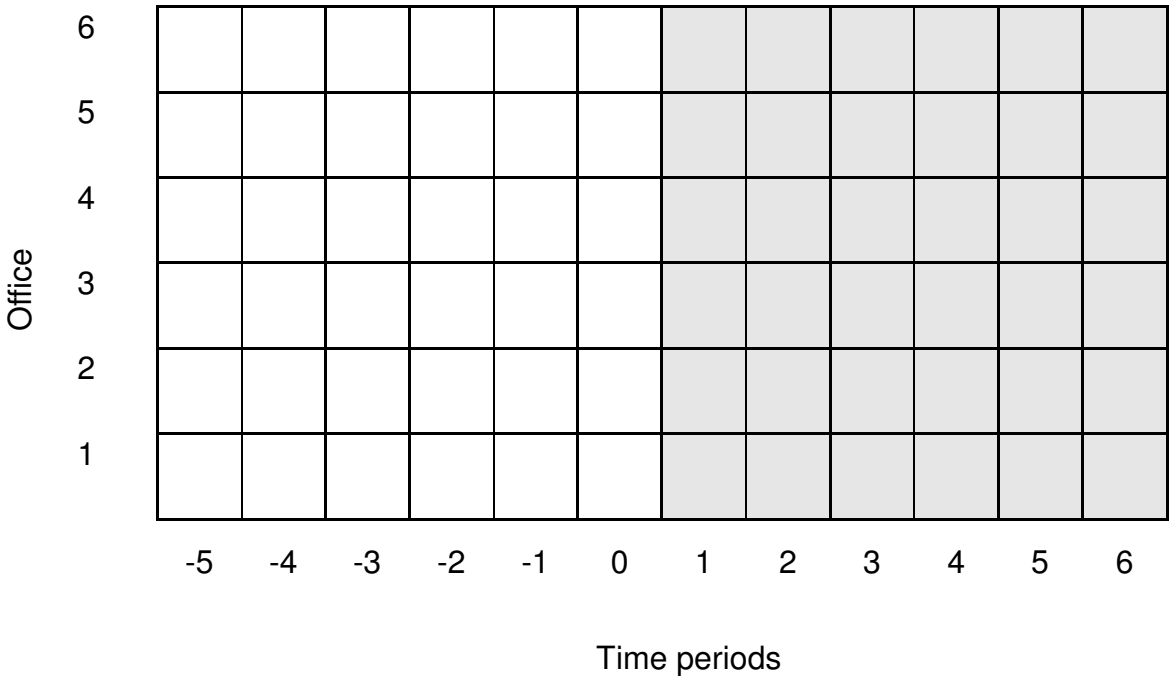


TABLE 2
PROGRAMME IMPLEMENTATION IN PROBATION AREAS

Area	*Use of programme*	*Total eligible*	*Percent*
A	3,072	4,078	75.3%
*B	188	426	44.1%
C	2,325	8,439	27.6%
Total	5,585	12,943	43.2%

NOTE: * Area B targeted the programme at a more troublesome sub-set of those targeted in the other two Areas.

TABLE 3
PROGRAMME IMPLEMENTATION IN AREA B
(randomized design)

Step / Office	*Use of programme*	*Total eligible*	*Percent*
1	26	65	40.0%
2	50	98	51.0%
3	66	144	45.8%
4	14	35	40.0%
5	21	67	31.3%
6	11	17	64.7%
Total	188	426	44.1%