College of Policing

college.police.uk

College of Policing stop and search training experiment

Design of the randomised controlled trial

Paul Quinton

This publication is available at: http://whatworks.college.police.uk/Research/Pages/Published.aspx.

The College of Policing will provide fair access to all readers and, to support this commitment, this document can be provided in alternative formats.

Any enquiries regarding this publication, including requests for an alternative format, should be sent to us at: contactus@college.pnn.police.uk.

Contents

The design of the trial

Overview

This paper describes the design of a randomised controlled trial (RCT) carried out by the College of Policing (the College), which was used to evaluate the implementation and impact of a training programme on stop and search in six pilot forces. The RCT involved assigning 1,323 police officers – who were 'regular users' of stop and search – to two groups at random and in roughly equal numbers. One group was a treatment group (to be trained) and the other was a control group (not to be trained).

The training consisted of a pre-read and assessment, followed by a one-day classroom training session. The focus of the training was on practical legal decision making, unconscious bias and procedural justice. Its aim was to have a positive impact on officers' knowledge and attitudes about stop and search, their reported anticipated behaviours, the proportion of searches that resulted in arrest and the quality of the written grounds for suspicion recorded by officers.

The impact evaluation that was subsequently carried out sought to determine whether the training had had an effect on these outcomes (ie, whether the training 'worked'). The related process evaluation aimed to explore the quality and nature of implementation, the context and mechanisms of change and officer behaviour after the training had been introduced (ie, why and how the training 'worked').

Summary results have been published separately (Quinton and Packham 2016) as have the more detailed findings of the impact evaluation (Miller and Alexandrou 2016a and 2016b) and the process evaluation (Giacomantonio et al 2016a and 2016b).

Overall design

The training pilot was implemented as an RCT. The design involved individual officers being assigned at random to the treatment group (to receive training) or control group (not to receive training). Despite randomisation taking place at the officer level, the treatment and control groups were stratified by virtue of the sampling process. The resulting hierarchy within the trial data was taken into account during the analysis for the impact evaluation (Miller and Alexandrou 2016a).

An RCT design was principally adopted because of the interest of the Stop and Search Project Board in the impact evaluation being able to demonstrate whether the training pilot had had a causal impact on outcomes. This decision was not without risks, particularly as randomisation at the officer level meant there was potential for contamination, resulting from treatment group officers routinely working with control group officers and sharing their practices, which would inevitably dampen the effect of the training.

Cluster randomisation was considered as it would have minimised the risks of contamination, but was not a feasible option in practice.[1] The structure of the six pilot forces varied markedly, meaning they did not share common organisational units to enable cluster

[1] It would have involved randomly assigning officers in groups (eg, forces, operational shifts or local policing units) to treatment or control conditions.

randomisation to be carried out in a straightforward fashion.[2] More importantly, most force leads reported that it was not possible to abstract whole shifts of officers from operational duties for training, but that individual officers could be abstracted from across the force area as operational resilience would be maintained. Furthermore, a recent cluster RCT in the MPS, which involved shifts of officers being assigned to treatment or control, highlighted that the shifts suffered from a high level of officer turnover (of up to nine per cent per month), which was difficult and costly to monitor and threatened trial implementation and the ability of the evaluation to detect whether the training had had a significant effects (Grossmith et al 2015).

The use of a quasi-experimental design – which would involve selecting the officers or operational units to be trained on a non-random basis – was considered but rejected for similar reasons. In addition, a quasi-experiment would have been more costly than an RCT because of the need for pre-test data to be gathered and analysed in the impact evaluation.[3] It would also be more difficult to attribute any change in outcomes to the training with a quasi-experimental design, because other explanations for the change could not be ruled out.

The selection of the pilot forces

The training pilot was implemented in six volunteer police forces:

- British Transport Police (BTP)[4]
- Cleveland Police
- Greater Manchester Police (GMP)
- the Metropolitan Police Service (MPS)[5]
- Sussex Police
- Thames Valley Police (TVP).

The decision to test the training programme in several forces, rather than in a single force, was informed by the EHRC's view that the pilot should be 'national' and seek to reflect a number of different operational contexts (eg, large and small forces, urban and more rural forces).

The original aim was for the College to identify seven or eight forces in the first instance, with the expectation that about five of these forces would eventually be able to take part in the pilot. The process involved the College contacting 44 police forces[6] about the training pilot in February 2015. The initial approach was made via the network of force stop and search leads, who attended the Police-Public Encounters Board.[7]

The College explained what the pilot was likely to entail and asked for volunteers to take

[2] Randomisation with unequal clusters is possible, but would have added to the overall complexity of the evaluation design, particularly as the absence of baseline data meant that it was not possible to estimate the intra-cluster correlation for the main outcome measures.

[3] In theory, the procedures for randomisation in an RCT should ensure that treatment and control groups are, on average, equivalent to one another prior to implementing the intervention, meaning pre-test outcome data are not essential even if they are often desirable.

[4] The training pilot was carried out with officers from A, B and C Divisions as they covered England and Wales. D Division was excluded as it covered Scotland, which has a different legal system.

[5] The training pilot was carried out with officers from four operational command units (OCUs), which were proposed by the MPS: Brent, Havering, the Roads and Transport Policing Command and Tower Hamlets. Each OCU provided a quarter of the total sample for the force.

[6] The 43 'Home Office' forces covering England and Wales, plus the British Transport Police.

[7] A national forum for the police and members of the public to discuss issues related to stop and search. This quarterly meeting is chaired by the national lead for stop and search, Deputy Chief Constable Adrian Hanstock.

part. A total of 13 forces[8] expressed an interest in participating as a result of the initial approach or subsequent contact from the College. Discussions were then held with each force about what was required from volunteers, namely the ability to:

- deliver classroom training between July and September 2015
- comply with experimental conditions up until the end of December 2015
- provide data and research access to support the impact and process evaluations.

Each force's overall use of stop and search also featured in the discussions, as the officers who were to participate in trial were required to be 'regular users' of the power so that it was feasible for the evaluation to detect whether the training had had a significant effect.

These discussions quickly revealed that it was unrealistic for six of the 14 forces to participate in the trial, largely because they did not have the capacity to deliver training to about 100 officers at relatively short notice in addition to what was already planned. Despite positive preliminary discussions, a further two forces had to withdraw from the pilot due to other training priorities or concerns about officer abstractions. Agreement was subsequently reached with the six forces listed above about their participation.

Table 3 provides a profile of the six pilot forces in terms of their use of stop and search based on the 2012/13 data that were available nationally at the time of their selection (Home Office 2014). While three of the six pilot forces (ie, Cleveland, MPS and Sussex) might be regarded as relatively 'high users' of stop and search, the other three forces were relatively 'low users' (ie, TVP, GMP and BTP).

Including forces with relatively low search rates in the training pilot was recognised as a risk, as doing so could have affected the ability of the evaluation to detect whether the training had had a significant impact on officer behaviour. This situation was very likely to have been exacerbated by large reductions in the number of recorded searches across England and Wales that were highlighted in the data that only became available nationally after the training pilot had been agreed (Home Office 2015).

The eligibility of officers to participate in the trial

To be eligible to participate in the trial, police officers needed to:

- be serving constables (including probationers)
- work in a role where use of stop and search was ordinarily expected (eg, neigbourhood, response or roads policing)
- have used stop and search 'regularly' during 2014/15 (so that it was feasible for the impact evaluation to detect any changes in practice resulting from the training)[9]
- be available for the pilot training and ordinary duties afterwards (eg, not on annual maternity or long term sick leave, or on restricted duties).

[8] The forces included: Bedfordshire, British Transport Police, Cambridgeshire, Cleveland, Derbyshire, Greater Manchester, Hampshire, Metropolitan Police, Northamptonshire, North Yorkshire, Nottinghamshire, Surrey, Sussex and Thames Valley.

[9] The original focus was going to be on 'high' use rather than 'regular' use of stop and search, but was changed so that the evaluation results had greater external validity (ie, they were more generalisable).

Table 3. Profile of the pilot forces at time of selection (2012/13 unless otherwise stated)

Pilot forces	Recorded s1 searches (n)	Constables* (n)	Search rate (per constable)	Ranked use (out of 44)	Arrests from s1 searches (n)	Arrest rate (%)	Disproportionality ratio**
BTP	10,206	1,960	5.21	36th	881	8.63	–
Cleveland	31,567	1,118	28.24	1st	5,036	15.95	1:1.8
GMP	33,742	5,611	6.01	31st	2,524	7.48	1:2.2
MPS	359,287	23,283	15.43	4th	43,141	12.01	1:3.2
Sussex	23,559	2,101	11.21	11th	4,098	17.39	1:5.5
TVP	23,227	3,403	6.83	30th	1,619	6.97	1:2.8

*Full time equivalent constables on 31 March 2013.
**Ratio compares stop and search rates for 'Black' or 'Black British' people to rates for 'White' people in 2011/12 (the last available force-level data). Data for BTP were not available.
Sources: Home Office 2014, Dhani 2014 and Ministry of Justice 2013.

Sample size

The decision about the number of officers to be included in the trial was a pragmatic one. Following discussion with members of the Stop and Search Project Board, it was agreed that each pilot force should train about 100 officers and have a similar number in the control group. It was felt that 600 represented a substantial, yet manageable, number of officers for the pilot forces to train in total and was likely to be sufficiently large for the evaluation to be able to detect whether the training had had a significant effect at a programme level and possibly even at a force level.

A power calculation was not carried out because the size of the effect the training might have was not known and because officer survey data on stop and search were not available for the pre-test period.[10] A larger sample size would have enabled relatively small effects to be detected and countered the impact of any contamination that would reduce the effect of the pilot training.

A similar training RCT carried out by the College had a total sample size of 600 officers, of which 240 officers were assigned to the treatment and 360 to the control (Wheller et al 2013). As this RCT was carried out in a single force, it would have been unaffected by clustering in the data. There was a concern that clustering in the data might affect the current trial, even though cluster randomisation was not used (Torgerson and Torgerson 2013). Clustering usually requires the overall sample size to be increased to counteract the effect of the individual units within each cluster sharing similar characteristics simply by virtue of them being within that cluster (eg, all officers in one pilot force being exposed to similar organisational factors, which are likely to be different from those in another pilot force). The adjustment to the sample size that was required to take account of the intra-cluster correlation was unknown, however. The fact that a stratified sample was used may have increased statistical power, although improved power tends to occur in smaller trials (Kernan et al 1999). The analysis was adjusted for the stratification at the force level. .

Assuming a simple random sample, a sample size of 600 tends to produce confidence intervals of up to +/- four per cent around a survey estimate (meaning differences of more than about eight percentage points between treatment and control are likely to be significant). A sample of 100 would produce confidence intervals of up to +/- 10 per cent (meaning differences between treatment and control of at least 20 points are likely to be significant). Assuming a three-month post-test period and given the relatively low search rates in some of the pilot forces (see table 3), it was estimated that both the treatment and control groups in each force would carry out a minimum of 150 searches during the evaluation.

The above sample sizes were increased by 10 per cent to take account of expected training non-attendance and sample attrition (eg, due to officers leaving the force or changing roles). The resulting sample sizes at pilot force and programme levels are presented in table 5.

Table 5. Estimated sample sizes

	Estimated sample sizes (n)	
	Force level	Programme level
Control group officers	110	660
Treatment group officers	110	660
Total	220	1,320

[10] The primary outcome measures in the impact evaluation were to be derived from an officer survey

The procedures for randomly selecting officers

The following procedures were used to identify 220 eligible officers from each force to be included in the trial. Randomly selecting officers from those eligible to take part should, in theory, reduce bias within the sample and help ensure that the results of the impact evaluation are generalisable to that wider group of officers. The procedures were followed by a College researcher (Shayan Moftizadeh), under the supervision of the principal investigator (Paul Quinton), in July 2015. The process was effectively blinded because, even though the researcher and principal investigator had access to each officer's collar number, they had no prior knowledge of any of the officers.

1. To identify officers who had used stop and search 'regularly', analysis was carried out of each pilot force's stop and search dataset, which contained the encounter level details of all recorded searches in 2014/15. Before the datasets were shared with the College, the pilot forces were asked to remove any variables that would enable the person searched to be identified (eg, name, contact details).

   For each dataset, a pivot table was created in Excel showing the number of recorded searches carried out each month and in total, and the monthly average, for each officer (identified by the collar number on each search record). The number of months in which each officer had carried out a search was then calculated to create a 'regular use' estimate (ranging from 1 to 12).

   This approach sought to place greater emphasis on the consistency of use, rather than on the overall volume of use (which would have been emphasised if the mean was used). For example, an officer who searched one person every month would have scored the maximum on the scale (12), whereas an officer who carried out 24 searches in one month but none during the rest of the year would have scored the minimum (1). The officers were then ranked in descending order according the estimate.
2. A minimum 'regular usage' threshold was set for each pilot force within which it was possible to identify at least 250 officers, anticipating that around 30 officers would not be eligible to be included in the trial (see table 6).[11] Differing use levels of stop and search across the pilot forces precluded use of a consistent threshold.
3. The required number of officers for each force was then selected at random from within each threshold. First, a random number was generated for each officers. Second, the officers were sorted, in ascending order, using these random numbers. Third, the first 250 officers from each force were then selected.
4. The force/BOCU samples were then checked against other data to make sure they did not contain officers other than constables. The pilot forces were then asked to carry out further eligibility checks on the officers in the sample, based on the other criteria described above (ie, role and availability).
5. If the number of eligible officers fell below 220 in a force, an additional sample was drawn. The process for doing so was the same as before and involved selecting officers at random within the 'regular usage' threshold for that force. If the officers in the threshold ever became exhausted, officers were selected at random from the next tier down.

[11] A similar threshold was set in each of the four participating OCUs in the MPS, to identify 63 officers from each (at least 250 in total).

Table 6. The minimum ‘regular usage’ thresholds

Pilot force	Officers who carried out S&S in 2014/15 (N)	Minimum ‘regular usage’ threshold	Officers in threshold (n)
BTP	1,396	5	259
Cleveland	595	5	258
GMP	2,285	6	306
MPS – Brent	529	9	72
MPS – Havering	354	6	73
MPS – Roads and Transport	662	7	71
MPS – Tower Hamlets	543	7	76
TVP	2,882	7	296
Sussex	1,479	6	273
Total	10,727	-	-

Table 7 shows how many officers were subject to eligibility checks and the resulting number who were subsequently included or not included in the trial. The pilot forces have been anonymised from this point onwards, so that they cannot be identified in the impact evaluation report (Miller and Alexandrou 2016a).

Table 7. The number of officers subject to eligible checks included / not included in the trial

Pilot force	Officers subject to eligibility checks (n)	Officers included in trial (n)	Officers not included in trial (n)
Force A	250	220	30
Force B	296	224	72
Force C	252	220	32
Force D	250	220	30
Force E	267	220	47
Force F	327	219	108

The procedures for randomly assigning officers

Eligible officers included in the trial were assigned to the treatment and control groups using a stratified block randomisation process. The procedures for randomisation that were adopted had previously been used and peer reviewed as part of an earlier RCT testing the impact of procedural justice training (Wheller et al 2013).

The following procedures were followed by the principal investigator in July/August 2015. As before, the process was effectively blinded because the principal investigator had no prior knowledge any of the officers.

1. An ‘eligible officers’ dataset was created for each pilot force, which contained details of the 220 officers who met the criteria for participation in the trial. These datasets included additional basic data on the eligible officers (eg, on their role and location) that had been provided by the pilot forces.
2. Each dataset was stratified, which involved sorting the officers according to the

following variables (in order and where data were available):[12]

- general role – a recoded variable referring to the officer's general operational role (eg, response, neighbourhoods or specialist)[13]
- specific role – a variable referring to the officer's operational role in more specific terms (eg, firearms, dogs or roads policing)[14]
- location – a variable referring to the officer's local policing unit or station[15]
- a random number generated in Excel.

The stratification process aimed to ensure that the treatment and control were reasonably balanced in their composition (eg, including different types of officers from across the force). Achieving such balance would reduce the likelihood of any particular officer characteristic dominating either group, which could otherwise result in any post-test differences in outcome data being attributed wrongly to the pilot training.

3. The officers in each dataset were then assigned alternately to the treatment or control groups (eg, ABABAB…). A random number was generated in Excel to decide whether to start with the treatment or control (i.e. treatment < 0.5, control ≥ 0.5).
4. The resulting lists of treatment and control group officers were shared with the pilot forces, who carried out a final set of checks. Occasionally, these checks identified one or two officers who had since become ineligible to participate (eg, they had left the organisation, been promoted to sergeant, or were duplicates) and who were substituted with new officers selected at random from the original lists of those who were eligible. The pilot forces were asked to share information with all 220 officers about the trial and invite those assigned to the treatment group to take part in the training. They were also reminded of the importance of maintaining the experimental conditions (eg, not to switch officers between the treatment and control).

The pre-test equivalence of the treatment and control groups

In theory, randomisation should ensure that the treatment and control groups are, on average, equivalent to one another before the intervention is implemented in terms of all known and unknown attributes. This means that any differences identified afterwards can be directly attributed to the intervention. Analysis of the additional basic data provided by the pilot forces indicated that the randomisation procedures had been successful in this respect and did not point to any systematic biases that were likely to affect the results of the RCT.

As table 8 shows, the treatment and control groups were very similar in terms of officer characteristics, such as length of service, sex and role (the latter being expected because it was a stratifying variable). The differences between the two groups were generally no greater than one or two percentage points and did not point to any systematic biases. Furthermore, the average number of searches carried out each month by officers in the two groups was almost identical in 2014/15. Finding equivalence in search rates in the pre-test period is arguably more important than for other officer characteristics as it is directly related to the research questions but was not controlled for in the randomisation process.

[12] A decision was made not to stratify the datasets according to the officer's shift or team. Doing so would have unnecessarily increased the risk of contamination between treatment and control by, in effect, 'forcing' officers from the same shift or team into different groups.

[13] These data were unavailable in one pilot force.

[14] These data were unavailable in one pilot force.

[15] Location data were not used in the randomisation process in the MPS as location had already been taken into account as a result of the four participating OCUs each providing a quarter of the total sample of officers.

Table 8. Pre-test differences between treatment and control

	Control group officers (n=661)	Treatment group officers (n=662)
Pilot force (%)		
Force A	17	17
Force B	17	17
Force C	17	16
Force D	17	17
Force E	17	17
Force F	16	17
Total	100	100
Service length (%)		
0-4 years	24	25
5-9 years	39	35
10-14 years	22	26
15-19 years	7	5
20-24 years	2	3
25+ years	6	5
Missing data*	1	2
Total	100	100
Sex (%)		
Female	11	10
Male	89	90
Total	100	100
Role (%)		
Neighbourhoods	27	27
Response	45	46
Specialist	11	11
Missing data*	17	17
Total	100	100
Search rate (2014/15)		
Officer monthly average	1.91	1.88

*Missing data were from one pilot force, where they were not available.

Treatment levels and intention-to-treat

The aim was for all officers in the treatment group to be trained and for all the officers in the control group not to be trained. As with most RCTs, there were some differences between what was expected and what happened in practice. As table 9 shows, 87 per cent of officers assigned to the treatment group attended the classroom training (the same treatment level achieved by Wheller et al (2013) in their police training RCT). Markedly lower treatment levels among this group would have put at risk the ability of the impact evaluation to detect whether the training had had a significant effect.

Table 9. The number and proportion of officers who attended a classroom session

Pilot forces	Control group officers				Treatment group officers				Total officers (n)
	Assigned (n)	Trained (n)	Not trained (n)	Treated (%)	Assigned (n)	Trained (n)	Not trained (n)	Treated (%)	
Force A	110	0	110	0.00	110	92	18	83.6	220
Force B	112	0	112	0.00	112	96	16	85.7	224
Force C	111	5	106	4.50	109	99	10	90.8	220
Force D	110	0	110	0.00	110	97	13	88.2	220
Force E	110	0	110	0.00	110	94	16	85.5	220
Force F	108	0	108	0.00	111	98	13	88.0	219
Total	661	5	656	0.76	662	576	86	87.0	1,323

Despite the attendance rates among treatment group officers, five control group officers from one force also attended a classroom session as a result of their supervisors mistakenly instructing them to book onto the course (less than one per cent of the control group). While their attendance would represent a direct form of contamination between treatment and control, the numbers of officers mean it was likely to have a minimal effect on the results of impact evaluation.

Intention-to-treat analysis was carried out as standard in the impact evaluation (Miller and Alexandrou 2016). This meant that all officers were included in the analysis – where data were available – regardless of whether they completed the pilot training or not. Including untrained treatment group officers and trained control group officers would provide more of a 'real world' assessment of the training pilot.

The reasons for treatment group officers not attending a classroom session were recorded by the pilot forces (see table 10). In a large majority of cases, officers reportedly did not attend for practical reasons (eg, being on leave, changing roles, or scheduling difficulties) or because they were no longer eligible to be included in the trial. The reason for non-attendance was unknown for 12 officers (14 per cent of non-attenders). The extent and nature of training non-attendance did not seem to point to any systematic biases that were likely to affect adversely the impact evaluation.

Table 10. The reported reasons for officers not attending classroom training sessions

Reason for non-attendance	Percentage of non-attending officers (n=86)
Annual leave	8.14
Sick leave or restricted duties	20.93
Other leave*	6.98
Unavailable to attend	33.72
Ineligible**	16.28
Unknown	13.95

*Such as maternity or adoption leave, special leave due to bereavement, or being on a career break.
**Such as changing roles or leaving the force.

Limitations

There are a number of limitations with the design of the trial and evaluation framework. These limitations are summarised below. The limitations associated with the specific research carried out for the impact and process evaluations are detailed in final reports for those two pieces of research (respectively Miller and Alexandrou 2016a and Giacomantonio et al 2016a).

- **Direct and indirect contamination** – The trial is very likely to have suffered from a degree of contamination, which could have reduced the chances of the training having an impact on outcome measures. There was some evidence of 'direct contamination' in that five control group officers from one pilot force mistakenly attended classroom training sessions. The small number of officers affected by this treatment error (representing less than one per cent of the control group) was, however, unlikely to affect the results of the impact evaluation, particularly as the level of treatment achieved in the treatment group was relatively high.

  More important was likely to have been the influence of 'indirect contamination', which may have resulted from treatment group officers working (possibly) with control group officers and (definitely) with officers not participating in the trial. Contamination would have resulted if treatment group officers failed to change their attitudes or behaviour because of the influence of untrained colleagues or, conversely, if control group officers changed their attitudes or behaviour because of the influence of trained officers.

The possibility of indirect contamination was an inevitable result of the design of the trial, which was required to randomise at the individual officer level as cluster randomisation was not feasible. The effect of indirect contamination should, however, be taken as read.

First, officers were selected from across the pilot force areas to participate in the trial. Thus, while control group officers would have routinely worked with non-participants, the extent to which they worked with control group officers is unknown.

Second, it is not clear exactly how contamination would affect officer attitudes and behaviour. Talking to a colleague about the contents of training is unlikely to have the same effect as actually attending the training. Nevertheless, it is probable that officers will learn from each other – through observation – when working together. Moreover, when officers return from training to their usual work setting, and without the presence of peers who have been through the same training, there is a chance they will also return to the same work routines as before. This may be exacerbated if elements of the training go against established practice and require the officer to 'go against the grain'.

Third, some outcome measures may be more or less susceptible to indirect contamination than others. While the behavioural outcome measures (based on police data) may be particularly prone to the influence of other officers, this is arguably less of an issue with the attitudinal outcome measures (based on the officer survey) as they reflect more of the private domain.

While indirect contamination remains a possibility, the involvement of a large number of officers in the trial should have helped to mitigate its impact (which would be to reduce the chances of the training having an effect) because larger sample sizes enable smaller effects to be detected.

- **Blinding** – The trial was not formally blinded in that, for the most part, the researchers and pilot forces were aware of which officers had been assigned to the treatment and control groups. It is possible, therefore, that bias may have affected decisions during implementing the training pilot, designing the trial and conducting the evaluation. It should be noted, however, that, while researchers had access to participating officers' collar number and relevant data, they had no prior knowledge of any of them.

- **Internal and external validity** – Pre-test comparisons between the treatment and control suggested that the two groups were, on average, equivalent to one another before the intervention was implemented. As there was little evidence of any systematic bias, it can reasonably be argued that the RCT had internal validity (not that differences at the baseline necessarily point problems with internal validity).

  There are, however, limits as to the extent to which the results of the trial can be generalised. The random selection of participants (rather than their random assignment) means that the results of the impact evaluation are likely to be generalisable to 'regular users' of stop and search in the six pilot forces. It would not be possible to claim that the results have external validity beyond this specific population.

  Irrespective of the results of the impact evaluation, it is not known what effect the training would have had with different groups of officers (eg, high or lower than average users, or public order units) in a different context (eg, non-volunteer forces who may have less interest in changing stop and search practices, or in forces with high rates of stop and search).

- **Construct validity** – The pilot training cannot be described as a fixed and singular intervention. It consisted of central guidance that was used locally to develop training. Hence, there was considerable scope for both training content and delivery to vary between forces. This lack of consistency could have reduced the chances of the training having an impact on outcomes, and raises question as to whether the RCT provided a good test of the ‘training’.

  Moreover, because randomisation took place at the individual officer level, the pilot only consisted of training delivered to frontline police officers. It lacked any of the other supporting mechanisms that would usually accompany a training programme and which would be used during national roll-out (eg, the training of supervisors and senior officers, follow-up briefings, reminders). The evidence suggests that multiple methods tend to be more effective than single methods when seeking to change behaviour (Wheller and Morris 2010).

- **Speed of design and implementation** – The nature of the commissioned work programme meant there was relatively limited time to design the training intervention, trial and evaluation. This resulted in several pragmatic decisions having to be made (eg, sample size), which would usually be subject to more consideration. The timescales also meant that it was not possible to develop a trial protocol before piloting and that some aspects of the impact and process evaluations had to be agreed before the pilot training had been fully developed.

- **Follow-up period** – While the impact evaluation aimed to assess whether the pilot training had sustained impact on outcomes via its Wave 2 survey, the follow-up period was limited to about three months. Irrespective of the results of the impact evaluation, it is not known what effect stop and search training might have on police practices in the longer term.

- **Wider context** – As previously discussed, after work on the trial commissioned work programme was well underway, newly released national data pointed to marked changes in police stop and search practices having taken place in the two years beforehand. These changes may have inadvertently limited the ability of the training pilot to affect change and the ability of the evaluation to detect it, although the RCT design should have ameliorated their impact.

References

Dhani, A. (2014) Police service strength England and Wales, 31 March 2012. London: Home Office.

Giacomantonio, C., Jonathan-Zamir, T., Litmanovitz, Y., Bradford, B., Davies, M., Strang, L. and Sutherland, A. (2016a) College of Policing stop and search training experiment: Process evaluation. Final report. Ryton-on-Dunsmore: College of Policing.

Giacomantonio, C., Jonathan-Zamir, T., Litmanovitz, Y., Bradford, B., Davies, M., Strang, L. and Sutherland, A. (2016b) College of Policing stop and search training experiment: Process evaluation. Appendices. Ryton-on-Dunsmore: College of Policing.

Grossmith, L., Owens, C., Finn, W., Mann, D., Davies, T. and Baika, L. (2015) Police, camera, evidence: London’s cluster randomised controlled trial of Body Worn Video. Ryton-on-Dunsmore and London: College of Policing and MOPAC.

Home Office (2014) Police powers and procedures: England and Wales, 2012 to 2013. London: Home Office.

Home Office (2015) Police powers and procedures: England and Wales, year ending 31 March 2015. London: Home Office.

Kernan W., Viscoli, C., Makuch, R., Brass, L. and Horwitz, R. (1999) Stratified randomization for clinical trials. Journal or Clinical Epidemiology, 52(1):19–26.

Miller, J. and Alexandrou, B. (2016a) College of Policing stop and search training experiment: Impact evaluation. Final report. Ryton-on-Dunsmore: College of Policing.

Miller, J. and Alexandrou, B. (2016b) College of Policing stop and search training experiment: Impact evaluation. Report appendices. Ryton-on-Dunsmore: College of Policing.

Ministry of Justice (2013) Statistics on race and the criminal justice system 2012. London: Ministry of Justice.

Torgerson, C. and Torgerson, D. (2013) Randomised trials in education: An introductory handbook. London: Education Endowment Foundation.

Wheller, L. and Morris, J. (2010) What works in training, behaviour change and implementing guidance? London: National Policing Improvement Agency.

Wheller, L., Quinton, P., Fildes, A. and Mills, A. (2013) The Greater Manchester Police procedural justice training experiment: The impact of communication skills training on officers and victims of crime. Ryton-on-Dunsmore: College of Policing.