College of Policing stop and search training experiment: Impact evaluation

Final report

Joel Miller and Banos Alexandrou

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.

Foreword

The research described in report was commissioned by the College of Policing in July 2015 and conducted by the Research Advisory Service. The report presents the findings from a randomised controlled trial of a pilot stop and search training programme that was conducted in six police forces in England and Wales in 2015. The research produced mixed results concerning the impacts of the training. While the research showed that the training had some effects on officers' knowledge, attitudes and anticipated stop and search behaviours, it also showed that the training had few measurable impacts on recorded street-level behaviour. The findings provide a useful reference point for further development of stop and search training.

For more information, please contact Banos Alexandrou of the Research Advisory Service at banos@researchadvisoryservice.org.uk or visit the Research Advisory Service website: www.researchadvisoryservice.org.uk.

Acknowledgements

We would first like to express our gratitude to the many police officers and support staff across the six participating police forces who participated in this study. We would like to offer our thanks to staff at the College of Policing who commissioned and supported this work, including Paul Quinton, Daniel Packham and Rory McKenna. We also wish to thank research scientists Nicole Sachs and Sarah Trocchio who played a critical role in the project by conducting extensive coding of police stop and search records. Finally, we would like to thank our academic reviewers, Tracey Meares and Wes Skogan, who provided invaluable feedback which substantially strengthened this paper.

About the authors

Joel Miller is an associate professor at Rutgers University School of Criminal Justice (USA). He has worked previously at the UK Home Office, Spain's University of Malaga, and the Vera Institute of Justice, New York. His research is focused on strategies to improve the effectiveness of, and public confidence in, criminal justice policies. His research has examined a diverse range of criminal justice areas, including police tactics and accountability, crime prevention, risk assessment and juvenile justice.

Banos Alexandrou is director of the Research Advisory Service and has been designing and delivering market and social research and evaluation projects for the public and private sectors for the past 20 years.

Summary

This report presents results from a randomised controlled trial of a pilot stop and search training programme. The training was designed to promote the non-discriminatory use of police stop and search powers, strengthen officers' knowledge and skills in applying reasonable suspicion, improve the treatment of members of the public and improve outcomes from encounters. It was led by the College of Policing (the College) in collaboration with the Equality and Human Rights Commission (EHRC).

The pilot was carried out in six police forces within England and Wales. A total of 1,323 uniformed officers were included in the study. They were selected because they were regular users of stop and search powers. They were then randomly assigned to a treatment group that was targeted for the pilot training (87 per cent ultimately received training) or a control group that was not intended to receive the training (0.8 per cent received training).

Here, we present the findings from an impact evaluation of the pilot based on three sources of data. These are:

- a Wave 1 survey, carried out a few days or weeks from the end of officers' pilot training
- a Wave 2 survey, initiated between about three and five months following the end of officers' training
- data generated from police search records, drawing from the three calendar months prior to the beginning of training and the three calendar months following the completion of the bulk of training in each force.

Analysis of the survey results tested hypotheses concerning the training's effects on officers' knowledge, attitudes and anticipated behaviours, while actual behaviours were measured through police stop and search records. Key findings, focused primarily on programme-level effects across the six forces, are presented below.

Impacts on officers' preparation and knowledge

- Compared to the control group, officers in the treatment group were a little less likely to report that prior stop and search training (including both pilot and other past training) had prepared them with relevant knowledge and skills, based on the Wave 1 survey. This suggests the pilot training compared unfavourably with officers' recollections of training earlier in their careers.

- Officers' knowledge of stop and search regulations and policy was generally high and was a little higher in the treatment than control groups, as measured in both Wave 1 and Wave 2 surveys. This suggests that the pilot training improved officers' already strong stop and search knowledge and that this improvement was sustained over time.

- In the Wave 2 survey, officers in the treatment group reported less confidence in the adequacy of grounds in written search records than the control group officers did. This suggests that they

had become more stringent in their evaluation of grounds for searches. This effect was found for stronger and weaker grounds although there was some evidence that the effect was greater for weaker grounds.

Impacts on officers' attitudes

- In the Wave 1 survey, treatment group officers averaged slightly less support for police ethnic/racial stereotyping than control group officers, suggesting a small impact of pilot training, although support for ethnic/racial stereotyping was already low among officers. This variable was not measured in the Wave 2 survey.

- In the Wave 1 survey, officers in the treatment group were a little less cynical about the prospect of policies regulating officer stop and search practices, suggesting a modest pilot training impact. This effect did not, however, endure to the Wave 2 survey.

- In both Wave 1 and Wave 2 surveys, there was a somewhat lower level of support for high volume stop and search strategies in the treatment group compared to the control group. This suggests the pilot training prompted officers to favour a more selective use of stop and search in crime control. This effect was sustained to Wave 2

- In the Wave 1 survey, there were no differences between treatment and control groups in their support for procedural justice (ie, being respectful and fair) during stop and search. Consistent with this finding, the process evaluation showed that procedural justice was not a central feature of the training that was delivered (see Giacomantonio et al, 2016). Support for procedural justice was not measured in Wave 2.

Impacts on officers' anticipated behaviours

- When presented with a scenario involving the searching of a confrontational suspect in the Wave 1 survey, there were no clear differences between treatment and control group officers in how they said they would treat the suspect. This applied in relation to both procedural justice principles and the legal procedures used in encounters.

- As expected, when asked the likelihood of them **questioning** potentially suspicious people in a range of different scenarios, there was little difference in response between treatment and control group officers. This was true for examples in both Wave 1 and Wave 2 surveys. This suggests the training did not adversely affect officers' anticipated willingness to intervene in situations.

- When they were asked how likely they were to **search** suspicious people in the same scenarios, however, officers in the treatment group reported notably lower probabilities of doing so, in both Wave 1 and 2 surveys. This was true for scenarios involving suspected robbery or drugs offences as well as for both weaker grounds (as initially hypothesised) and stronger grounds. The effect was strongest for searching when the scenario involved the smell of cannabis as a key basis for suspicion, considered to represent weaker grounds. This may, in part, be due to the

emphasis placed in the training on the smell of cannabis, in isolation, as constituting inadequate grounds for a search (see Giacomantonio et al, 2016).

- Perhaps explaining the pilot training's effects on officers' anticipated search decisions, the Wave 2 survey showed that officers in the treatment group were less likely than those in the control group to evaluate grounds in the scenarios as adequate to justify a search. The survey, however, showed no differences between the groups in officers' declared willingness to conduct searches, provided grounds were present.

- The Wave 1 survey randomly varied the scenario suspects' descriptions between 'black' and 'white' when asking officers about their stop and search decision-making. Officers were generally more likely to say they would question or search white suspects than black suspects.

- There was, however, no statistically significant effect of training on ethnic/racial disparities in officers' anticipated stop and search decision-making.

Impacts on recorded behaviours

- Police data provided no strong evidence of a reduction in officers' search rates directly attributable to the training. There was, however, a small effect that was close to statistical significance, meaning it is possible the training had a marginal effect, in line with officers' responses to survey questions.

- An analysis of officers' written grounds indicated no differences in their quality between treatment and control groups. This suggests the training had no impact on the types of searches being conducted or the detail provided by officers when recording their grounds.

- Police data showed no effects of training on the proportion of searches resulting in arrests, suggesting that the training has produced no improvements in officer effectiveness.

- Police data showed no effects of training on the ethnic/racial distribution of people searched. This was consistent with survey findings showing that the training had no effects on the use of ethnic/racial appearance in officers' decision-making when responding to written scenarios.

Force-specific effects

- Training was associated with more pronounced effects in some forces than others, although variations were not consistent across types of outcome. Key findings include:

    - Force E registered almost no statistically significant effects on the range of outcome variables.
    - Force D experienced the largest number of significant effects of treatment on knowledge and attitudes.
    - Statistically significant effects were found on at least some **anticipated** search behaviours for all forces, as measured by the surveys, except Force E.

    - There was, however, a lack of clear and consistent effects of training on **actual** officer behaviours, as recorded in police data, for any of the forces (apart from two isolated statistically significant effects).

- Force-level differences may reflect variations in the implementation of training between sites. Forces, however, also varied in their geography and organisation which may have influenced how the training was received. Some differences may also be the product of chance variations between forces.

Conclusions

- While the training had some intended effects, these effects were not found for all objectives, were often modest when they were found, and were often inconsistent across forces. Moreover, there were few concrete effects of training found in measured street-level practice. This raises some questions about the utility of the training as it was formulated for the pilot.

- Future stop and search training might usefully give greater emphasis to modelling behaviours in stop and search encounters, alongside abstract teaching about the use and regulation of stop and search powers. This could involve the use of role-plays, for example.

- Future stop and search training should probably place greater emphasis on improving how officers interact with suspects, paying particular attention to procedural justice principles, given that the pilot training had no effects related to procedural justice.

- A training package that also targeted force supervisors and managers might be more effective. Such an approach could involve education in auditing and supervising officers' use of stop and search, and developing supervisors' and managers' skills in encouraging and directing officers to adopt more effective and fairer stop and search practices.

Contents

(Appendices available in a companion document: Miller and Alexandrou, 2016)

1. Introduction

This report presents results from an evaluation that tested the impact of a pilot training programme on stop and search. The pilot training was introduced as a randomised controlled trial (RCT) in six police forces within England and Wales and was led by the College of Policing in collaboration with the Equality and Human Rights Commission (EHRC). See Quinton (2016) and Quinton and Packham (2016) for further details on the background and design of the trial.

Background

Section 1 of the Police and Criminal Evidence Act (PACE) 1984 introduced a new national police power which enabled the police to search members of the public for stolen or prohibited articles where there were 'reasonable grounds for suspicion'. PACE simultaneously sought to authorise, standardise and regulate police practices by repealing the numerous local and national powers that existed previously, the extensive and discriminate use of which was criticised in the Scarman Report (1981) as a major cause of the Brixton riots. Police powers to stop and search were further extended with the introduction of section 60 of the Criminal Justice and Public Order Act 1994 and section 44 of the Terrorism Act 2000 (since repealed) which permitted officers to search members of the public without reasonable grounds when authorised to do so under certain conditions.

Since the introduction of these search powers, their use has been the subject of considerable public debate. The Stephen Lawrence Inquiry (Macpherson, 1999) concluded that the police were, in part, institutionally racist because of their disproportionate searching of people from ethnic/racial minorities. The Equality and Human Rights Commission (EHRC, 2010 and 2013) has since raised concerns about the excessive and disproportionate use of police search powers. More recently, and prompted by research into the 2011 riots (Lewis et al,2011), Her Majesty's Inspectorate of Constabulary (HMIC, 2013) questioned whether stop and search was being used effectively and fairly by forces, highlighting that 27 per cent of sampled search records did not appear to contain sufficient grounds to justify the search.

Last year, in response to the HMIC inspection and a national consultation exercise, the home secretary announced a series of reforms, stating that the misuse of stop and search wasted police time, was unfair (especially to young black men) and damaged public confidence. As part of the reform package, the home secretary (2014) commissioned the College of Policing to 'review the national training of stop and search with a view to developing robust professional standards…[and] unconscious bias awareness training to reduce the possibility of prejudice informing officers' decisions.'

To this end, the College entered into a partnership with the EHRC to develop new National Policing Curriculum learning standards on stop and search and a new pilot training curriculum. This effort has culminated in the implementation of a training pilot in six police forces taken by over 600 officers. This study represents part of a wider evaluation of that training pilot, looking specifically at its impact on outcomes. Findings on the nature and quality of the training pilot's implementation can be found in a complementary process evaluation (see Giacomantonio et al, 2016).

The training pilot

The training pilot was targeted at officers who regularly carried out searches and included the following topics:

- Non-discriminatory and human rights compliant encounters.
- Unconscious bias (including defining and understanding unconscious bias, and how it may affect a range of police-initiated encounters).
- Practical understanding and skills in:
    - applying reasonable suspicion
    - handling interactions in line with procedural justice
    - improving encounter outcomes.
- Skills to improve the capacity of the police to treat people with dignity and respect, and promote and protect human rights in line with the police's statutory obligations.

Hypotheses

This impact evaluation tested hypotheses about the pilot training's effects on officers' knowledge, attitudes and behaviours. Hypotheses were grouped according to whether they reflected primary, secondary, or other outcomes as designated by the trial's designers (see Quinton, 2016) and are described below.

Primary outcomes

- Trained officers have greater knowledge and feel more prepared for stop and search, including:
    - a more positive assessment of their experience of training in stop and search practices and policies
    - improved knowledge of stop and search policies
    - a greater recognition of when grounds for suspicion are inadequate.
- Trained officers have attitudes more favourable to good practice in police–public interactions, including:
    - less support for stereotyping in police suspicion
    - less cynicism toward the regulation of stop and search practice
    - less support for high volume stop and search strategies
    - greater support for procedural justice (Tyler, 2004) in stop and search encounters.
- Trained officers **say** they will behave in line with training standards on stop and search decision-making and practice, including:
    - a greater tendency to follow legal procedures during stop and search
    - a greater tendency to treat people with procedural justice during encounters
    - a lesser tendency to search when grounds are weak

- a lesser ethnic/racial bias in stop and search decision-making.

Secondary outcomes

- Based on search records, trained officers' practice is more professional and effective, including:
  - higher quality recording of grounds
  - higher arrest rates from searches.

Other potential outcomes (with no hypothesised direction of effect)

- Trained officers show other differences in recorded practices compared to untrained officers, including differences in:
  - proportions of searches carried out on people from black and minority ethnic groups
  - the overall number of searches carried out.

2. Data and methods

Six police forces participated in the pilot training programme. In four of the forces, the trial was force-wide and for two it focused on operational subunits of the larger forces. Officers who carried out searches regularly in 2014/15 were selected at random for inclusion in the trial. They were then randomly assigned between a treatment group that was targeted for training and a control group that was to receive no training. Random assignment was conducted after sorting officers by stratifying variables relating to geography and job type. Quinton (2016) and appendix A shows the distributions of key variables for the treatment and control groups respectively at the point of random assignment and in the post-training period. Both indicate close equivalence or balance between treatment and control groups across the aggregate six-force sample.[1]

Training for forces took place between 27 August and 21 October 2015. For five out of six forces most officers were trained during September, while for one force (Force F) substantial numbers of officers were also trained in October. Across forces, nine out of ten treatment group officers attended the training (87 per cent attended overall), and five officers in the control group (0.8 per cent) attended the training in error (in one force). Overall, this represents a high level of adherence to the experimental design.

Wave 1 and 2 surveys

Research data included two online surveys of officers participating in the research. A Wave 1 survey was initiated on the 20 October 2015 for most officers, with first invitations issued at least six days, and up to 54 days, after treatment group officers had had their training session. The start of the survey was held back by a week for the few officers in one force where training ran late. A Wave 2 survey was subsequently initiated on 19 January 2016, with invitations first going out between 85 and 145 days (ie, between about three and five months) following treatment group officers' training sessions. In practice, officers may have filled out their surveys a further three or four weeks after their initial invitation.

A secondary random assignment process was incorporated into the Wave 1 survey. This assessed whether the ethnic/racial description of a suspect affected officers' anticipated decision-making and, furthermore, whether these tendencies were changed by the training. Specifically, the scenarios presented to officers were randomly varied between suspects described as 'white' and suspects described as 'black', and officers were asked whether they would question and whether they would search the suspect. This random assignment was independent of the allocation to treatment and control groups, although the same stratified random assignment method was used. The different ethnic/racial descriptions ('black' and 'white') were, therefore, evenly distributed within both

[1] The overall design and management of the RCT, including random assignment, was carried out by the College of Policing (see Quinton, 2016, for details). Survey design was conducted by the Research Advisory Survey (RAS) and administration of the survey (with some College support) was conducted by Research Support and Marketing (RSM) under contract to RAS. Police data was provided by participant forces (via the College) to RAS. All analysis and reporting was conducted independently of the College by RAS.

treatment and control groups (see table 1 below and appendix D for more details of the questionnaire variations).

Survey measures

The two surveys included various Likert scales, scores and single-item measures. A number of measures were common to both surveys, allowing us to assess whether initial effects identified in Wave 1 were sustained during Wave 2. The Wave 2 survey also included some additional questions that probed issues not addressed in Wave 1. Core measures included the following:

- **Preparation and knowledge**
    - **Perceived contribution of prior training to stop and search knowledge (Wave 1 only)** – This scale focused on officers' evaluation of how any previous training they had received (pilot training or otherwise) had prepared them in terms of legal and procedural knowledge in relation to stop and search.
    - **Perceived contribution of prior training to interpersonal skills (Wave 1 only)** – This scale focused on officers' evaluation of how training has prepared them to interact with members of the public.
    - **Knowledge of stop and search policies (Waves 1 and 2)** – This was a summary score based on response to true/false statements focused on stop and search policies.
    - **Knowledge of PACE Code A**[2] **(Waves 1 and 2)** – This was a summary score based on answers to a slightly reduced set of the same true/false statements used in the 'knowledge of stop and search policy' measure above. It left out a single statement that did not relate to PACE Code A.
    - **Assessed strength of written grounds (Wave 2 only)** – A series of five single items was included measuring officers' assessment of the strength of written grounds for a search, informed by real-life examples of written grounds. The five examples include two 'stronger' grounds and three 'weaker' grounds. Strength of grounds is determined by the inclusion of clearly specified factors that support suspicion according to PACE Code A.
- **Attitudes**
    - **Support for police non-ethnic/racial stereotyping (Wave 1 only)** – This scale assessed officers' support for the use of general stereotypes (relating to young males, dress, models of car, prior offending, people who do not 'fit in') in the formation of suspicion.
    - **Support for police ethnic/racial stereotyping (Wave 1 only)** – This scale assessed officers' support for the use of ethnic/racial stereotypes (relating to white people, Asian people and young black people) in the formation of suspicion.

[2] PACE Code A governs the exercise by police officers of their statutory powers of stop and search.

- **Cynicism toward the regulation of stop and search (Waves 1 and 2)** – This scale focused on officers' scepticism about the way regulations are supposed to govern search decision-making.
- **Support for high volume stop and search strategies (Waves 1 and 2)** – This scale measured officers' support for high frequency search practices to address crime problems.
- **Perceived value of procedural justice principles in stop and search (Wave 1 only)** – This scale measured officers' support for procedural justice principles in stop and search (ie, being respectful and fair).

- **Anticipated stop and search behaviours**

  - **Procedurally just treatment of a stop and search suspect (Wave 1 only)** – This scale measured the emphasis officers placed on acting in a procedurally just fashion during a search when presented with a vignette focused on a confrontational suspect.
  - **Legal treatment of a stop and search suspect (Wave 1 only)** – This scale measured the emphasis officers placed on acting according to legal procedure during a search when presented with a vignette describing a confrontational suspect.
  - **Probability of initiating an encounter (across multiple scenarios) (Wave 1 version)** – In Wave 1, eight single-item measures were used to gauge officers' likelihood of questioning, and then searching, male suspects in four scenarios which varied in terms of: (i) the strength of grounds that were present (ie, stronger or weaker grounds) and (ii) the type of crime that was suspected (ie, drugs or robbery). The ethnic/racial appearance of the suspect in the scenarios was also varied at random between survey respondents in order to test the relevance of ethnicity/race in an officer's anticipated decision-making. In Version A, the suspects in the first two vignettes were described as 'black', and the second two 'white'. In Version B, the ethnic/racial descriptions were reversed (ie, 'white' in the first and 'black' in the second) across the same sets of scenarios. Table 1 below presents the different scenarios provided in the Wave 1 survey.

Table 1. Decision-making scenarios relating to encounters with suspects, A and B assignment, Wave 1 survey

	Survey scenario (each questionnaire includes all scenarios)			
	1	**2**	**3**	**4**
Suspected crime	Drugs	Drugs	Robbery	Robbery
Strength of grounds	Weaker	Stronger	Weaker	Stronger
Suspect appearance varies for officer by A/B random assignment	Varies between: (A) black (B) white	Varies between: (A) black (B) white	Varies between: (A) black (B) white	Varies between: (A) black (B) white
Scenario questions	Everyone asked about: (i) question (ii) search	Everyone asked about: (i) question (ii) search	Everyone asked about: (i) question (ii) search	Everyone asked about: (i) question (ii) search

- **Probability of initiating an encounter (across multiple scenarios) (Wave 2 version)** – In Wave 2, two of these scenarios and questions were reproduced (stronger robbery and weaker drug scenarios) although with no information on the suspect's ethnicity/race.
- **Evaluation of the strength of grounds in the scenarios (Wave 2 only)** – For each of the scenarios used in Wave 2, a follow-up question asked respondents how strong they thought the grounds were to give insights into the reasons for their anticipated behaviours.
- **Propensity to search when grounds are present (Wave 2 only)** – Shedding further light on motivations for search behaviours, this scale assessed how inclined officers are to search when they believe grounds to be present.

- **Background variables**

  - The survey also collected information on age, gender and professional role. These data were supplemented by force records and were available for both survey waves.

Principal components factor analysis was used to check and refine the Likert items included in the scales described above. Generally speaking, we sought factor loadings of at least 0.5 and reliability scores of at least about 0.7 for inclusion, although in practice we made occasional slight compromises on reliability scores. Detailed descriptions of these scales (including factor loading and reliability) are provided in the discussion of results and in appendix C.

Survey testing

The Wave 1 questionnaire was developed using existing literature (eg, Wheller et al, 2013; Rosenbaum and Lawrence, 2012) to identify scales and questions that could be used or adapted, and by creating original scales and questions where prior literature did not provide examples. Questions and scales were refined through an extended pilot testing process. This involved testing questionnaire drafts on officers not involved in the trial. After completing the survey, pilot subjects were debriefed using a cognitive interviewing method. This allowed researchers to learn how officers interpreted questions and to identify obstacles they faced providing valid responses. Successive refinements to the questionnaire were made across multiple testing sessions.

The Wave 2 questionnaire kept many of the measures in Wave 1 that registered effects, while dropping others to allow the introduction of additional measures. New measures sought to clarify reasons for some of the changes seen in Wave 1 as well as providing opportunities for broader analysis of force policies and practices. A much shorter testing process was deployed to assess the Wave 2 questionnaire and to make refinements.

More detail on pilot testing is provided in appendix B. The final surveys, as formatted for electronic distribution, are provided in appendices J and K.

Survey administration

The College emailed all officers assigned to the treatment or control groups (irrespective of whether or not they attended the pilot training) with a unique link to the online survey, although the survey was managed by a third-party research company (RSM). The survey was initially sent to officers at least a week following their training (six days was the minimum), although for the earliest officers trained the gap was a little over six weeks.

In Wave 1, most officers had a period of up to 29 days to complete the survey following the initial email invitation. The College emailed the questionnaire on a Tuesday and sent subsequent reminder emails to non-responders one, two and three weeks thereafter. A final email was also sent two days before the end of the survey period. Two additional emails were sent by a senior officer from each pilot force to encourage officer participation (the first was sent the day before the College's initial email and the second to non-responders the day before the week two reminder). For the handful of officers who attended the last training session in Force F (which was delayed), the survey went out about a week later than for the other trial participants; they received just two reminders.

In the Wave 2 survey, all the officers in the trial had up to 22 days to complete the survey. The procedure followed a similar pattern to Wave 1 but used one fewer College reminder. One force had their initial reminders staggered due to initial technical difficulties that meant officers could not access the survey. Wave 2 also offered an incentive to participation, not included in Wave 1, to combat anticipated fatigue among officers who were being asked a second time to participate in the study. Specifically, in the invitation emails, RAS promised to make a £1 donation to a police charity for each survey completed, up to a total of £400.

Survey response rates

Table 2 provides details of the number of trial officers who completed the Wave 1 and Wave 2 surveys. The table indicates that the overall response rate for Wave 1 was 74 per cent, with a higher response rate for the treatment group (79 per cent) than the control group (69 per cent). Wave 2 response rates were lower, but still credible, with an overall response rate of 49 per cent, with 55 per cent in the treatment group and 42 per cent in the control group. Response rates in individual forces also showed some variations.

While random assignment, on average, produces equivalent treatment and control groups, the chance possibility of some differences between groups in the achieved samples is always present. This can be exacerbated by subject non-response. To assess the equivalence of these samples among survey respondents, tables A2 and A3 in appendix A provide breakdowns of age, gender, job-type and force for the achieved treatment and control groups in the two surveys, relying on the survey data (though survey gaps in gender information were supplemented by force data).

Table 2. Experimental samples and survey responders

Pilot Force	Total							Control group							Treatment group						
	All trial officers	Survey responders						All control officers	Survey responders						All treat-ment officers	Survey responders					
		Wave 1		Wave 2		Both			Wave 1		Wave 2		Both			Wave 1		Wave 2		Both	
		n	%	n	%	n	%		n	%	n	%	n	%		n	%	n	%	n	%
A	220	139	63	94	43	80	36	110	67	61	40	36	33	30	110	72	65	54	49	47	43
B	224	158	71	117	52	102	46	112	73	65	56	50	49	44	112	85	76	61	54	53	47
C	220	192	87	130	59	123	56	111	96	86	62	56	60	54	109	96	88	68	62	63	58
D	220	178	81	99	45	95	43	110	85	77	43	39	42	38	110	93	85	56	51	53	48
E	220	155	70	107	49	100	45	110	64	58	38	35	35	32	110	91	83	69	63	65	59
F	219	160	73	99	45	95	43	108	69	64	41	38	40	37	111	91	82	58	52	55	50
All	**1323**	**982**	**74**	**646**	**49**	**595**	**45**	**661**	**454**	**69**	**280**	**42**	**259**	**39**	**662**	**528**	**80**	**366**	**55**	**336**	**51**

Both show that the aggregate treatment and control groups (ie, all six forces) are reasonably similar, notwithstanding a few differences (for example, appendix tables A2 and A3 indicate some modest differences in age between Wave 1 and 2 treatment and control groups, more pronounced in Wave 2). Larger differences are evident within individual force samples. For example, there are some notable age differences between treatment and control groups in Waves 1 and 2 for Force E. In large part, differences probably arise because the smaller sample sizes within forces allow chance to play a stronger role shaping group characteristics. Statistical tests help take account of this, by requiring greater effect sizes in smaller samples to achieve statistical significance.

Overall, we can say that we have been reasonably successful in producing a balanced treatment and control group in the two surveys, based on available measures.

Police stop and search data

A second source of data, available for all officers involved in the trial, was the electronic data generated by forces based on their officers' completion of search forms. These forms should be completed every time an officer searches a member of the public and are used to record:

- the grounds used to justify the search
- the power used
- the object of the search (eg, drugs, stolen property or a knife)
- the search outcome
- basic demographic information about the suspect (eg, officer-defined and self-defined ethnic/racial group).

The analysis focused on those searches that required officers to have reasonable grounds for suspicion. Other kinds of searches or contacts included in the original force databases were excluded during data cleaning prior to analysis. Unfortunately, there is no national requirement for police officers to record stops without searches, so data on stops that did not lead to searches were not available for analysis.

Basic measures

For most measures, we focused on records of searches conducted in the three-month period before and the three-month period after the training was implemented. The pre-test period for all forces included search records from June to August 2015, which were almost entirely before force training (with a few minor exceptions). The post-test period for five of the pilot forces was defined as October to December 2015, after the bulk of training in these forces had been completed. In Force F, which had extensive training running into October, the post-test period was defined as November 2015 to January 2016 to allow for this difference.

Core measures for the pre- and post-test period used in the analysis were:

- numbers of searches per officer
- arrest rates of searches[3]
- ethnic/racial characteristics of search suspects.

Genuine non-response was not an issue for the above measures, as police data were theoretically available for all officers in the trial (apart from a handful who had left the pilot forces, were on an extended period of leave or restricted duties, or had changed to a different role). Not all officers in the trial, however, carried out searches in the pre- and post-test periods. For arrest rates of searches and the ethnic/racial breakdown of searches, we therefore focused on searches produced only by officers completing searches, and considered the search as the primary unit of analysis (rather than the officer).

Measuring the quality of recorded grounds

The written grounds present in officers' completed search forms represent their formal account of their reasons for suspicion that justified the search they conducted. In order to assess whether training had improved the quality of these grounds, we developed quantitative measures from the grounds narrative. These indicated whether grounds included legitimate factors for suspicion based on PACE Code A. They were applied to the written grounds taken from search forms as they were recorded on force databases.

Through an initial testing exercise, using pre-pilot recorded grounds, researchers developed the coding frame that could be used to generate quantitative measures relating to the quality of written grounds. This framework was then applied by two researchers to force databases of recorded grounds in the post-pilot period. Grounds were assigned randomly to one or other of the coders who were blinded as to their association with treatment or control groups. A small proportion of grounds was also cross-coded by both coders (and the lead author) allowing for quality checking and control.

The final grounds variables used in this research, based on the coding frame developed, were a set of binary variables indicating which of the following legitimate factors were included in each of the written grounds:

- suspicious behaviours observed by police
- appearance of carrying something suspicious
- match with suspect description/intelligence on suspect
- directly identified by victim/witness
- suspicious vehicle
- drug intoxication
- drugs/drug paraphernalia witnessed by officers
- location/temporal factors link suspect to crime risk
- seen with suspicious associates

[3] Other outcomes (eg, fixed penalty notices) were not recorded consistently across the six forces.

- self-incrimination during conversation/questioning
- group/gang membership
- other.

More details of the coding strategy and coding frame are found in appendices H and I.

Analysis strategy

The analysis followed an 'intention to treat' method in which all the officers assigned to the treatment and control groups were compared regardless of their ultimate participation (or non-participation) in the training pilot. This method avoids any selection bias that would otherwise arise by excluding cases that did not adhere to the training protocol; the latter might be systematically different from those who participated in the training.

Most analyses relied on simple comparisons of officer-level averages or 'means' between treatment and control groups (both at the overall programme level, and the individual force level). For these comparisons, effect sizes were calculated in the form of Cohen's d, which assesses the difference in means after dividing by the standard deviation across respondents. Significance tests for force-level comparisons relied commonly on t-tests, while for comparisons of the programme-level six-force sample, significance was assessed using multiway analyses of variance (ANOVA) that controlled for binary (or dummy) variables representing the different forces. These controls provide more accurate significance results, because they take account of the force as a primary stratifying unit within the random assignment (Bloom, 2006). They also help adjust for any clustering of values within forces.

A variety of other bivariate and multivariate strategies were, however, used for measures that did not fit with mainstream parametric test assumptions. Thus, the analysis also included count models and non-parametric tests of associations, such as the Mann-Whitney U test and its variants.

Basic comparisons and more complex multivariate analyses can be seen in appendices E, F and G.

Limitations

Random assignment to treatment and control groups, as part of an RCT, provides the strongest method available to researchers to draw conclusions about the causal impact of an intervention on a set of outcomes. That is why this evaluation used an RCT, and why we are confident, overall, in the study design. Nonetheless, it important to highlight some inevitable limits to our confidence in the results reported (see also Quinton, 2016).

First, it is possible this study, like any RCT, may produce some misleading results because random assignment sometimes produces differences in treatment and control groups purely as a product of chance. Secondly, when RCTs involve subject attrition—as we have here because some officers declined to respond to surveys—this too can produce differences between treatment and control groups that may bias conclusions. Our analysis of characteristics of treatment and control groups,

including the subsets responding to surveys, does not provide strong cause for concern but we cannot entirely rule out the possibility that differences exist.

Additionally, in the presentation of our results, we place primary emphasis on treatment effects that achieve 'statistical significance'. An inherent problem when relying on multiple tests of statistical significance, however, is that a small minority of results will be statistically significant by chance alone in the absence of an underlying relationship. Conversely, some real relationships between variables may not be detected through significance testing, either because their effects are too small or because chance variation hides evidence of their effects. These possibilities should also be taken into account when reviewing results.

Results reporting

In the findings presented in this report, we focus primarily on the aggregate programme-level effects of training across the six participating forces. Individual force-level results, however, are also presented to provide context to the aggregate results and basic differences between forces are discussed.

3. Impacts on officers' preparation and knowledge

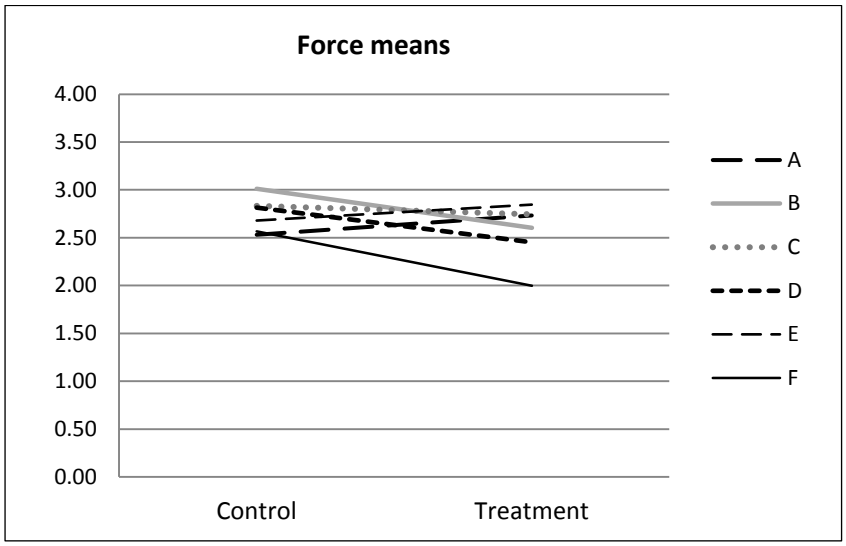
The training was hypothesised to have effects on officers' knowledge and how prepared they felt in relation to stop and search. Our results showed some impacts on these outcomes. In particular, there was evidence of sustained improvements in trained officers' knowledge of stop and search policies. Additionally, treatment group officers were less likely to see weaker written grounds as adequate to support a search, although this was also evident for stronger written grounds. Despite these improvements, trained officers had a more negative impression of the stop and search training they had previously received. Results also showed some variations across individual forces. For example, Forces A and D showed the most pronounced effects of training on knowledge and Force B registered no clear effects of training on knowledge or assessment of grounds.

Assessment of prior training

We hypothesised that the pilot would produce a more positive assessment of officers' prior experience of training in stop and search, whether from pilot training or from other training experiences. Figure 1 assesses this hypothesis by presenting results for the scale: **perceived contribution of prior training to stop and search knowledge**, measured in the Wave 1 survey. This scale is based on statements such as 'Training has helped me identify situations when it is legal to conduct a search' and 'I know what I'm required to say to somebody I'm searching because of the training I've received.'

Contrary to our hypothesis, figure 1 indicates a relatively small **negative** overall programme-level treatment effect on the reported contribution of training to officers' knowledge on stop and search. This means that officers who were trained in the pilot were less enthusiastic about their overall stop and search training experience. This finding is driven by significant negative effects in three out of five forces (Forces B, D and F).

Figure 1. Pilot training effects on perceived contribution of prior training to stop and search knowledge (0–4 scale), Wave 1 survey

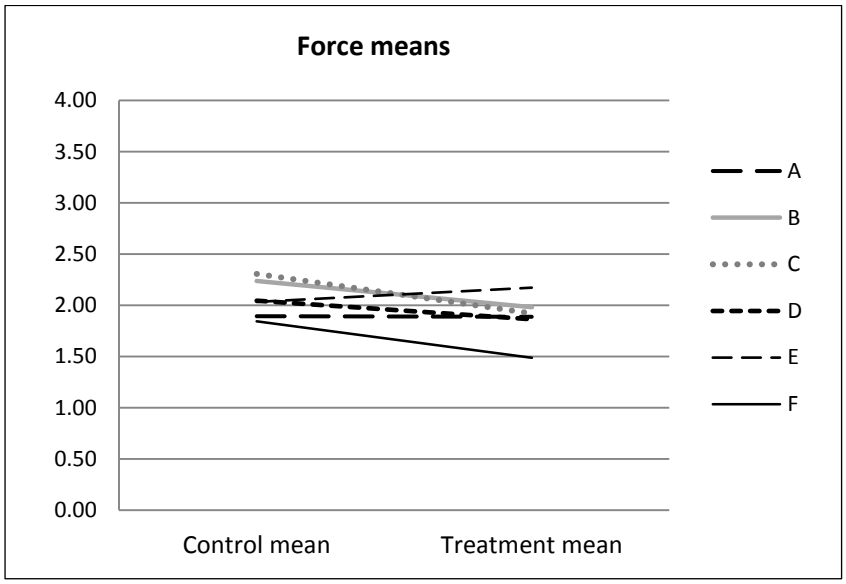


Force	Control means	Treatment means	Effect size	Sig.
A	2.53	2.73	0.29	
B	3.01	2.60	-0.57	***
C	2.83	2.74	-0.14	
D	2.81	2.45	-0.42	**
E	2.68	2.85	0.23	
F	2.57	2.00	-0.59	***
All	2.75	2.56	-0.24	***

Notes: Effect size is measured using Cohen's d; 0.2 is considered 'small', 0.5 'medium' and 0.8 'large' (Cohen, 1988). Force-level significance levels relied on t-tests. Whole sample significance levels relied on two-way ANOVAs that included binary force dummy variables alongside a treatment factor.
* p<0.05, ** p<0.01, *** p<0.001

Figure 2 presents further results for the **perceived contribution of prior training to interpersonal skills** scale. This scale uses statements such as 'Training has prepared me to deal with difficult interactions with the public' and 'I've received training that has helped me think about how I talk to the people I stop and search.' The scores are lower than the previous scale on knowledge, with force means around the middle of the scale.

Figure 2. Pilot training effects on perceived contribution of prior training to interpersonal skills (0–4 scale)



Force	Control means	Treatment means	Effect size	Sig.
A	1.89	1.89	-0.01	
B	2.24	1.98	-0.31	
C	2.31	1.92	-0.47	**
D	2.05	1.86	-0.22	
E	2.03	2.17	0.16	
F	1.84	1.49	-0.39	*
All	2.08	1.89	-0.22	***

Notes: Effect size is measured using Cohen's d; 0.2 is considered 'small', 0.5 'medium' and 0.8 'large' (Cohen, 1988). Force-level significance levels relied on t-tests. Whole sample significance levels relied on two-way ANOVAs that included binary force dummy variables alongside a treatment factor.
* p<0.05, ** p<0.01, *** p<0.001

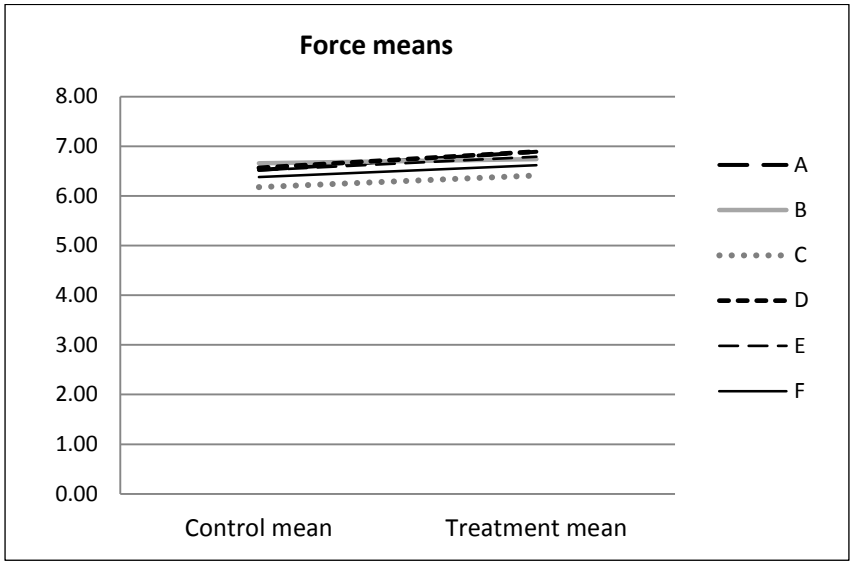
Again, contrary to our hypothesis, figure 2 indicates a negative overall programme-level treatment effect, suggesting that officers who were trained in the pilot tended to be less enthusiastic about its contribution to their interpersonal skills. This was largely driven by significant negative effects in Forces C and F.

Knowledge of stop and search policy

We hypothesised that training would improve officers' knowledge of stop and search policies. Officers' knowledge was, therefore, measured in both surveys. Figure 3 presents findings on **knowledge of stop and search policies** in the Wave 1 survey, conducted within a few days or weeks of the training. This is a summary score, based on responses to a series of true/false statements focused on stop and search policies. It included statements primarily regarding PACE Code A requirements such as 'The law allows me to search somebody because they have recently been charged with a crime' and 'I am normally required to give my name, number and station before searching somebody.'

As figure 3 illustrates, there were modest statistically significant positive effects at the programme level, in line with expectations. At the force level, all effects were also positive, although only statistically significant for Forces A and D. Officer scores were, however, high overall (averaging over six out of eight within force treatment and control groups), which suggested officers were generally very knowledgeable about stop and search policies, even in the absence of the pilot training.

Figure 3. Pilot training effects on knowledge of stop and search policy (0–8 scale), Wave 1 survey

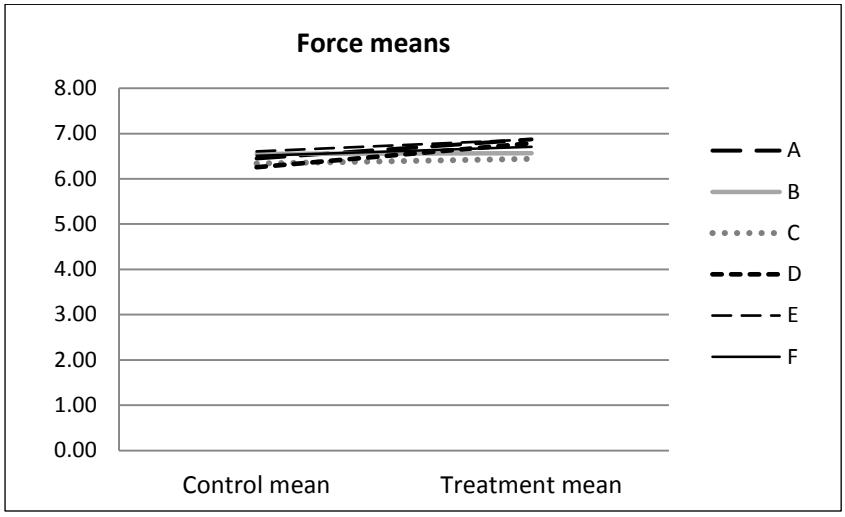


Force	Control means	Treatment means	Effect size	Sig.
A	6.52	6.89	0.37	*
B	6.66	6.74	0.08	
C	6.18	6.41	0.22	
D	6.56	6.89	0.35	*
E	6.52	6.79	0.26	
F	6.38	6.62	0.21	
All	6.46	6.71	0.25	***

Notes: Effect size is measured using Cohen's d; 0.2 is considered 'small', 0.5 'medium' and 0.8 'large' (Cohen, 1988). Force-level significance levels relied on t-tests. Whole sample significance levels relied on two-way ANOVAs that included binary force dummy variables alongside a treatment factor.
* p<0.05, ** p<0.01, *** p<0.001

The Wave 2 survey, conducted approximately three months after training, used the same knowledge measure. Results, presented in figure 4, suggest that training effects on knowledge were sustained over the period. Again, there was a modest but statistically significant effect of treatment at the programme level, echoed in statistically significant effects in Forces A and D.

Figure 4. Pilot training effects on knowledge of stop and search policy (0–8 scale), Wave 2 survey



Force	Control means	Treatment means	Effect size	Sig.
A	6.45	6.87	0.43	*
B	6.55	6.57	0.01	
C	6.34	6.44	0.09	
D	6.26	6.79	0.49	*
E	6.61	6.87	0.30	
F	6.51	6.71	0.18	
All	6.45	6.70	0.25	**

Notes: Effect size is measured using Cohen's d; 0.2 is considered 'small', 0.5 'medium' and 0.8 'large' (Cohen, 1988). Force-level significance levels relied on t-tests. Whole sample significance levels relied on two-way ANOVAs that included binary force dummy variables alongside a treatment factor.
* p<0.05, ** p<0.01, *** p<0.001

In appendix E, additional analysis is presented on the **knowledge of PACE Code A** score. This score is based on answers to seven out of the eight statements included in the **knowledge of stop and search policy** variable above. It excluded, however, the statement 'The "reasonable grounds" threshold is exactly the same for a search and an arrest'. This was not part of PACE Code A at the time of the training pilot (though it was part of earlier iterations). As in the prior measure, at Wave 1 and 2, there were statistically significant programme-level treatment effects. At Wave 2, Forces D and E also showed statistically significant effects not seen in Wave 1. Overall, these additional findings further support the idea that training had an overall effect on stop and search knowledge that was sustained in the medium term.

Applying knowledge to written grounds

The research hypothesised that trained officers would be more able to recognise when grounds for suspicion were inadequate, given that they would be able to apply knowledge and skills learned during training. The Wave 2 survey, therefore, asked officers to evaluate examples of five written grounds for searches. Two of these examples were constructed to have stronger grounds and three

had weaker grounds. We hypothesised that treatment group officers would, as a result of training, express less confidence in the weaker written grounds than the control group officers.

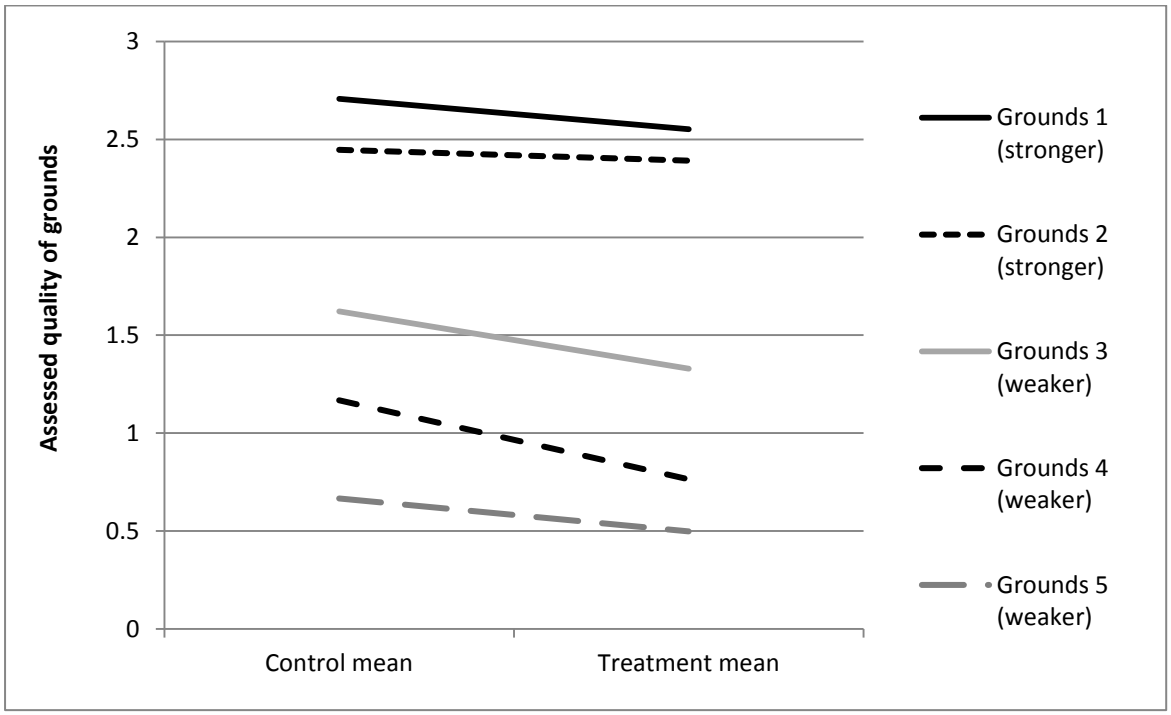
Specifically, officers were asked to assess the quality of written grounds by indicating whether they were 'definitely good enough', 'probably good enough', 'probably not good enough' or 'definitely not good enough'. Programme-level findings (ie, including all forces) are presented in figure 5, with the grounds assessment presented as a mean 0 to 3 score, with 0 meaning 'definitely not good enough' and 3 indicating 'definitely good enough'. The figure also presents the wording of each of the weaker and stronger written grounds that officers were asked to assess.

The figure indeed shows that treatment group officers were less confident in weaker grounds than control group officers, with all effects statistically significant. There were also unexpected negative effects on one of the two stronger grounds (the second was also in the same direction, but did not achieve statistical significance). Nevertheless, the treatment effects appeared **larger** for the weaker grounds and multivariate models presented in appendix G provide borderline statistical support for this idea.[4]

Individual forces results (reported in tables E4 to E9 in appendix E) also indicate mostly negative effects of treatment on the evaluation of weaker grounds, meaning that trained officers reported less confidence in them than untrained officers. Specifically, all forces except Force B showed one or more statistically significant effects of this kind. For stronger grounds, treatment effects tended to be less pronounced or absent. Only Forces A and C registered one or more statistically significant effects on stronger grounds. This again indicated that trained officers were less confident in the adequacy of written grounds than untrained officers.

[4] Further statistical testing using an ordered logistic regression suggested a programme-level interaction of p=0.057 between grounds strength and treatment, suggesting that training may have had greater negative effects on the assessment of weaker grounds (although results fell short of conventional significance levels). Force-specific models for individual forces showed non-significant interaction effects, except the model for Force E, which indicated a similar negative effect where treatment coincided with weaker grounds (p=0.004) (see table G6 in appendix G).

Figure 5. Pilot training effects on officer assessed strength of five different written 'grounds for suspicion' (0–3 value item), Wave 2 survey



Grounds Strength	Wording	Control mean	Treatment mean	Sig.
1. Stronger	Observed for a few minutes, seen looking into cars, hanging around for no reason, fidgeting and patting pocket, avoided police, concealed items on seeing police (Section 1, PACE)	2.71	2.55	***
2. Stronger	Male seen hanging around, not moving on, in location of known drug supply. Walked away when saw police. Fits suspect description of active drug dealer (distinctive clothing, facial features). (Section 23, Misuse of Drugs Act)	2.45	2.39	
3. Weaker	Searched for going equipped items because was in area with recent burglaries. Was walking at night with an individual with previous for vehicle break-ins. Evasive during questioning. (Section 1, PACE)	1.62	1.33	***
4. Weaker	Seen in drug hotspot talking to known drug dealer. (Section 23, Misuse of Drugs Act)	1.17	0.76	***
5. Weaker	Call that robbery suspect was in park (Section 1, PACE)	0.67	0.50	**

Notes: No effect sizes were calculated because response categories were single four-level items, not suitable for conventional effect size calculations. Significance tests were based on Van Eltern tests, a non-parametric Mann-Whitney U test adapted for stratified samples such as this one. The test was stratified by police force.
* p<0.05, ** p<0.01, *** p<0.001

4. Impacts on officers' attitudes

In this chapter, we examine the effects of the pilot training on officers' attitudes that are relevant to the exercise of stop and search powers. We hypothesised that the training would produce attitudes more favourable to good practice and the chapter finds some support for this outcome. As we go on to demonstrate, the research showed that treatment group officers reported slightly less support for police stereotyping and a lower level of support for high volume stop and search strategies compared to the control group. Treatment group officers also had less cynicism towards the regulation of stop and search, although this effect was not sustained. The research, however, showed no meaningful differences between the groups in officers' support for procedural justice in stop and search. Consistent with this, the process showed in fact that procedural justice was not a central feature of the training as delivered (see Giacomantonio et al, 2016). Some force-specific differences were also evident. Notably, Force D stands out for registering strong and significant effects of training on a variety of attitude measures, while Forces C and E registered no clear effects.

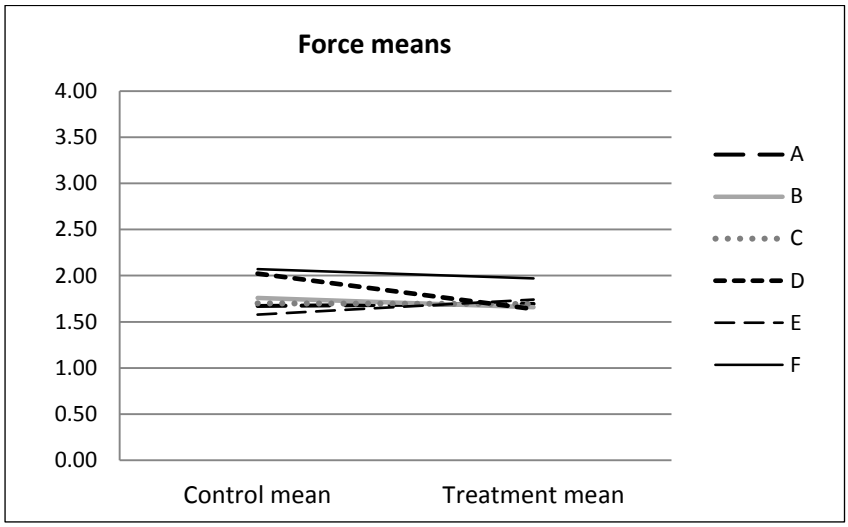
Support for stereotyping

We hypothesised that training would reduce the level of officer stereotyping. Therefore, in Wave 1 survey only, we created two scales to measure aspects of stereotyping in relation to police suspicion.

The first of these measured **officers' support for police non-ethnic/racial stereotyping** developed from officers' agreement with statements such as 'Officers should pay attention to the way people dress, because this provides clues as to whether they are involved in crime' and 'Officers should give more attention to people they know have committed crime regularly in the past.'

The results from this first measure are presented in figure 6. They indicate no statistically significant programme-level effects of treatment on non-ethnic/racial stereotyping. Nonetheless, Force D showed a substantial and significant negative treatment effect in line with expectations (ie, training appeared to reduce support for stereotyping). It is also notable that scores were consistently at, or lower than, the mid-point of the scale in both groups, suggesting only limited support for stereotyping regardless of training.

Figure 6. Pilot training effects on support for police non-ethnic/racial stereotyping (0–4 scale), Wave 1



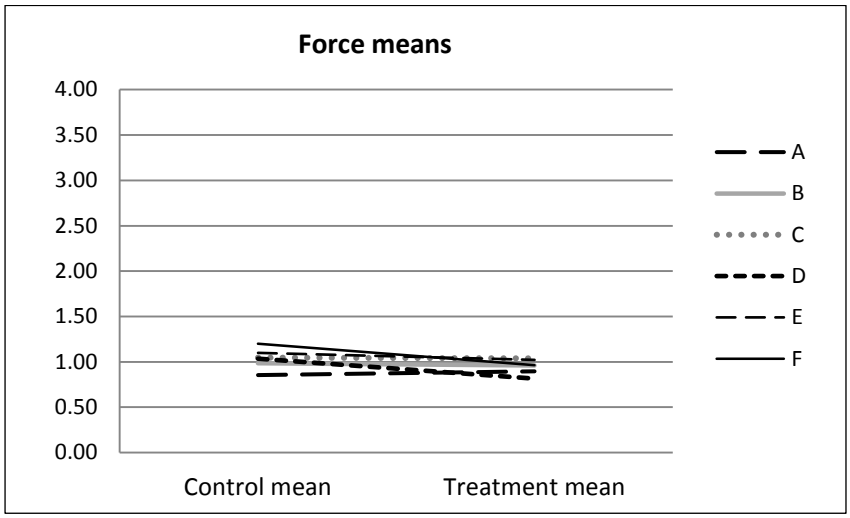
Force	Control means	Treatment means	Effect size	Sig.
A	1.67	1.69	0.03	
B	1.76	1.66	-0.13	
C	1.70	1.69	-0.01	
D	2.02	1.64	-0.46	**
E	1.58	1.74	0.22	
F	2.07	1.97	-0.14	
All	1.80	1.73	-0.09	

Notes: Effect size is measured using Cohen's d; 0.2 is considered 'small', 0.5 'medium' and 0.8 'large' (Cohen, 1988). Force-level significance levels relied on t-tests. Whole sample significance levels relied on two-way ANOVAs that included binary force dummy variables alongside a treatment factor.
* p<0.05, ** p<0.01, *** p<0.001

The second stereotyping scale measured **officers' support for police ethnic/racial stereotyping**. This was developed from officer agreement with statements such as 'When tackling street drug-dealing, it makes sense for officers to pay greater attention to young males from black and Asian backgrounds because they're more likely to be involved' and 'Officers should pay more attention to young black males when policing street robbery because they're more often involved in this type of crime.'

Figure 7 presents findings for this variable. It shows a very small but statistically significant, negative programme-level treatment effect in line with expectations (ie, training appeared to reduce ethnic/racial stereotyping). This was driven in particular by Forces D and F, which showed notable and statistically significant negative treatment effects (ie, an apparent reduction in support for stereotyping). It also showed that support for ethnic/racial stereotyping was generally very low (around one out of four points) across both groups.

Figure 7. Pilot training effects on support for police ethnic/racial stereotyping (0–4 scale), Wave 1



Force	Control means	Treatment means	Effect size	Sig.
A	0.85	0.90	0.06	
B	0.99	0.96	-0.04	
C	1.05	1.04	-0.01	
D	1.04	0.82	-0.31	*
E	1.10	1.02	-0.10	
F	1.20	0.96	-0.36	*
All	1.04	0.95	-0.12	*

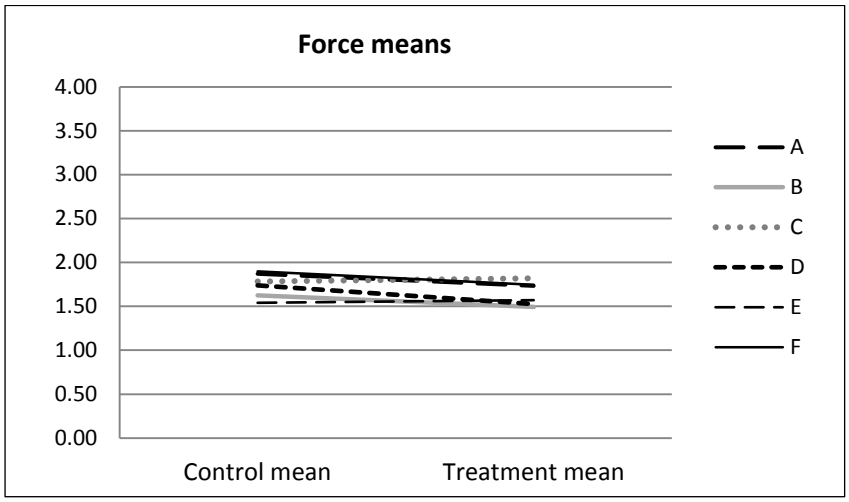
Notes: Effect size is measured using Cohen's d; 0.2 is considered 'small', 0.5 'medium' and 0.8 'large' (Cohen, 1988). Force-level significance levels relied on t-tests. Whole sample significance levels relied on two-way ANOVAs that included binary force dummy variables alongside a treatment factor.
* p<0.05, ** p<0.01, *** p<0.001

Cynicism toward the regulation of stop and search

Recognising that the pilot training sought to challenge wider attitudes in police culture about the regulation of police behaviours, it was hypothesised that the pilot training would result in more positive attitudes to the formal regulation of stop and search. Both Wave 1 and 2 surveys, therefore, measured **officers' cynicism toward the regulation of stop and search,** a scale comprised of statements such as 'An officer's gut feelings are more valuable than official rules when it comes to knowing when to search someone' and 'The bureaucrats who write stop and search policies are out of touch with how things work on the street.'

Findings from Wave 1 are presented in figure 8. They show a small significant negative programme-level treatment effect (ie, the training appeared to reduce cynicism). This was also echoed in a statistically significant negative effect in Force D. In general, the scale showed a moderate level of cynicism about the regulation of searches, with scores in general slightly below the middle of the scale.

Figure 8. Pilot training effects on cynicism toward the regulation of stop and search (0–4 scale), Wave 1

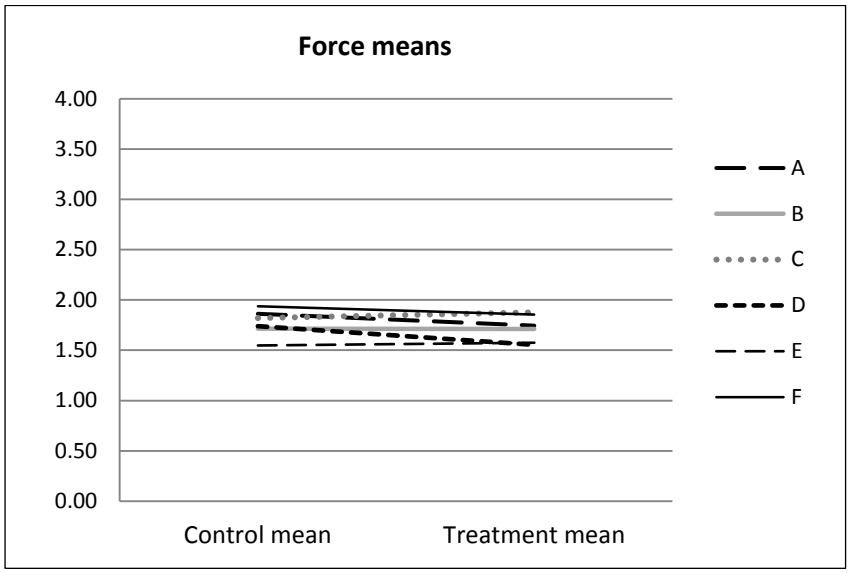


Force	Control means	Treatment means	Effect size	Sig.
A	1.87	1.73	-0.20	
B	1.63	1.49	-0.23	
C	1.78	1.82	0.06	
D	1.74	1.53	-0.34	*
E	1.54	1.57	0.05	
F	1.90	1.74	-0.25	
All	1.75	1.65	-0.16	*

Notes: Effect size is measured using Cohen's d; 0.2 is considered 'small', 0.5 'medium' and 0.8 'large' (Cohen, 1988). Force-level significance levels relied on t-tests. Whole sample significance levels relied on two-way ANOVAs that included binary force dummy variables alongside a treatment factor.
* p<0.05, ** p<0.01, *** p<0.001

Figure 9 presents results for the same measure (**cynicism toward the regulation of stop and search**) for the Wave 2 survey. The results here indicate no statistically significant programme-level or force-level effects, which would suggest the benefits of training on this variable largely subsided with the passage of time.

Figure 9. Pilot training effects on cynicism toward the regulation of stop and search (0–4 scale), Wave 2



Force	Control means	Treatment means	Effect size	Sig.
A	1.86	1.74	-0.18	
B	1.71	1.71	0.00	
C	1.82	1.88	0.09	
D	1.74	1.55	-0.28	
E	1.55	1.57	0.04	
F	1.94	1.86	-0.12	
All	1.77	1.72	-0.08	

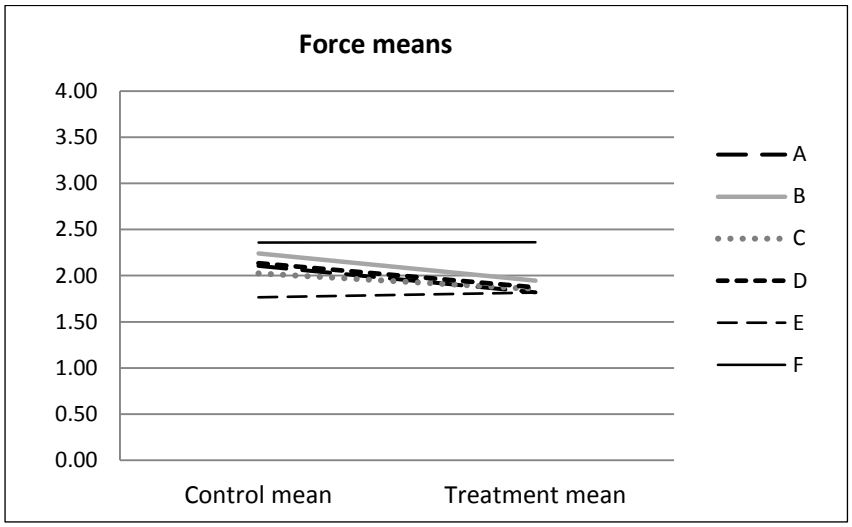
Notes: Effect size is measured using Cohen's d; 0.2 is considered 'small', 0.5 'medium' and 0.8 'large' (Cohen, 1988). Force-level significance levels relied on t-tests. Whole sample significance levels relied on two-way ANOVAs that included binary force dummy variables alongside a treatment factor.
* p<0.05, ** p<0.01, *** p<0.001

Support for high volume stop and search strategies

Also focusing on the pilot training's effects on wider police culture, the research tested the hypothesis that training would reduce support for high volume stop and search strategies. Surveys at Wave 1 and Wave 2 included the scale **support for high volume stop and search strategies.** This scale incorporates statements including 'The police can reduce crime simply by doing more stop and search' and 'Police need to conduct searches in large numbers to keep some order on the streets.'

Results for this scale at Wave 1 are presented in figure 10. They show a small statistically significant negative programme-level effect, indicating that trained officers were less supportive of high volume stop and search strategies and suggesting they favoured a more selective approach instead. Three forces (A, B and D) also showed statistically significant negative effects in line with expectations.

Figure 10. Pilot training effects on support for high volume stop and search strategies (0–4 scale), Wave 1

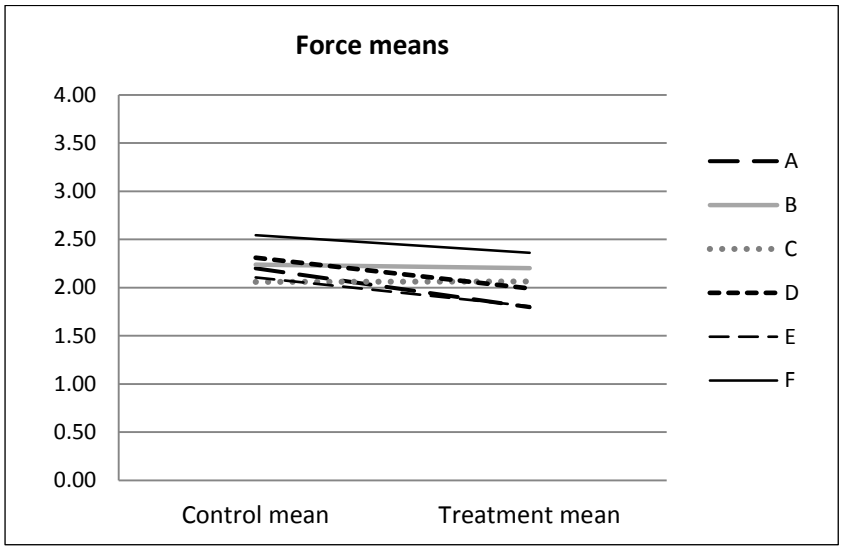


Force	Control means	Treatment means	Effect size	Sig.
A	2.11	1.82	-0.34	*
B	2.24	1.95	-0.38	*
C	2.02	1.85	-0.21	
D	2.14	1.87	-0.31	*
E	1.77	1.82	0.07	
F	2.36	2.36	0.00	
All	2.11	1.95	-0.19	**

Notes: Effect size is measured using Cohen's d; 0.2 is considered 'small', 0.5 'medium' and 0.8 'large' (Cohen, 1988). Force-level significance levels relied on t-tests. Whole sample significance levels relied on two-way ANOVAs that included binary force dummy variables alongside a treatment factor.
* p<0.05, ** p<0.01, *** p<0.001

Results for the same scale at Wave 2 are presented in figure 11. These indicate a similarly strong programme-level effect at Wave 2 as at Wave 1, suggesting that the impacts of training on this support for high volume stop and search strategies persisted over time. At the force level, however, only Force A showed a statistically significant treatment effect, which was in line with expectations.

Figure 11. Pilot training effects on support for high volume stop and search strategies (0–4 scale), Wave 2



Force	Control means	Treatment means	Effect size	Sig.
A	2.20	1.80	-0.47	*
B	2.24	2.20	-0.06	
C	2.06	2.06	0.01	
D	2.31	1.99	-0.41	
E	2.11	1.81	-0.36	
F	2.54	2.36	-0.21	
All	2.23	2.03	-0.24	**

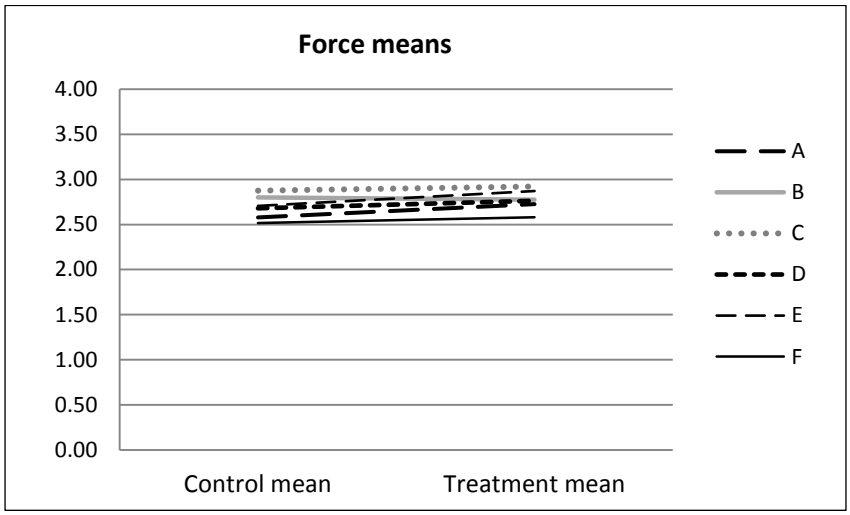
Notes: Effect size is measured using Cohen's d; 0.2 is considered 'small', 0.5 'medium' and 0.8 'large' (Cohen, 1988). Force-level significance levels relied on t-tests. Whole sample significance levels relied on two-way ANOVAs that included binary force dummy variables alongside a treatment factor.
* p<0.05, ** p<0.01, *** p<0.001

Support for procedural justice in stop and search

Finally, we hypothesised that training would increase officers' support for procedural justice in relation to stop and search. To this end, the Wave 1 survey included a **support for procedural justice in stop and search** scale, drawing on agreement with statements such as 'People will show more respect for the law if they think we listen to their opinions during a stop and search' and 'Treating someone who's angry during a stop and search with respect increases the community's confidence in the police.'

Figure 12 presents results for this scale. Contrary to expectations, it indicates the treatment was not associated with any significant change in levels of support for procedural justice. This was true at the programme level and at the individual force level. In both groups, however, there was a moderately high level of support for procedural justice. Consistent with this, the companion process evaluation has shown that procedural justice was not a central feature of the training that was delivered in the forces (see Giacomantonio et al, 2016).

Figure 12. Pilot training effects on support of procedural justice in stop and search (0–4 scale), Wave 1 survey



Force	Control means	Treatment means	Effect size	Sig.
A	2.58	2.73	0.16	
B	2.80	2.78	-0.04	
C	2.88	2.92	0.06	
D	2.68	2.76	0.10	
E	2.71	2.87	0.19	
F	2.51	2.58	0.07	
All	2.70	2.78	0.09	

Notes: Effect size is measured using Cohen's d; 0.2 is considered 'small', 0.5 'medium' and 0.8 'large' (Cohen, 1988). Force-level significance levels relied on t-tests. Whole sample significance levels relied on two-way ANOVAs that included binary force dummy variables alongside a treatment factor.
* p<0.05, ** p<0.01, *** p<0.001

5. Impacts on officers' anticipated behaviours

In this chapter, we examine how officers said they would act in response to a variety of scenarios related to stop and search. Study hypotheses anticipated improvements in behaviours. Some of these, though not all, were borne out in study findings. In particular, the chapter shows that there were no programme-level differences between treatment and control group in how officers said they would treat a suspect, in respect to legal or procedural justice principles. There were, however, substantial training effects on whether officers said they would search a suspect across scenarios, although this applied to situations both with stronger and weaker grounds. It seemed to happen because trained officers were less likely to evaluate grounds as adequate for a search and not because they reported being less motivated to conduct searches when grounds were present. As expected, the effect was not present when officers were asked whether they would question suspects, suggesting training affected **how** officers said they would intervene in a situation rather than **whether** they said they would intervene or not.

Finally, by comparing officers' responses to scenarios according to the ethnic/racial appearance of suspects, it was evident that training had no effects on the relevance of ethnicity/race for decision-making. Irrespective of training, officers were generally **less** likely to say they would question or search black suspects compared to white.

Individual forces showed some variations in effects. Notably, in Force D the training increased officers' anticipated use of procedural justice and legal principles in their treatment of a suspect, while it reduced this for Force B. Force E stood out for registering no clear effects on any measures assessed.

How a suspect would be treated during a search encounter

We hypothesised that trained officers would say they are more inclined to follow legal procedure and to treat people according to procedural justice principles when conducting stop and search. To address their anticipated treatment of suspects, we used two measures in the Wave 1 survey based on responses to a scenario that involved dealing with a confrontational suspect during a stop and search encounter. [5]

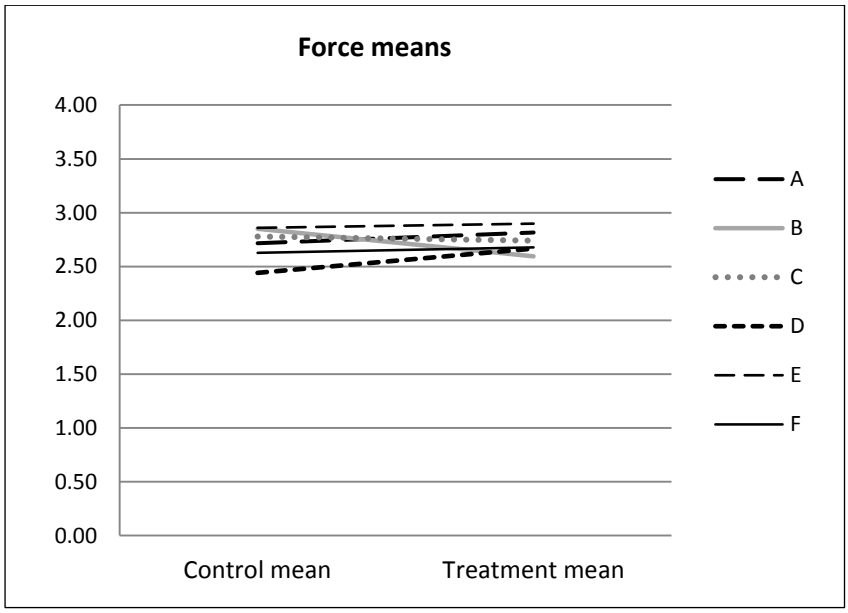
Procedurally just treatment of a suspect

The first scale, measuring the **procedurally just treatment of a stop and search suspect,** was based on how much priority officers say they would give to considerations such as 'Acknowledging the suspect's feelings' and 'Letting the suspect tell their side of the story' when engaging in a stop and search with a confrontational suspect.

[5] The scenario read: 'Imagine you have seen a man looking into some parked cars about 10pm in the evening. He appears to be holding a tool in the sleeve of his jacket. You have stopped him and asked him what he is doing. He becomes angry and starts shouting. He says you are picking on him for no reason. He says he is just taking a walk and he wants to be left alone. You decide to conduct a search.'

Figure 13 presents results for this scale. It indicates no programme-level effect of treatment. At the individual force level, there were two statistically significant effects, but in opposing directions (which cancelled each other out at the programme level). While Force D experienced a positive impact of treatment (indicating that training made officers more likely to say they would treat a suspect in a procedurally just way), Force B experienced a negative effect (suggesting training made them less likely to indicate procedurally just treatment). Overall, however, scores were in the upper rather than the lower range of the scale suggesting that officers are generally inclined towards procedurally just treatment of suspects, irrespective of training.

Figure 13. Pilot training effects on procedurally just treatment of a stop and search subject (0–4 scale), Wave 1 survey



Force	Control means	Treatment means	Effect size	Sig.
A	2.72	2.82	0.14	
B	2.85	2.60	-0.38	*
C	2.78	2.74	-0.05	
D	2.44	2.67	0.31	*
E	2.86	2.90	0.05	
F	2.63	2.68	0.07	
All	2.71	2.73	0.03	

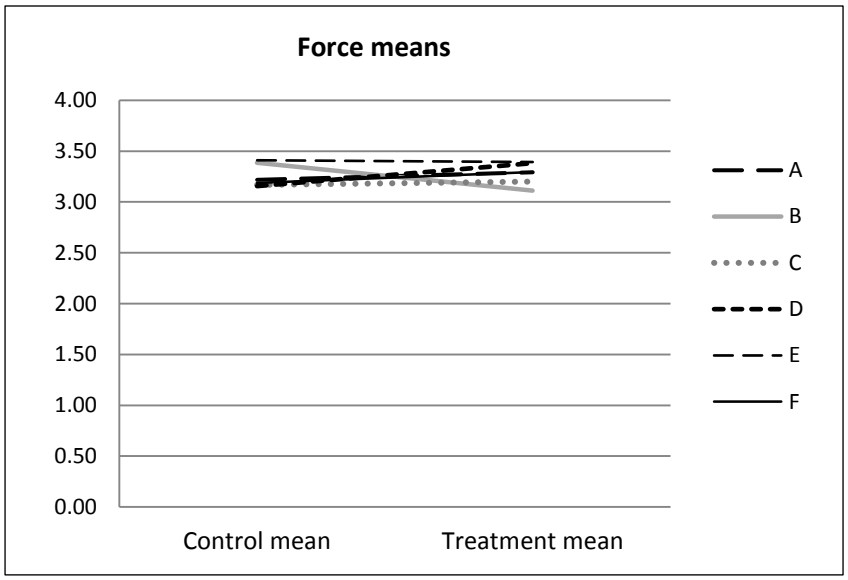
Notes: Effect size is measured using Cohen's d; 0.2 is considered 'small', 0.5 'medium' and 0.8 'large' (Cohen, 1988). Force-level significance levels relied on t-tests. Whole sample significance levels relied on two-way ANOVAs that included binary force dummy variables alongside a treatment factor.
* p<0.05, ** p<0.01, *** p<0.001

Legal treatment of a suspect

In figure 14, we present findings for the **legal treatment of a stop and search suspect.** This scale drew on the priority officers said they would give, in the same scenario as before, to legal requirements when searching, including 'Explaining why you are conducting a search' and 'Telling the suspect what you are searching for.'

The figure again shows that treatment had no programme-level statistically significant effects. Again, at the force level, effects were mixed: Force D experienced a statistically significant positive effect of treatment and Force B experienced a statistically significant negative effect. These findings must be set against the generally high scores on the scale, measuring about three out of four, irrespective of force or treatment group.

Figure 14. Pilot training effects on legal treatment of a stop and search subject (0–4 scale), Wave 1 survey



Force	Control means	Treatment means	Effect size	Sig.
A	3.22	3.29	0.11	
B	3.39	3.11	-0.40	*
C	3.17	3.20	0.04	
D	3.16	3.38	0.32	*
E	3.41	3.39	-0.02	
F	3.19	3.29	0.14	
All	3.25	3.28	0.05	

Notes: Effect size is measured using Cohen's d; 0.2 is considered 'small', 0.5 'medium' and 0.8 'large' (Cohen, 1988). Force-level significance levels relied on t-tests. Whole sample significance levels relied on two-way ANOVAs that included binary force dummy variables alongside a treatment factor.
* p<0.05, ** p<0.01, *** p<0.001

Stop and search decisions and the strength of grounds

We hypothesised that training would make officers less likely to conduct a search when the grounds for doing so are weak. To investigate officers' anticipated decision-making in relation to stop and search, and to test this hypothesis, we examined whether the pilot training affected officers' responses to a range of potential stop and search scenarios.

Wave 1 survey

In the Wave 1 survey, a series of scenarios was given to each officer. The scenarios varied in terms of the situation (ie, suspected robbery or drug offence) and the strength of grounds (ie, weaker or stronger). For each scenario, officers were asked how likely they would be, first, to question the suspect(s) and, then, to search them (see appendix D for more details).

Table 3 summarises officers' estimated likelihood of initiating each type of action for both treatment and control groups in the whole six-force sample.

Table 3. Officer indicated probabilities of questioning or searching by crime type and grounds strength for both treatment and control groups (all forces combined), Wave 1 survey

Possible crime type	Strength of grounds	Type of encounter	Mean estimated probability of encounter (%)		Effect size	Sig.
			Control	Treatment		
Robbery	Weaker	Question	69.7	67.8	-0.06	
		Search	42.3	31.2	-0.35	***
	Stronger	Question	90.2	88.3	-0.09	
		Search	66.5	57.5	-0.27	***
Drugs	Weaker	Question	70.3	68.0	-0.07	
		Search	80.0	53.9	-0.78	***
	Stronger	Question	86.6	86.7	0.00	
		Search	66.3	58.4	-0.24	***

Notes: Effect size is measured using Cohen's d; 0.2 is considered 'small', 0.5 'medium' and 0.8 'large' (Cohen, 1988). Force-level significance levels relied on t-tests. Whole sample significance levels relied on two-way ANOVAs that included binary force dummy variables alongside a treatment factor.
* p<0.05, ** p<0.01, *** p<0.001

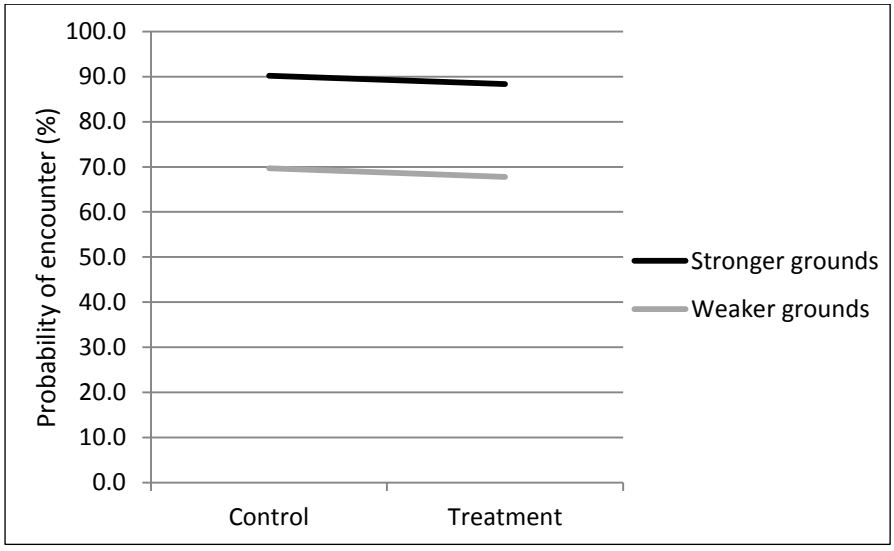
Overall, the table shows no significant programme-level changes associated with the training pilot in the willingness of officers to question suspects. The result was not a surprise; we had not predicted any change as officers were still expected to stop people if they were suspicious.

There were, however, notable statistically significant reductions in officers' declared likelihood of conducting searches for both stronger and weaker grounds and across crime types. This was in line with our expectations, particularly for the weaker grounds scenarios. To better illuminate treatment effects, figures 15 to 18 illustrate variations in effects across the scenarios.

Figure 18 shows something particularly interesting. In the control group, officers were more inclined to say they would search for drugs in the 'weaker grounds' scenario (involving the smell of cannabis) than in the 'stronger grounds' scenario (involving an exchange between two people in a drug-dealing area). In the treatment group, the opposite was true. This result would suggest that searches prompted by the smell of cannabis were normally very common and that training produced a large reduction in these searches, at least in terms of officers' anticipated behaviour. This substantial effect may follow from the fact the pilot training placed direct emphasis on the smell of cannabis in isolation as constituting inadequate grounds for a search (Giacomantonio et al, 2016).

Multivariate statistical models (presented in table G1, appendix G) confirmed the disproportionate effect of the treatment on officers' decision-making in relation to weaker (rather than stronger) drug searches, an effect not found for other scenario variations (ie, drug questioning, robbery questioning and robbery searching).[6]

Figure 15. Officers' indicated probabilities (means) of questioning suspects in relation to robbery, stronger and weaker grounds, Wave 1 survey



[6] Specifically, multivariate models measured the effects of an interaction term between grounds and treatment. For drug search scenarios, this interaction term was statistically significant ($p<0.001$) in its prediction of officers' declared inclination to conduct a search. It was not significant for other models.

Figure 16. Officers' indicated probabilities (means) of searching suspects in relation to robbery offences, stronger and weaker grounds, Wave 1 survey

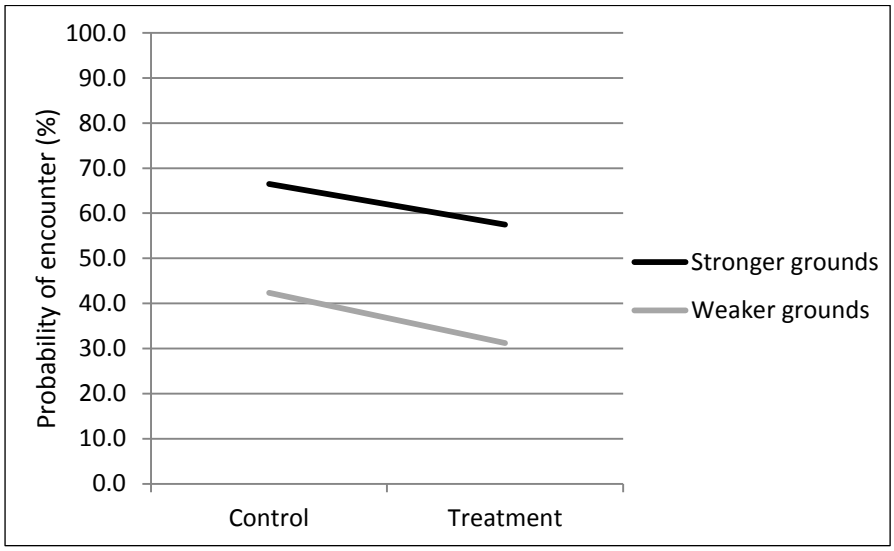


Figure 17. Officers' indicated probabilities (means) of questioning suspects in relation to drug offences, stronger and weaker grounds, Wave 1 survey

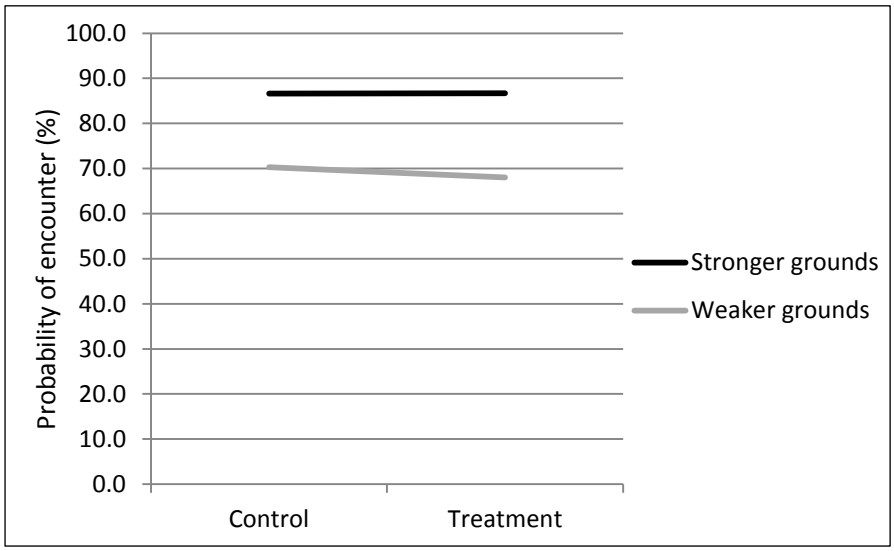
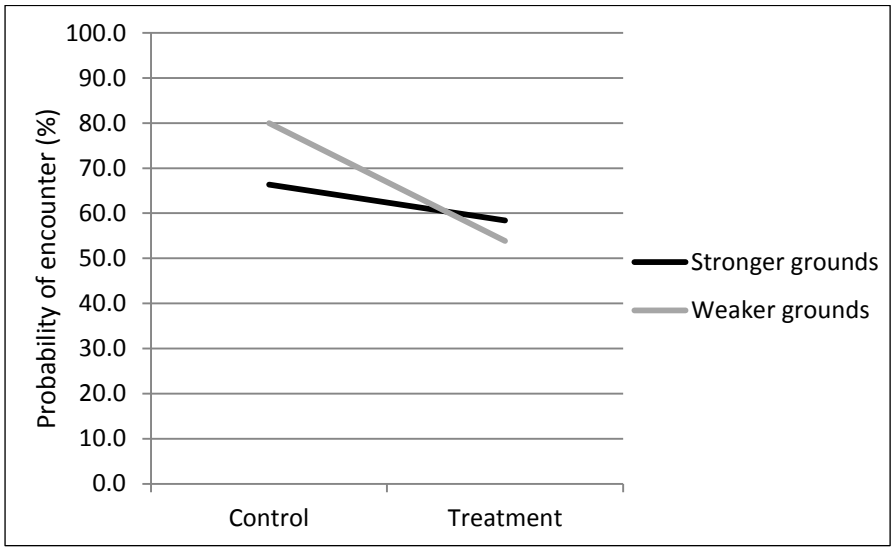


Figure 18. Officers' indicated probabilities (means) of searching suspects in relation to drug offences, stronger and weaker grounds



Force-level analyses of the above measures are presented in table E1 (appendix E). They show few effects of treatment on anticipated questioning, but indicate many statistically significant negative effects on anticipated searches, with the exception of Force E. Significant effects on searching were more frequent for weaker grounds than stronger, in line with expectations.

Wave 2 survey

The Wave 2 survey provided an opportunity to see if the training's effects on anticipated search behaviours were sustained. It, therefore, reproduced two core scenarios from Wave 1, focused respectively on respondents' declared likelihood of questioning and searching suspects in relation to the weaker grounds drugs-based scenario and the stronger grounds robbery-based scenario.

Table 4 presents the results of these questions for the whole six-force sample. Again, they indicate little difference between the treatment and control group in officers' declared probability of questioning of suspects, suggesting training did not affect their declared willingness to intervene when they were suspicious. Once again, there were, however, statistically significant negative effects on officers' anticipated likelihood of searching in relation to both scenarios. This suggested that training continued to produce effects on officers' declared inclination to search, first identified in Wave 1.

Wave 2 incorporated some additional questions not present in Wave 1 that helped explain the strong effects training had on anticipated searching. For each of the stop and search scenarios presented in Wave 2, a follow-up question asked how strong officers thought the grounds were, regardless of their likelihood of searching. Additionally, the survey included a scale measuring officers' **propensity to search when grounds were present**, based on phrases such as 'If I have grounds, I usually conduct a search' and 'I think twice before carrying out a search, even if strong grounds are present.'

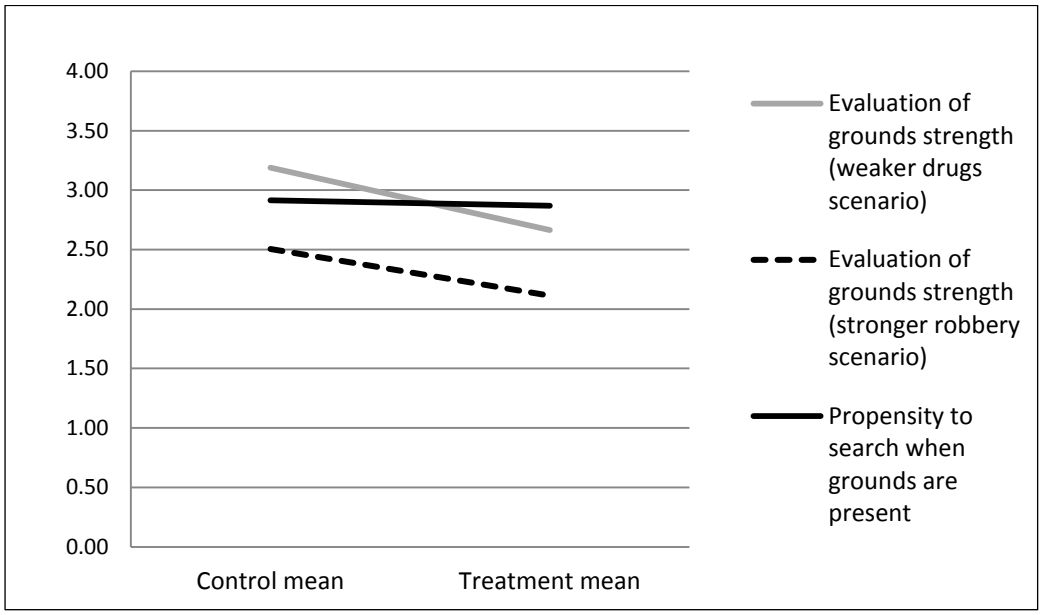
Table 4. Officers' indicated probabilities of initiating questioning or searching by crime type, and grounds strength, for both treatment and control groups (all forces combined), Wave 2 survey

Possible crime type	Strength of grounds	Type of encounter	Mean estimated probability of encounter (%)		Effect size	Sig.
			Control	Treatment		
Robbery	Stronger	Question	93.0	91.9	-0.07	
		Search	60.0	50.0	-0.34	***
Drugs	Weaker	Question	80.6	79.8	-0.04	
		Search	81.7	64.8	-0.58	***

Notes: Effect size is measured using Cohen's d; 0.2 is considered 'small', 0.5 'medium' and 0.8 'large' (Cohen, 1988). Force-level significance levels relied on t-tests. Whole sample significance levels relied on two-way ANOVAs that included binary force dummy variables alongside a treatment factor.
* p<0.05, ** p<0.01, *** p<0.001

Figure 19 presents findings for these measures aggregated across forces. It shows that training had no effects on officers' motivation to conduct searches when grounds were present. Meanwhile, it did have statistically significant negative effects on whether officers assessed that adequate grounds were present in each situation. These results would suggest that the pilot training's negative effects on officers' anticipated likelihood of conducting a search was a function of their more stringent interpretation of grounds for suspicion rather than a reduced motivation to conduct searches.

Figure 19. Officers' assessment of grounds in two survey scenarios, and their propensity to conduct searches when grounds are present (all scales are 0-4), Wave 2 survey



Measure	Control mean	Treatment mean	Effect size	Sig.
Evaluation of grounds strength (weaker drugs scenario)	3.19	2.66	-0.53	***
Evaluation of grounds strength (stronger robbery scenario)	2.51	2.11	-0.41	***
Propensity to search when grounds are present	2.91	2.87	-0.07	

Notes: Effect size is measured using Cohen's d; 0.2 is considered 'small', 0.5 'medium' and 0.8 'large' (Cohen, 1988). Significance for the propensity to search measure relies on a two-way ANOVA that includes binary dummy variables for forces alongside a treatment factor. Significance of grounds evaluation measures relied on Van Eltern tests, a non-parametric Mann-Whitney U test adapted for stratified samples such as this one. The test is stratified by police force.
* p<0.05, ** p<0.01, *** p<0.001

Whether training affects the relevance of ethnic/racial appearance to stop and search decisions

Finally, we assess whether the training produced changes in ethnic/racial patterning of stop and search decision-making. This would be indicative of training having an effect on officer **bias**. To do this, we take advantage of the randomly assigned variation in 'black' and 'white' suspect descriptions present in both treatment and control groups for the Wave 1 survey.

First, it is useful to examine officers' declared probabilities for initiating different categories of encounters according to the ethnic/racial characteristics of the suspect description given in a particular scenario. These results are presented in table 5. Force-specific versions of this table are available in appendix F.

Table 5. Officers' indicated probabilities of initiating an encounter by encounter type, strength of grounds and crime type, across treatment and control groups and white and black suspect descriptions (all forces combined), Wave 1 survey

Ethnicity/ race of suspect	Type of possible crime	Strength of grounds	Type of encounter	Mean estimated probability of encounter (%)		Effect size	Sig.
				Control	Treatment		
White	Robbery	Weaker	Question	76.6	71.1	-0.18	
			Search	44.7	33.4	-0.35	***
		Stronger	Question	92.7	90.4	-0.13	
			Search	68.3	60.9	-0.22	*
	Drugs	Weaker	Question	75.9	74.5	-0.05	
			Search	82.1	53.8	-0.86	***
		Stronger	Question	87.7	87.1	-0.03	
			Search	67.6	61.3	-0.20	*
Black	Robbery	Weaker	Question	62.6	64.5	0.06	
			Search	39.9	28.9	-0.35	***
		Stronger	Question	87.6	86.3	-0.06	
			Search	64.6	54.0	-0.32	***
	Drugs	Weaker	Question	64.8	61.5	-0.10	
			Search	78.0	53.9	-0.70	***
		Stronger	Question	85.6	86.3	0.03	
			Search	65.2	55.6	-0.28	**

Notes: Effect size is measured using Cohen's d; 0.2 is considered 'small', 0.5 'medium' and 0.8 'large' (Cohen, 1988). Force-level significance levels relied on t-tests. Whole sample significance levels relied on two-way ANOVAs that included binary force dummy variables alongside a treatment factor.
* p<0.05, ** p<0.01, *** p<0.001

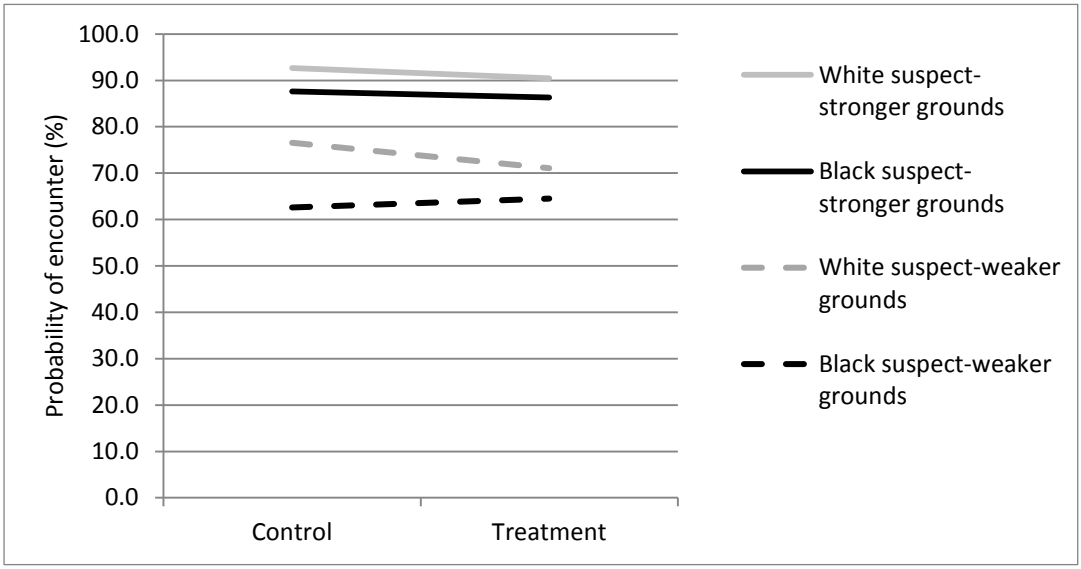
The table shows that effects found for the aggregate sample (in table 4) were also found for the sub-groups of officers presented with either white or black suspect descriptions. As before, officers in the treatment group were significantly less likely than control group officers to say they would search either white or black suspects. Also, in line with the overall programme-level findings, no effect was found for decisions to question for scenarios with either white or black suspects.

Interestingly, the table further suggests that black suspect descriptions were associated with lower anticipated likelihoods of initiating an encounter. Statistical testing using multivariate models (see

tables G2 and G3, appendix G) confirmed this result. Specifically, models showed that a black (rather than white) suspect was typically associated with officers having a lower anticipated likelihood of initiating an encounter. This tendency was statistically significant in most scenario variations (ie, in relation to crime, strength of grounds and whether considering questioning or searching).[7]

Irrespective of these patterns, the key question for the evaluation was whether training changed the relevance of a suspect's ethnic/racial characteristics to officer decision-making. To help assess this, figures 20 to 23 illustrate white/black differences in officers' reported probabilities of initiating encounters for different scenario variations.

Figure 20. Officers' indicated probabilities (means) of questioning suspects in relation to robbery offences, stronger and weaker grounds, white and black suspects, Wave 1 survey



[7] In all scenarios, coefficients suggested encounters were less likely to be initiated for black people compared to white people. Six out of eight scenarios, when modeled, showed statistically significant effects of ethnicity/race on officers' likelihood of searching. Only searching for weaker drugs scenarios and questioning for stronger drugs scenarios were not significant.

Figure 21. Officers' indicated probabilities (means) of searching suspects in relation to robbery offences, stronger and weaker grounds, white and black suspects, Wave 1 survey

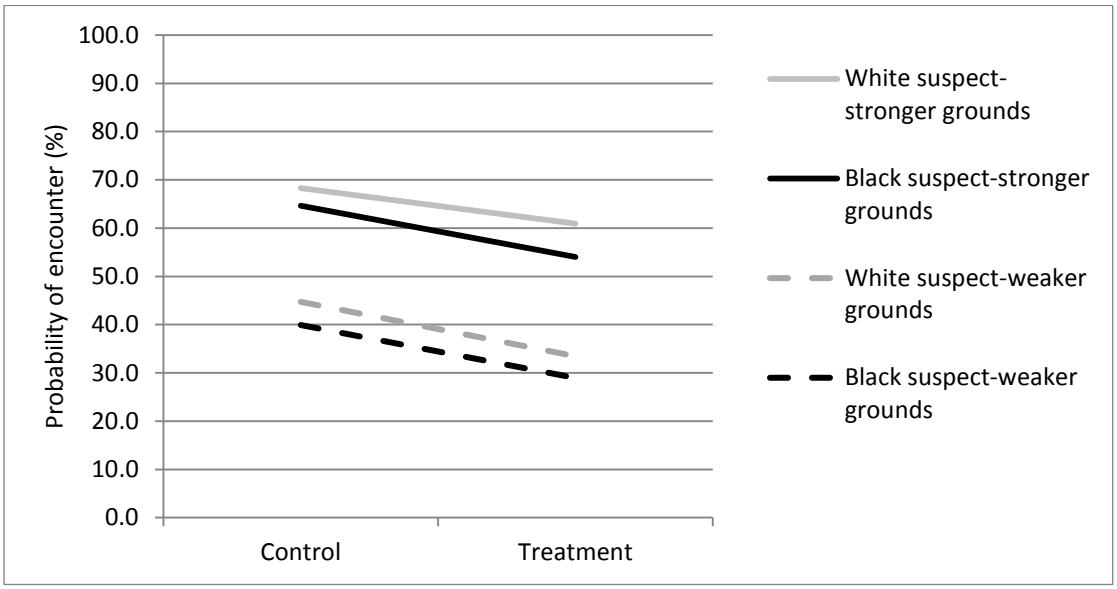


Figure 22. Officers' indicated probabilities (means) of questioning suspects in relation to drugs offences, stronger and weaker grounds, white and black suspects, Wave 1 survey

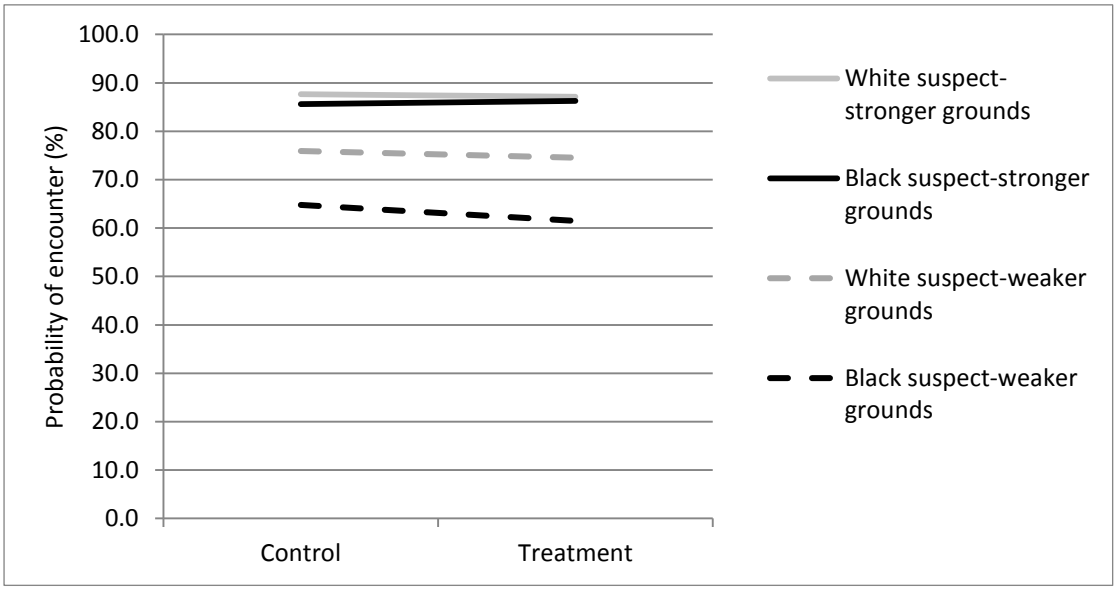
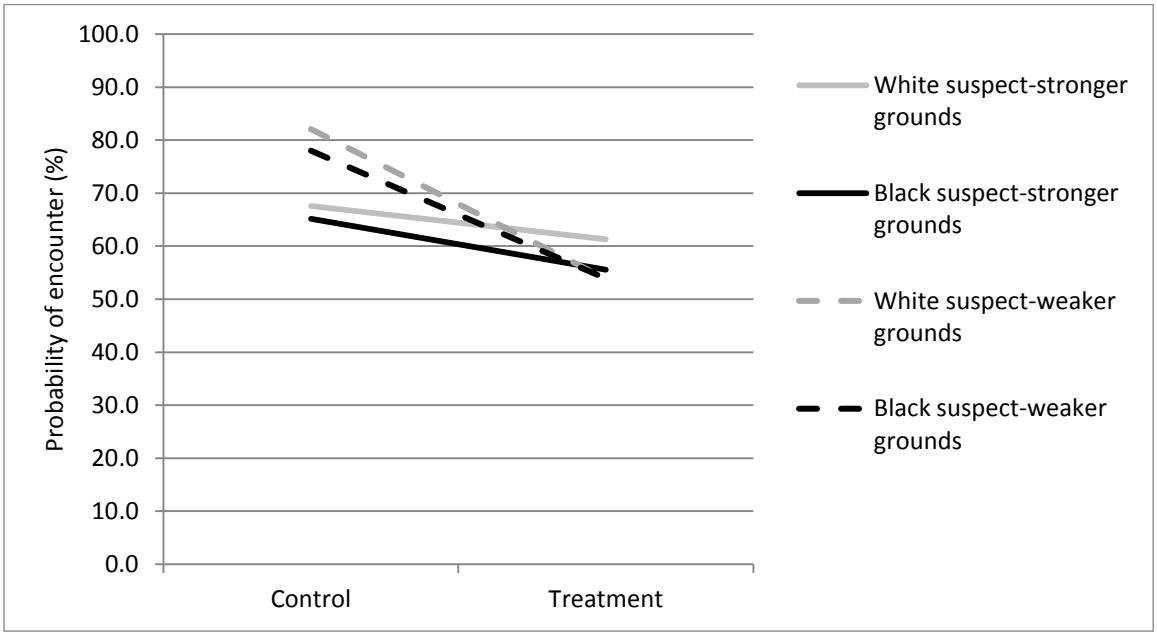


Figure 23. Officers' indicated probabilities (means) of searching suspects in relation to drugs offences, stronger and weaker grounds, white and black suspects, Wave 1 survey



Visually, the patterns of ethnic/racial bias did not show much difference between treatment and control group. Nevertheless, in the weaker grounds scenario in figure 20 (suspected robbery), differences in officers' anticipated likelihood of questioning of black suspects compared to white suspects were less in the treatment group than control group. This suggested training **may** have had an impact on the relevance of race/ethnicity to officer decision-making in relation to these particular decisions. Statistical testing using multivariate models, however, indicated the effect was not statistically significant (tables G4 and G5, appendix G). In short, there is no convincing evidence of an impact of training on the use of ethnic/racial appearance in officers' anticipated stop and search decision-making.

Individual force-level findings are reported in tables F2 to F7 (appendix F). They suggest there were some force-level variations in specific patterns of bias. Separate multivariate modelling across forces and scenario variations, however, found no systematic evidence, at the force-level, that training affected the relevance of ethnicity/race of suspects to anticipated decision-making.

6. Impacts on officers' recorded behaviours

In this chapter, we examine the impact of the training pilot on officers' search behaviours, as measured using police stop and search records. In particular, we were interested in testing hypotheses that the quality of recorded grounds improved as a result of training and that searches by trained officers led to arrests more often. We were also interested in whether the pilot affected the rate at which officers actually conducted searches and the ethnic/racial breakdown of those searches. In short, our findings provide no convincing evidence that training had any impacts on these outcomes, although there was marginal evidence of a small impact on the overall rate of searches conducted by officers. There was also no indication that individual forces differed from the programme-level picture.

Effects on overall search rates

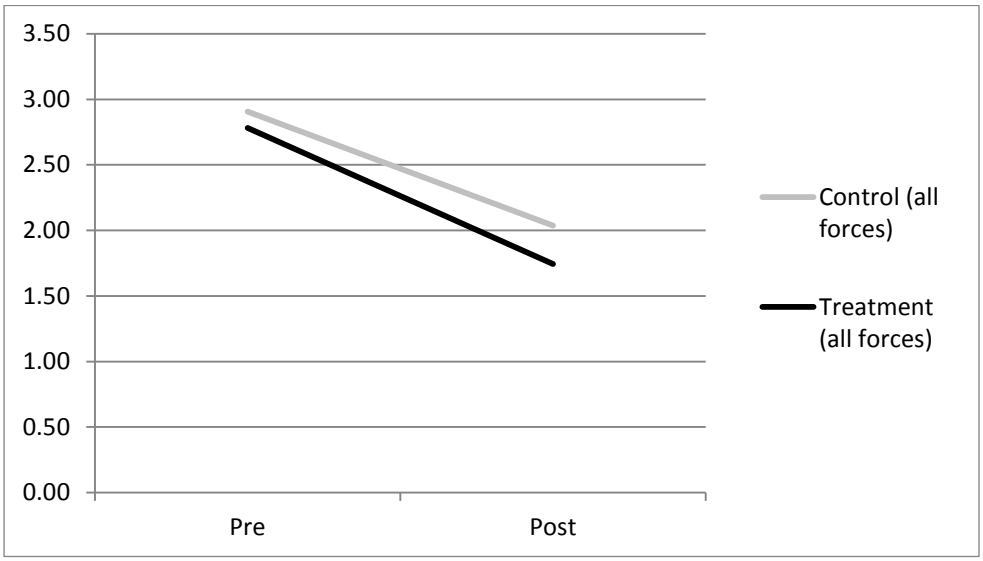
It is useful first to assess the impact of treatment on rates of searches as this sets the stage for other analyses. It also provides a check on survey responses in which treatment group officers indicated they were less likely to conduct searches than control group officers.

Figure 24 presents the mean search rates for officers in treatment and control groups for the three-month periods pre- and post-training. It also includes results of statistical models that assess the size and statistical significance of training effects.

The figure shows a programme-level decline in searches between the pre- and post-pilot periods for both treatment and control groups. This decline was much larger than any differences between the two groups and consistent with the recent national trend (see Quinton, 2016). The decline was slightly sharper in the treatment group, perhaps reflecting the effect of training on officers' reduced tendency to carry out searches, as suggested by their answers to survey questions. Statistical testing of the treatment effect on search rates, however, indicated that the effect fell slightly short of statistical significance.[8] There was only one statistically significant force-level effect, with a statistically significant decline in the search rate in Force C.

[8] A zero-inflated negative binomial model indicated that the average marginal effect of treatment on search rates per officer has a significance level of p=0.098.

Figure 24. Pre- and post-training means of searches per officer over three-month periods for treatment and control groups, based on recorded police searches



	Control		Treatment			
Force	Pre-test mean	Post-test mean	Pre-test mean	Post-test mean	Average marginal effects	Sig.
A	2.34	0.76	2.22	0.65	-0.31	
B	2.72	2.37	3.53	2.21	-0.42	
C	2.81	1.42	2.25	0.69	-0.44	*
D	3.97	2.66	3.89	2.85	-0.01	
E	1.45	0.98	0.96	0.72	-0.02	
F	4.18	4.06	3.82	3.32	0.07	
All	2.91	2.04	2.78	1.74	-0.25	

Note: Significance levels and average marginal effects are calculated using zero-inflated negative binomial regression models with counts of searches in the three-month post-test period as the dependent variable, and treatment as an independent variable. The models also included multiple measures of searches in the three-month pre-test period and (in the aggregate model) force dummies as control variables (see model G7 in appendix G for model specification). Average marginal effects represent the average model-predicted change in searches per officer produced by treatment across all cases. Thus, it is estimated that treatment would produce 0.25 fewer searches per officer on average across the sample (though this is not significant, at p=0.089).
* p<0.05, ** p<0.01, *** p<0.001

Effects on the quality of recorded grounds

A key research hypothesis was that training would improve the quality of officers' written grounds. Table 6 allows us to assess this hypothesis by comparing the legitimate factors included in written grounds for all six forces combined. Variables are based on researcher blind-coding of the most recent search post-training for each officer, where available. The table also shows the proportion of officers recording a search, which provided the basis for the analysis.

The table indicates there were no programme-level statistically significant differences between the treatment and control groups in the prevalence of legitimate factors used in grounds or in the cumulative number of these factors within grounds. The table, however, also indicates that this analysis was based only on search records for slightly more than half of officers. The analysis, therefore, missed the effects of training on officers who had yet to conduct a search in the post-pilot period. Conceivably, they may have reacted differently to training.

Table 6. Legitimate suspicion-generating factors in written grounds in most recent search by officers who conducted a search in the post-training period, treatment and control groups

	Control	Treatment	Sig.
Officers recording searches in post-training period	54.6% (n=361)	51.7% (n=342)	
Legitimate suspicion-generating factor in written grounds:			
Suspicious behaviours observed by police	34.1%	38.0%	
Appearance of carrying something suspicious	8.0%	10.8%	
Match with suspect description/intelligence on suspect	26.0%	24.9%	
Directly identified by victim/witness	15.2%	17.8%	
Suspicious vehicle	12.2%	14.3%	
Drug intoxication	9.4%	7.9%	
Drugs/drug paraphernalia witnessed by officers	31.3%	31.6%	
Location/temporal factors link suspect to crime risk	20.8%	23.4%	
Seen with suspicious associates	24.1%	25.1%	
Self-incrimination during conversation/questioning	26.9%	30.4%	
Group/gang membership	0.3%	0.3%	
Other factor	2.2%	0.6%	
Count of total number of factors (mean)	2.11	2.25	

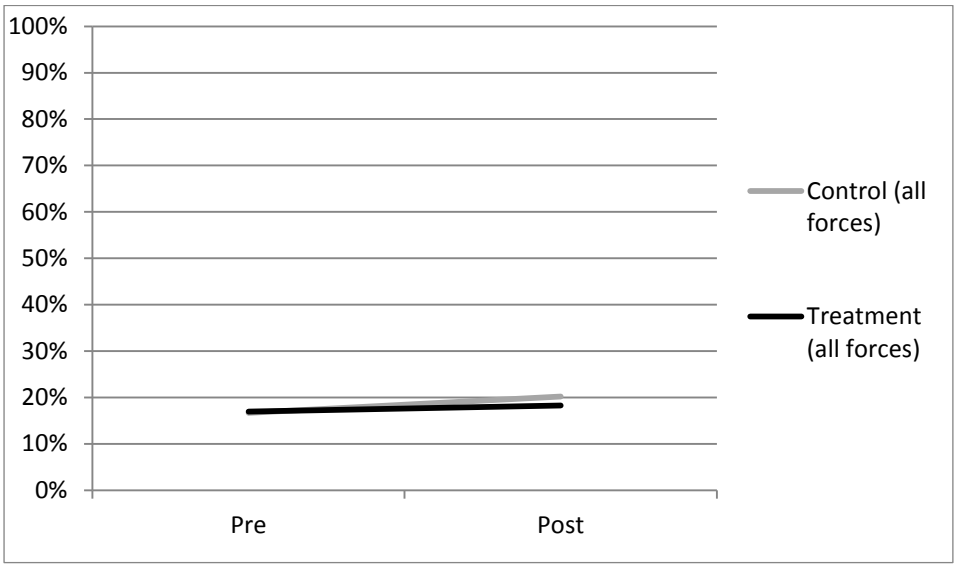
Notes: Significance tests based on logistic regression with force dummies, and (for the count of factors) on a Van Eltern test, stratified by force. Nothing here is significant.
* p<0.05, ** p<0.01, *** p<0.001

Tables E13 to E19 in appendix E provide force-level tables equivalent to table 6. None of these show any evidence of systematic force-level differences between treatment and control groups.

Effects on arrest rates

Another hypothesis is that training would improve arrest rates from searches. To assess this, figure 25 presents the percentage of searches that led to arrests among the searches generated by treatment and control group officers in the pre- and post-pilot period. It also includes results of statistical models that tested the effects of training on arrest rates.

Figure 25. Proportions of searches leading to arrest, among searches conducted by treatment and control group officers in the three-month pre- and post-training periods



	Control		Treatment			
Force	Pre % (n=1,921)	Post % (n=1,346)	Pre % (n=1,842)	Post % (n=1,154)	OR	Sig.
A	8.6	17.9	9.0	18.1	0.96	
B	14.4	15.1	11.7	15.8	1.35	
C	11.5	22.8	20.4	29.3	0.72	
D	17.4	18.1	18.0	19.2	1.03	
E	23.3	17.6	23.6	25.3	1.56	
F	23.3	24.9	21.7	15.5	0.61	*
All	16.7	20.2	16.9	18.3	0.85	

Note: Significance levels and odds ratios (ORs) reflect the interaction effects of treatment and the post-period. They are calculated using logistic regression models run on arrests combined across the three-month period pre- and three-month period post-training. Independent variables included: pre-/post-test time period, treatment group and treatment/post-test period interaction. Aggregate models also included a force dummy. All models calculate cluster-adjusted robust standard errors to allow for officer-level clustering.
* p<0.05, ** p<0.01, *** p<0.001

The figure shows little difference between aggregate treatment and control groups in how arrest rates changed over time. Moreover, there is no statistically significant programme-level effect of treatment on arrest rates. Among individual forces, there was a significant effect in only one force (Force F), suggesting a reduction in the proportion of searches leading to arrest as a result of training. Overall, there was no evidence of any sizeable and consistent training effect at the programme level and few effects at the force level.

Effects on the ethnic/racial distribution of searches

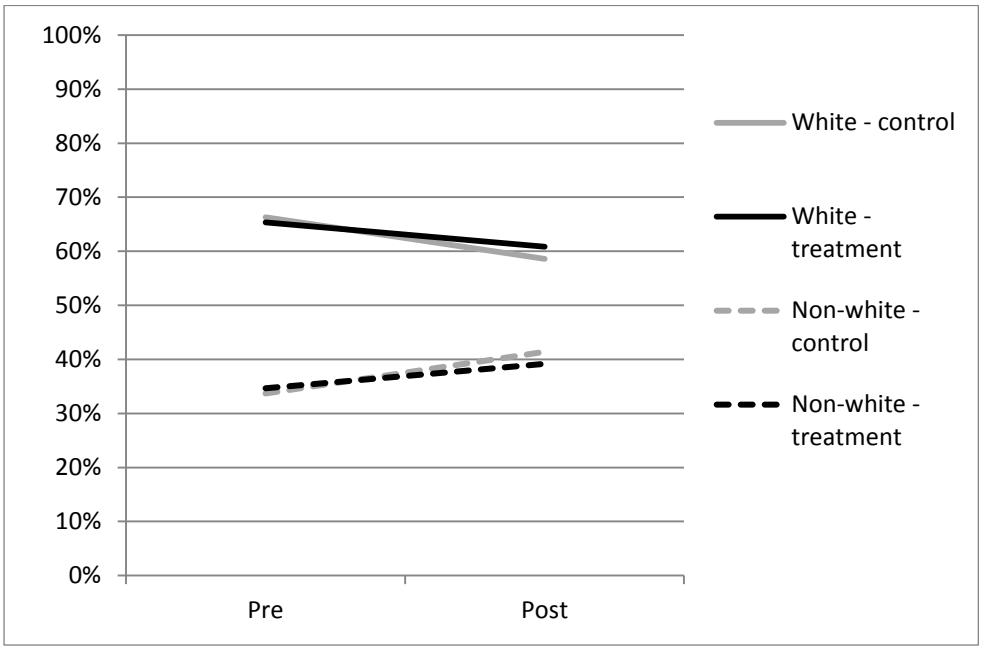
If training has had impacts on race disproportionality, it would necessarily be reflected in changes to the proportion of searches for different ethnic/racial groups conducted by trained officers.[9] Here, therefore, we examine the effect of training on the distribution of searches among people from different ethnic/racial backgrounds. It is, however, important to bear in mind that there was no expectation at the outset that training would change this distribution in any particular direction.

Figure 26 presents the ethnic/racial breakdown for the searches produced by treatment and control group officers in all six forces in the pre- and post-training periods. Once again, there was no evidence of a training impact. Both control and treatment groups showed increases in the proportion of searches conducted on people from black and minority ethnic groups between pre- and post-training periods, but these changes were similar for the two groups. Moreover, statistical testing of effects for different ethnic/racial groups showed no statistically significant effects.

The force-level ethnic/racial distribution of searches is presented in table E12 in appendix E, although small cell counts for some forces and ethnic/racial groups meant we did not produce force-level models. There is no clear evidence from the table, however, of any strong force-level effects of training on the ethnic/racial distribution of searches.

[9] It would have been challenging to directly calculate rates of disproportionality for this research because this would have also required population measures of ethnicity/race of residents for the specific geographies to which participant officers were assigned. It was, however, unnecessary to do so because any effects of training on disproportionality would also be registered as changes in the raw breakdown of searches by ethnic/racial group. It is the latter which we focus on here.

Figure 26. Ethnic/racial patterns of searches conducted by officers in treatment and control groups in three-month pre- and post-training periods for officer recorded searches (all forces)



	Control		Treatment			
	Pre % (n=1,921)	**Post % (n=1,346)**	**Pre % (n=1,842)**	**Post % (n=1,154)**	**OR (ref: white)**	**Sig.**
White	66.3	58.6	65.4	60.8	reference	n/a
Black	12.6	17.5	10.2	12.7	0.78	
Asian	10.1	12.6	13.0	15.3	0.81	
Other	1.2	1.9	1.1	1.3	0.66	
Mixed	3.4	3.1	3.2	2.4	0.71	
Unknown	6.5	6.4	7.2	7.5	0.91	

Note: Significance levels and odds ratios (ORs) reflect the interaction effect of treatment and the post-test period. They were calculated using multinomial logistic regression run on ethnic/racial group combined across the three-month period pre- and three-month period post-training. Independent variables included: pre-/post-test time period, treatment group, treatment-post period interaction and a force dummy. The model calculated cluster-adjusted robust standard errors to allow for officer-level clustering. The table shows no evidence of statistically significant effects.

* p<0.05, ** p<0.01, *** p<0.001

7. Conclusions

This research has examined the impact of a pilot stop and search training programme in six forces on officers' knowledge, attitudes and behaviours in relation to stop and search. It has used an RCT – the 'gold standard' of programme evaluation – to assess the impact of the training. Overall, the research indicated that training had some of its intended effects. Nevertheless, these effects were not found for all outcomes and were often modest when they were found. There was also evidence of some differences between forces in their response to training. Ultimately, however, there were few concrete effects found in recorded street-level practice which raises some questions about the utility of the training as it was formulated for the pilot.

Key findings

Primary hypotheses tested in this study were that training would improve officers' knowledge and attitudes about stop and search and that officers would show improvements in their self-reported, anticipated behaviours. In line with the first of these, training apparently had some modest and enduring positive effects on officers' knowledge of stop and search policy. Trained officers were more likely than control group officers to think weakly written grounds for a search were inadequate, in line with training goals. They were also more likely to think the same for stronger written grounds, although the effect may have been smaller.

There was also evidence that training affected some attitudes towards stop and search. This included reduced support for police stereotyping involving ethnicity/race and high volume stop and search strategies as well as less cynicism toward the regulation of stop and search. No effect was found, however, on officers' support for procedural justice principles in stop and search, which was perhaps not surprising given it was not prominent in the training forces delivered. While not all attitudes were measured over the medium term, some were, indicating that some effects on attitudes were more persistent than others.

The evaluation, however, found its strongest effects on officers' anticipated stop and search decision-making, as measured by surveys. While the research found no impact at the programme level on **how** officers said they would treat suspects in an encounter nor whether they would **question** a suspect, there was an impact on whether officers said they would carry out a **search**. Specifically, when presented with written scenarios about possible drugs or robbery crimes, officers in the treatment group reported being less likely to search suspects. We had specifically hypothesised this would be the case for situations where grounds were weaker and evidence supported this prediction. The effect, however, was also evident for scenarios with stronger grounds; an effect that persisted over time. Moreover, supplementary survey questions suggested that this effect reflected officers being more stringent when assessing the strength of grounds rather than them being less motivated to conduct searches when grounds were present.

Meanwhile, the survey showed that training had no programme-level effects on officers' decision-making in relation to the ethnic/racial appearance of a suspect. Specifically, the research showed that officers generally reported a greater likelihood of questioning and searching white, rather than

black, suspects and this pattern did not change as a result of training. While surprising perhaps, the lack of bias against black and minority ethnic groups is somewhat consistent with empirical work concluding that searches of people from black and minority ethnic groups are not disproportionate, once the characteristics of street populations are taken into account (MVA and Miller, 2000; Waddington, 2004). We should, however, also acknowledge that there may be biases in the way officers responded to sensitive survey questions that reference ethnic/racial characteristics which could also contribute to the finding.

In a set of secondary hypotheses, it was anticipated that there would be improvements in the quality of grounds recorded by officers conducting searches and increased arrest rates from recorded searches. There was no evidence of a change in the quality of grounds recorded, suggesting, in turn, that there were no changes to the kinds of searches being conducted or officers' record-keeping. Similarly, programme-level arrest rates showed no impacts from training. Additionally, analysis of search rates found no clear effects from training. It is possible that there was a small negative effect, although this did not achieve statistical significance. An effect, however, would be in line with the survey responses given by officers. Training had no apparent effect on the ethnic/racial distribution of search suspects in line with responses to the survey findings reviewed above.

While the research has focused on overall programme-level training effects across the six forces, it has also reviewed the outcomes of individual forces. This analysis suggests that training was associated with more pronounced effects in some forces than others, although variations were not consistent across types of outcome. Thus, Force E registered almost no statistically significant effects on the full range of outcome variables. Meanwhile, Force D stood out for experiencing the most significant effects of treatment on knowledge and attitudes. When looking at effects on anticipated search behaviours, statistically significant effects were found in all forces except Force E. Despite these differences, a finding at the force level, in line with programme-level findings, was a lack of clear and consistent effects of training on recorded officer behaviours.[10] Force-level differences may reflect some variation in the implementation of training between sites. Forces, however, also varied in their geography and organisation which may have influenced how training was received. There was also the possibility that some differences in results between forces are the product of chance variations.

Implications

It should be possible to improve stop and search training to take account of some of the shortcomings highlighted by this research, although the development of training should also take into account the detailed findings of the companion process evaluation (Giacomantonio et al, 2016).

[10] Two isolated significant effects of training on recorded behaviours were identified at force level: Force C registered a significant decline in search numbers and Force F a statistically significant decline in arrest rates. Other measures of recorded behaviours within these forces, however, did not show significant differences. Moreover, a few statistically significant results may have been simply as a result of chance given the multiple comparisons made at force level, meaning we should be sceptical about reading too much into these results.

The current research highlights, in particular, the gap between knowledge, attitude and anticipated behaviours on the one hand and actual street-level officer behaviours on the other. While training apparently affected the former, it did not influence the latter. This suggests that future stop and search training should perhaps give greater emphasis to modelling behaviours in stop and search encounters, alongside more abstract teaching of the use and regulation of stop and search powers. This could involve the greater use of role-plays, for example, in which officers enact decisions about when to stop and search and how to engage with suspects.

Future stop and search training should probably also place greater emphasis on how officers interact with suspects, in particular their application of procedural justice principles. This is an area for which training produced no impact, either in relation to officers' articulated support for procedural justice principles or in their anticipated behaviours when confronted with a challenging stop and search encounter.

A final consideration is that, by focusing only on frontline officers, the pilot training may have missed an opportunity to influence supervisors and managers in their strategies for overseeing frontline officers' stop and search activity. A training package that also targeted force supervisors and managers might have been be more effective. Such an approach could involve education in the auditing and supervision of officers' use of stop and search and developing supervisors' and managers' skills for encouraging and directing officers to adopt more effective and fairer stop and search practices.

References

Bloom, H. S. (2006). The core analytics of randomized experiments for social research. MDRC working paper. New York: MDRC.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Equality and Human Rights Commission (EHRC) (2010). Stop and think: A critical review of the use of stop and search powers in England and Wales. London: EHRC.

Equality and Human Rights Commission (EHRC) (2013). Stop and think again: Towards race equality in police PACE stop and search. London: EHRC.

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

Her Majesty's Inspectorate of Constabulary (HMIC) (2013). Stop and search powers: Are the police using them effectively and fairly? London: HMIC.

Home Secretary (2014). Stop and search: Comprehensive package of reform for police stop and search powers. Oral statement to Parliament, 30 April 2014. London: Home Office.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159-174.

Lewis, P. Newburn, T., Taylor, M., Mcgillivray, C., Greenhill, A., Frayman, H. and Proctor, R. (2011) Reading the Riots: Investigating England's summer of disorder. London: The London School of Economics and Political Science and the Guardian.

Macpherson, W. (1999). The Stephen Lawrence Inquiry. London: HM Stationery Office.

Miller, J. and Alexandrou, B. (2016). College of Policing stop and search training experiment: Impact evaluation. Appendix. London: College of Policing.

MVA and J. Miller (2000). Profiling populations available for stops and searches. London: Home Office.

Quinton, P. (2016). College of Policing stop and search training experiment: Design of the randomised controlled trial. London: College of Policing.

Quinton, P. and Packham, D. (2016). College of Policing stop and search training experiment: An overview. London: College of Policing.

Rosenbaum, D. P. and Lawrence, D. S. (2012). Teaching respectful police citizen encounters and good decision-making: Results of a randomized control trial with police recruits. Paper, National Police Research Platform.

Scarman, L. (1981). The Scarman Report: The Brixton disorders. London: HM Stationery Office.

Tyler, T. R. (2004). Enhancing police legitimacy. The Annals of the American Academy of Political and Social Science, 593(1), 84–99.

Waddington, P., Stenson, K. and Don, D. (2004). In proportion: Race, and police stop and search. British Journal of Criminology, 44(4): 1–26.

Wheller, L., Quinton P., Fildes A. and Mills A. (2013). The Greater Manchester Police procedural justice training experiment: Technical report. Ryton-on-Dunsmore: College of Policing.