**Pol-1304**

**Jamie Hobday**

**Wolfson College**

**Supervisor: Dr Katrin Muller-Johnson**

# TARGETING REASONS FOR REJECTING RANDOM ASSIGNMENT IN AN RCT

**Submitted in part fulfilment of the**

**requirements for the**

**Master's Degree in Applied Criminology and Police Management**

**[2014]**

*Abstract*

*The Randomised Control Trial (RCT) is seen as the 'gold standard' of evaluation work, but running such experiments with high fidelity and validity can be a challenge, especially in an operational policing environment. Researchers are continually seeking tools and methods of working that can support experimentation in front-line policing operations. An algorithmic triage approach to case selection and random allocation, utilising the Cambridge Gateway on-line tool, holds promise in this arena. This approach and tool has been operationalised in the Turning-Point Experiment, UK. Five custody suites in a busy urban police agency utilised this approach over a three year period. Case selection and random allocation was conducted by operational police officers on a 24/7 basis. Officers were instructed to use their discretion to exclude cases that were ostensibly eligible, but they felt would not have the public's support. The consistency and validity of their decisions is studied through these decisions to reject random allocation. The improving rate of rejection over the Turning-Point experiments life is investigated, and the method of project implementation is proposed as a likely explanation for improvements. Officer's rejections are categorised as either appropriate or inappropriate and inter-rate reliability is measured across four key project staff. The pattern of inappropriate rejections amongst individual officers is studied and compared to officers' characteristics identified from a survey. A moderate correlation is found between officers making better quality decisions and higher levels of educational achievement. The conclusion is reached that this model can be effective at utilising officers in case selection and random allocation roles in large scale RCTs in policing. Recommendations are made to any researchers wanting to replicate the experiment.*

## Acknowledgements

I would like to thank my supervisor Dr Katrin Muller-Johnson for her guidance and encouragement in the writing of this thesis.

I extend my appreciation and gratitude to the Cambridge University staff involved in the Turning-Point Experiment, especially Peter Neyroud, Molly Slothower, Larry Sherman, Heather Strang and Barak Ariel. The learning journey they have taken me on has been extraordinary.

I extend this appreciation to all of the police and partner colleagues who were also so committed to the experiment, especially Darren Henstock, Simon Williams, Allan Green and Tom Joyce: without whom we would not have experienced anything like the success we have. The custody officers and offenders managers –too many to mention – also deserve recognition; as do colleagues from the Youth Offending Service, Crown Prosecution Service and Probation, to name but a few.

I offer a huge thank you to my friends and colleagues who have helped me through this experience and have tolerated my distractions and absences.

However my gratitude is most strong for my family, particularly my wife Katie, who has had to endure 18 months of absences and neglect. Their love, support and patience has carried me through and shown me how lucky I am to have them by my side.

# Contents

# List of Tables

## List of Figures

## List of Abbreviations and Definitions

| | |
|---|---|
| RCT | Randomised Control Trial |
| CPS | The Crown Prosecution Service |
| AR | Appropriate Rejection (of random assignment) |
| IR | Inappropriate Rejection (of random assignment) |
| DPP | The Director of Public Prosecutions |
| IT | Information Technology |
| Weighted score | a score of between -1 and +2 given to each decision, where -1 is an appropriate decision, +1 is an appropriate decision and +2 is a clearly inappropriate decision |
| Mean IR score | the mean of the four assessors weighted scores for each decision |
| Composite IR score | the sum of the mean IR scores for every rejection decision each officer has made |
| IR Rate | the composite IR score for each officer, divided by the total number of entries they made into the Cambridge Gateway. |
| Overall Score | the combination of each officers IR ad AR Rates |

# Introduction

The basic tenet of this study is to determine if using an algorithmic triage model to support eligibility decisions in the Turning-Point Randomised Control Trial (RCT) led to high levels of consistency and validity in police officer decision-making, especially for more subjective selection criteria. The processes of case selection and randomisation are central to the validity of any RCT, so maintaining their fidelity is a key objective. This has historically led to these roles being preserved for researchers who were seen as independent, unaffected by operational considerations, current practice or cultural influences. This can lead to increased costs and therefore potentially fewer such experiments taking place.

An algorithmic triage model of decision-making is characterised by a series of eligibility questions, leading the operator logically through all the relevant selection criteria and recording their responses to each. Any measure not meeting the necessary criteria will lead to the case being identifed as unsuitable and unable to continue further. Importantly, cases which meet all the selection criteria can still be excluded where the operator identifies special characteristics that they believe make it unsuitable. This is the exercise of discretion. A rationale is required for these decisions. Suitable cases then go on to be randomly allocated to treatment or control groups, and dealt with accordingly.

This is *not* a study of whether cases were treated as assigned after randomisation, which is also an important question, but is being studied separately.

This study is designed as a piece of 'real-world' research (Robson, 2002) and thus is focussed on problem-solving, attempting to identify actionable, practical factors that those running randomised experiments in policing can take away and use. A flexible methodology has been utilised, building on a dataset collected during a live experiment of fixed design (Robson, 2002), supplemented with a survey of the decision-makers and their colleagues. Experimental data on reasons for rejecting ostensibly eligible cases is explored, with officers decisions categorised as either appropriate or inappropriate. The way that inappropriate decisions are made over the experimental life-span is examined, specifically in relation to developments in the decision-making tool and training/support given to the decision-makers. We have then determined if there are patterns of inappropriate decision-making by officer, and if these correlate to any characteristics of those officers (identified from the survey data). Finally, recommendations are made in relation to the use of the algorithmic triage model in future experiments in policing, and how such a model could be efficiently implemented.

## Relevance

Taking experimental approaches from the social sciences into the operational policing environment can be problematic, but also holds hope of increasing the effectiveness of police agencies and supporting them in better achieving their aims (Sherman, 1992). Understanding hurdles to the implementation of randomised control trials is essential in progressing and improving their use in this area of public service delivery (Sherman, 2010). In this study we use Turning-Point as a case study to assess the effectiveness of the algorithmic triage model for case selection.

The overall volume and distribution of inappropriate decision-making in the Turning-Point experiment will help us understand whether the algorithmic triage model was effective or not. This will have relevance to the operation of RCTs in police settings in the future, potentially giving a model to replicate (or avoid). Policing in the United Kingdom, and the wider western world, is increasingly adopting the RCT as a rigorous method of testing interventions – see Figure 1 (Braga et al. 2014). Discovering methods, tools and approaches to delivering such experiments cost-effectively, in volume and to high levels of validity is essential.

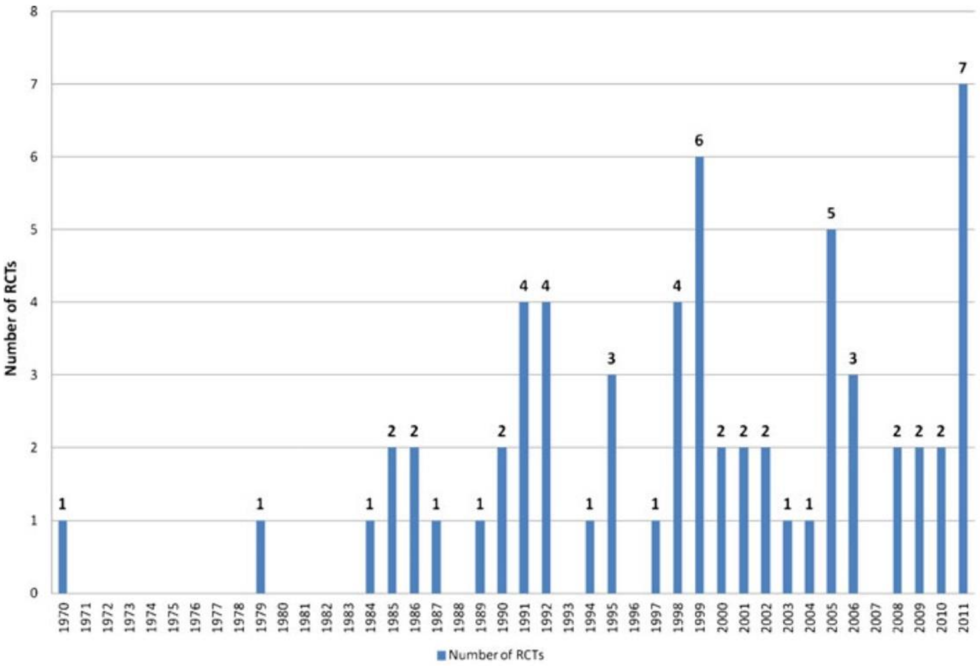


**Figure 1 Timeline of RCTs in Policing, 1970 – 2011 (Braga et al. 2014)**

The operational environment of policing is one where officers seek the best outcomes on a case by case basis. Quite rightly, their focus is on trying to the best in each individual set of circumstances. Officers will use their training, experience and the organisational processes/rules in place to seek the best outcome in each end every case, using their discretion and professional judgment (Kelling, 1999). The desire to do this will naturally raise a conflict with any process that randomly allocates cases to one treatment or another, as we are doing in Turning-Point and as all RCTs do. Whether their views are correct or not, officers will generally have a view as to which of the two treatments is likely to lead to the better outcome for the case in front of them and their instincts will push them towards trying to obtain that treatment. This has the potential to lead both to cases being kept away from randomisation altogether, and to cases failing to be treated as randomly assigned. These are problems besetting all randomised trials in operational policing (Sherman, 1992).

These two problems can and will lead to problems of validity: where cases are deliberately kept away from the experimental field the problems of external validity will occur; where they are not treated as assigned, problems of internal validity will occur (Ruane, 2005). This study examines the thorny problem of officers inappropriately deeming cases ineligible and thus potentially reducing the external validity of police led experiments.

Where external validity is low then it is difficult to generalise that the results or effects found during the experiment will be likely to occur in other similar places that deliver the same treatments. It is essential that experimental results are generalisable, as a large measure of the rationale for running experiments is that the lessons learnt can be taken and applied elsewhere.

This specific issue has relatively little written about from a policing perspective: therefore it is hoped that this study will make a significant contribution to knowledge in this area. It is possible that the Turning-Point Experiment will be replicated elsewhere and therefore these findings should help those responsible for any replications to increase the external validity of their experiment. This study will investigate this issue through the analysis of decision rationales, recorded contemporaneously, and the use of a survey of decision-makers and their colleagues.

## An Overview of the Experiment

The Turning-Point Experiment RCT took place in Birmingham, England: it commenced late 2011 and ceased allocating new cases in June 2014. Birmingham is policed by West Midlands Police, the second largest force in the country, staffed by 8,204 officers and 3,619 support staff (West Midlands Police, 2014). Birmingham is the forces largest city, with a population of 1.1 million, has high levels of ethnic diversity, and a noticeably young population profile (Birmingham City Council, 2014). Turning-Point involved the selection of a small number of low-risk criminal cases that would usually be prosecuted, randomly assigning these cases either to prosecution as usual or offering the offender an enhanced out of court disposal. The experiment was designed to test the relative cost effectiveness of the two disposals. Importantly, cases were selected and randomised by operational staff (custody officers), not researchers.

It was calculated by the principal researcher that 400 cases were required in the experiment to achieve the necessary statistical power to report on recidivism rates (Neyroud 2011). The initial field site included 2 commands; in 2012 this area was expanded by a further 2 commands to cover the whole of Birmingham. The project area contained approximately 1000 operational staff, 5 custody suites and 58 custody officers, processing an average of 524 eligibility decisions a month. The author filled the role of project manager. In comparison to other RCTs in an operational policing environment Turning-Point is therefore one of the largest, utilising a professional (as opposed to academic) distributed and devolved decision-making model. This makes the decision-making of particular interest.

The eligibility criteria for this experiment were designed jointly by the police, the Crown Prosecution Service and academics. An integral element of the case selection procedure was the use of professional judgment or discretion by custody officers. Whilst some basic eligibility criteria were easy to verbalise and to write into explicit instructions, (Appendix A – Prosecution Position' lays out the agreed criteria), others were less so. Cases more serious than would normally be diverted from court were deliberately in scope, so in order to protect the reputation of the force custody officers were expected to assess each case in the light of public expectation and reputational risk, as well as strict eligibility. Custody officers recorded every decision, using an on-line tool known as the Cambridge Gateway (see Ariel et al 2012 for a fuller explanation of the Gateway). Where ostensibly eligible cases were excluded from randomisation officers were asked to describe why in their own words - it is these cases and responses that give rich data to help answer our research question. In many cases the rejection of an eligible case will have been perfectly correct and evidence of custody officers making good decisions in line with the overall objectives of the experiment and protecting the forces reputation. These are described in this study as 'Appropriate Rejections' (AR). On other occasions the reasons for rejection will contradict or be at cross purposes to the experimental objectives: these are described as 'Inappropriate Rejections' (IR). We will pay these decisions pay extra attention due to their ability to undermine the validity of the experiment, and their power to shed light on the officers' decision-making process.

## Literature Review

In trying to determine the benefit of using an algorithmic triage model in an RCT, a number of different areas of literature are relevant. Work on the effective operation of experiments in policing is central. More general literature on program implementation is of great use as the way in which the model was introduced and staff supported in its use has been crucial. Due to the nature of this particular experiment we shall also briefly consider the literature on the use of discretion in police decision-making, and previous work on the particular culture within the police custody environment.

It will assist the reader if we start with a brief explanation of the evidence base for the approach adopted within the Turning-Point Experiment in the first instance.

### Turning-Points' Evidence Base

There is a considerable body of work on life-course criminology and desistance which suggests that 'turning-points' occur in the lives of offenders, leading to step changes in behaviour (Laub and Sampson, 2003). There is also evidence that prosecuting offenders can itself lead to increased rates of reoffending, certainly with juveniles (Petrosino, Turpin-Petrosino and Guckenburg, 2010). Project HOPE was a trial of a swifter method of dealing with offenders with substance misuse problems in the US (Hawken, 2009), suggesting this may increase desistance better than severe responses. Also supporting the hypothesis that severity may not be an effective lever at encouraging desistance are experiments in restorative justice (Sherman & Strang, 2007), and dealing with some domestic violence offenders without immediate arrest (Dunford, Huizinga and Elliot, 1990). These pieces of evidence led to the development of the approach on trial within Turning-Point (Sherman & Neyroud, 2012).

## Experimentation in Policing

RCTs in policing and justice settings are essential to ensure programs do no harm, and to measure the scale of their effectiveness (Sherman 2007), but the approaches to testing differ between academics and practioners. A small network of academics have been responsible for the lion’s share of police experiments, and their experience points to the necessity to select, coach and support the personnel involved (Braga et al. 2014). There is a growing body of research on the importance of random allocation of cases, but less on cases selection. Previous advice has been for the need for researchers to take control of case selection and randomisation as the only way to preserve the integrity of the system. Recent developments of an on-line tool using the algorithmic triage model now raise the possibility of operational staff conducting these roles with confidence and high levels of reliability (Ariel et al. 2012).

There is now a growing body of literature on the running of randomised experiments in social science settings. Oakley (2000) describes a 'paradigm war' between those who favour qualitative or quantitative approaches to social science. The use of random assignment in any operational criminal justice setting is ambitious and there is little context specific advice to help experimenters (Roman et al, 2012), though the difficulties are well documented (Weisburd 2000; Cordray 2000; Greene 2014; Abramowicz et al. 2011; Lunn et al. 2012; Sherman et al. 1992). Sherman (2007) argues that randomised control trials are not just essential to ensure that criminal justice interventions do no harm, but also that they are essential to measure the scale of the positive impact they have too. Additionally they can also help understand the severity of the problem being tackled and the best implementation strategies for programmes (Boruch et al. 2000). If conducted with high fidelity then randomised trials can achieve high levels of internal and external validity, making them the 'gold standard' of evaluation (Sherman, 1998).

Strang (2012) notes how the standpoint of researchers and practioners can differ in a fundamental way that is crucial to randomised trials. She identifies researchers as comfortable with doubt and uncertainty, but their operational counterparts as having more certainty in their practice. She advocates strong relationships between both groups and relentless briefing as key elements of a successful partnership. The reality of police experimentation is that the majority of it has been led by a small network of academics, whose influence has been widespread (Braga et al. 2014). There is a need to widen this network and this report talks to processes and tools that can support that expansion.

The limited advice there is for police staff and researchers agrees on some key themes. Getting the right core staffing is essential, as is getting the practioners on board: as the 'agents of selection and randomisation' (Strang & Sherman 2012; Roman et al. 2012). The Milwaukee Domestic Violence experiment lays much of its success down to both the leadership of key, influential individuals; and selecting the officers that would be conducting the experiment itself. No staff selection took place for the Turning-Point experiment - luckily there is also value placed on effective training, coaching and support for the team (Strang & Sherman 2012).

Kilburn (2012) describes the differing approaches to randomisation between researches and practioners well: as a case of the glass appearing half full or half empty. She describes the surprise of her research team to find practioners avoiding the study because they viewed the chance of participants not getting the treatment as undesirable, despite the fact this is exactly what they would get by being excluded. Although treating cases as assigned is generally considered an essential element of any RCT, it has been known for experimental design to allow for the deliberate selection of cases by the practioner. A health and justice experiment in Switzerland allowed doctors to 'purposely select' up to 25% of eligible participants to receive one particular treatment. This drastic action was felt necessary to reduce the resistance to random assignment, and to reduce the temptation of practioners to manipulate the process (Killias et al. 2000).

The literature identifies key issues for those planning randomised trials. Care not to overestimate likely case flow is advised, as is careful consideration of the agents and point of random assignment. Whereas there is a body of work on the problems associated with random assignment, there is very little written on the issue of case selection. Interesting and valuable information again comes from the Milwaukee Domestic Violence Experiment where the broad eligibility of misdemeanour domestic battery offences is further restricted in a similar way as occurred in Turning-Point. We find cases excluded where police officers are assaulted, restraining orders violated or offenders refuse to leave when requested (totalling 18% of eligible cases). In Milwaukee the officers were issued with laminated cards, listing the selection criteria, as a tool to aid accurate case selection. Increasingly technology can assist nowadays.

For experiments the point of random assignment is a, if not the, crucial question. Historically the preferred option is having research staff complete this to ensure independence (Strang & Sherman 2012), but this creates significant resourcing requirements in a 24/7 operational policing environment. This is understandable when one looks at the history of RCTs in operational policing, for example the Minneapolis DV Experiment, where researchers discovered that officers were able to discover the pre-planned sequence of randomisation and were thus able to manipulate circumstances to favour arrest in cases they felt deserved it more. This led to the arrest cohort having more prior arrests than the comparison group – an obvious threat to validity (Sherman 2009).

A relatively new and potentially helpful tool in these circumstances is the Cambridge Gateway (also known as the 'Randomiser'). By creating a web-based tool for operational staff which simultaneously collects data, identifies eligibility and randomly allocates to treatment or control we may have arrived in an era where these roles can reliably and confidently be undertaken by practioners (Ariel et al. 2012). This study investigates valuable information collected via this tool, using it to assess how it supported the algorithmic triage model of decision-making by operational staff in the West Midlands Police, and how reliable the results of working in this way are.

## Implementation

Any study of experimentation in operational policing must consider the issue of implementation. We will consider a synthesis of the literature on this issue which identifies three levels of components and four levels of support required for success. Viewing the project as an evolutionary process may help understand the development of Turning-Point over time. Again, we can draw parallels with medicine, relying on an overview of successful implementation strategies which highlights three key issues. By applying organisational justice theory, and treating the project as a change programme, we are able to benefit from a new perspective.

We are greatly assisted in our search for 'what works' in implementation of programmes (including experiments in this wide definition) by the work of Fixsen et al. Their synthesis of the literature to date produces an excellent overview of effective practice in this area (Fixsen et al. 2005).

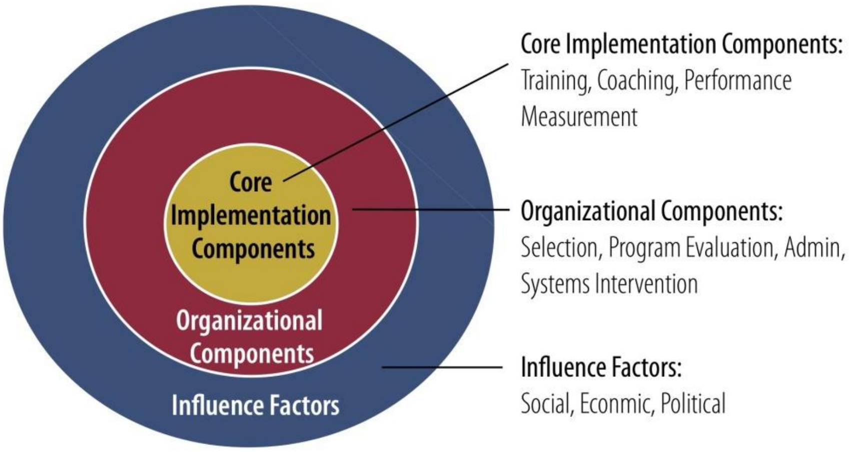

Core Implementation Components
Organizational Components
Influence Factors
Core Implementation Components:
Training, Coaching, Performance Measurement
Organizational Components:
Selection, Program Evaluation, Admin, Systems Intervention
Influence Factors:
Social, Econmic, Political

Figure 2 Multilevel Influences on Successful Implementation (Fixsen et al, 2005)

Fixsen finds that a range of components are necessary for successful programme implementation, and he describes these occurring at three levels: core, organisational and influence factors (Figure 2). He makes the point that no programme operates in a vacuum and there are factors beyond the core components that have a powerful bearing on implementation. Paying particular attention to developing evidence-based interventions within organisations, Fixsen creates a framework for implementation. This consists of a 'source', the set of core intervention components; the practioner who is changing their service to deliver the source in some way; and the communication link between the two. The communication link includes training, admin support etc. Importantly their findings are that training alone is ineffective and that ongoing coaching and support, especially in the field, is essential if real change is to be sustained. This is supported by a meta-analysis of the effects of training and coaching on teachers' actual implementation of programmes in the classroom. They identify 4 levels of support on a scale of increasing impact: theory and discussion; demonstration in training; practice and feedback in training; and finally coaching in the classroom. They found that programmes just using the lower level led to just 5% or less of participants being able to demonstrate the skill or using it in the classroom. However, when programmes utilised all four methods, both skill demonstration and actual use in the classroom rose to 95% of participants.

Fixsen developed a framework for implementation that links these components together, illustrating the key elements of feedback and fidelity measures (see Figure 3). They found that successful implementation utilised measures of integrity and fidelity to the new practices and created feedback loops with practioners so as they were able to adjust and improve their performance in an operational setting.

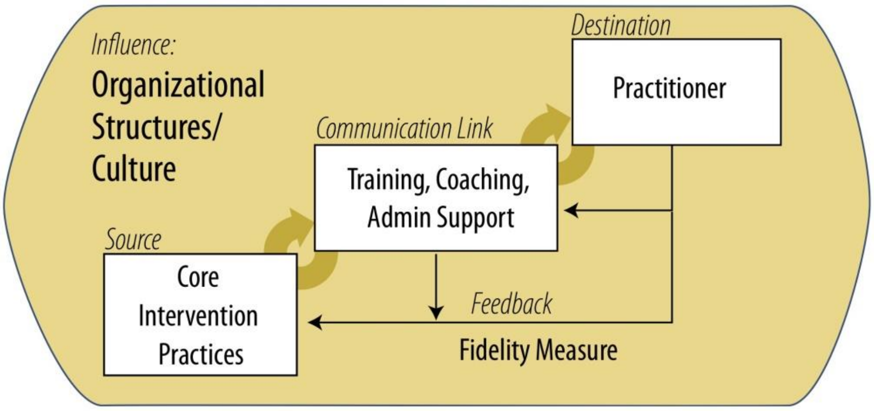

Influence:
Organizational Structures/ Culture
Destination
Practitioner
Communication Link
Training, Coaching, Admin Support
Source
Core Intervention Practices
Feedback
Fidelity Measure

Figure 3 Implementation Framework (Fixsen et al, 2005)

Congruent with Fixsens' 'multi-level' influences is earlier work by Pressman and Wildavsky (1973). In reviewing the implementation of a large governmental project in the US they identified the difficulties of bringing together workgroups to work on a shared initiative. They produced an interesting list of reasons why participants could agree to the objectives of an initiative, but still oppose or fail to implement it. This list included three worthy of mention: simultaneous commitments to other initiatives; differences of opinion on intermediate organisational goals; and dependence on others who lack the necessary sense of urgency. Turning-Point was implemented at a time of difficult fiscal constraints that were generating a multitude of initiatives and organisational changes running simultaneously, thus the environment was a significant factor. Pressman and Wildavsky describe implementation as an evolutionary process, with plans having to adapt to multiple changes and challenges. We may find this perspective helpful in the retrospective analysis of the Turning-Point experiment.

The difficulties of turning evidence of effective interventions and behaviour into standard organisational practice in medicine could be of relevance to the implementation of Turning-Point. Three key issues are identified in an overview of the medical literature on this matter (Grol & Grimshaw 2003). Firstly, the attributes of the evidence itself: including, amongst others, the compatibility of the recommendation with existing practioner values, the complexity of decision-making required, and the degree of organisational change required. Secondly, the specific barriers and facilitators to changing practice, where understanding these in a context-specific way is essential to success. Thirdly, developing specific strategies at multiple levels to deal with these challenges – all strategies have potential to be effective in the right circumstances: a finding similar to Fixsens'.

Organisational justice theory may assist in our understanding of implementation issues in the Turning-Point Experiment. A recent test of these theories in an operational police setting found support for them (Bradford et al. 2013). The authors conclusions are that in this respect the police service is not different to other organisations in that treating staff fairly and giving them a voice in change programmes developed their commitment to the stated objectives of both the programme and the organisation more widely: "in essence, people are motivated to support organisations in which they feel valued members" (Bradford et al. 2013, p113). We will consider later how the implementation of the experiment and the development of the Gateway tool within it relate to these findings.

## Discretion and Decision-Making

The use of discretion in the delivery of public services, such as policing, involves a continual balancing of policy requirements and the needs of the individual. The way in which the police exercise their discretion in the disposal of offenders is a source of regular public criticism, but there is a dearth of relevant evidence to base public policy on. Studying officer decision-making through the lens of it being either slow and deliberate or quick and intuitive may help us understand how officers come to their decisions.

The issue of police use of discretion is core to the study of the algorithmic triage model, which retained a significant discretionary element. A seminal piece of work by Michael Lipsky in 1980 ('Street Level Bureaucrats') is highly relevant. Like most of the public sector, police officers deal largely with 'non-voluntary' clients and operate in an environment where there is a need for continual balancing of broad policies and individual service. This can lead to phenomena such as the rationing of limited resources; 'creaming' cases most likely to succeed in organisational performance measures to certain interventions; and worker bias. Lipsky states worker bias occurs most when decisions are explicitly moral, considerations of 'worthiness' are present, and clients invoke feelings of hostility or sympathy. The reader may see this model as pertinent to the custody officers' prosecution/diversion decision-making in this study. Because of the need for discretion to ensure the best services are delivered to each individual, developing service providers' skills in this area is essential (Lipsky, 1980).

That different offenders should receive different sanctions for similar transgressions is not always widely understood; the important thing is that such decisions must be made "within the rules of reason and justice, not according to private opinion" (Wilcox, 1972, p. 22). Unfortunately, despite their having been over 50 years of concern over the way the police exercise their discretion there has been little produced in the way of useable research (Greene 2014), so the debate is still dominated more by opinion than fact. Many commentators view the police service as exercising their discretion in an arbitrary manner, replacing the public interest with an unofficial 'police interest' criteria (Sanders, et al., 2010). But police discretion is not absolute; it is always fettered or limited to ensure it is used in a principled and consistent manner (Bronitt & Stenning 2011). Exactly how we limit and guide that discretion, and how effective different strategies are, is of huge importance to policing.

It is instructive to note that there has only ever been one other RCT testing the effectiveness of alternatives to prosecution in the UK. We have to go back to 1970 in Greater Manchester: it failed to show any significant benefit to reoffending rates in a youth 'caution plus' style of programme, over simple cautioning (Rose & Hamilton 1970). (It is of interest to note that, without the availability of modern IT, this trial ensured the fidelity of random assignment with a combination of dice rolling and sealed envelopes opened by a senior officer). In 2009 38% of all offences 'solved' by police in the UK were dealt with by way of out of court disposals (Criminal Justice Joint Inspection Report 2011). It is of concern that there is so little reliable evidence of effectiveness on which to base policy. This is the evidence gap which the Turning-Point experiment was seeking to fill.

It is also possible to study the decision-making of custody officers through the lens of decision theory. This is a rapidly expanding area of study and I do not have the space to go into it in any detail in this study – but there is one factor that stands out. Kahneman

(2011) proposes the human mind has two basic styles of decision-making: one fast and intuitive, the other slower and more deliberate. He shows that we are often less rational than we would like to think, and that our decisions are often shaped more by our assumptions and inbuilt biases than we would like to admit. This is relevant as the training and guidance we put in place to help officers identify suitable cases relies on officers taking a rational and impartial view of each case. Understanding how officers in this study make decisions, and whether this is deliberate or unconscious, may help us understand why the results are as they are: a question specifically looking at this issue has been included I the survey.

## The Custody Environment and Culture

It is instructive to bear in mind the context in which eligibility decisions by Custody Officers are made in this study. Whilst little has been published on this matter specifically Skinns proposes the values of custody staff can be considered in four broad domains. A survey of officers themselves shows that they didn't consider strong moral values as a desirable attribute for the role. When investigating the role of custody staff in compliance with the Police and Criminal Evidence Act 1984, the Home Office found no evidence of systematic bias. The survey in this study includes one question specifically testing officers' outlooks against the binary categorisation utilised by Muir in his study of police culture in 1997.

More recently Skinns gives us a useful framework to look at custody values with, breaking these down into four domains: the tension between due process and crime control; adversarialism due to our system of justice; the protection of human rights; and procedural justice and legitimacy  (Skinns 2011). This framework can give us a useful reference with which to explore some of the decision-making in this study.

A comparative study of custody officers in 2 forces (Waddon & Baker 1993) investigated the custody role from the perspective of stress and role profile, and also speaks to the issue of values or moral standards. Interestingly, a survey carried out as part of this study asked officers to rank 20 characteristics by their desirability to the custody officers' role. 'Knowing the law' came out as the one they valued most, with 'calmness' second. Interestingly 'has strong moral values' came low down the list – 17th out of 20 options.

Four years later the Home Office conducted a review of research on the Police and Criminal Evidence Act 1984 (much of which pertained to custody and the investigative process whilst in custody) and although found evidence of some variation in the use of discretion the authors decided it was not due to deliberate bias or any coherent strategy of targeting (Brown, 1997).

This study could open up questions of the dominant police culture in custody suites in Birmingham, and how that may impact on decision-making. Again, limitations of this report mean we cannot do more than scratch the surface: just one question has been included in the survey to investigate a specific aspect of police culture. Muir (1977) studied officer culture in an American city, and developed a view of their outlooks as either 'cynical' or 'tragic'. A tragic understanding is a unitary one, where all people share basic similarities and officers' empathy is high; a cynical understanding is a dualistic one, where officers perceive parts of the community as very different from themselves, with some less deserving of a service.

# Methods

In an attempt to produce a piece of 'real-world' research (Robson, 2002) the existing dataset on case by case decision-making created during the operational phase of the Turning-Point Experiment was utilised. This data captures rationales for rejecting ostensibly eligible cases and these were used to classify individual decisions as either appropriate (AR) or inappropriate (IR) rejections of randomisation. It also identifies individual officers and so allows calculations of the overall ‘appropriateness’ of decision-making for each officer. By adopting a flexible methodology and conducting a survey of all custody staff in West Midlands Police it was possible to compare the responses of those with high levels of IR against the total population of officers involved. Additional data collected on the training of the officers with the highest IR rates will help determine if this was a likely factor in their performance.

## Research Questions

In seeking to answer the main enquiry of whether the algorithmic triage model of case selection can lead to high levels of consistency and validity in decision-making, this research sought to answer five specific questions:

**Question One: What is the rate of rejection of randomisation for cases that were ostensibly eligible?**

**Question Two: Does the overall rate of rejection of randomisation vary over the life of the experiment?**

**Question Three: Can we differentiate between rejections that were appropriate and inappropriate?**

**Question Four: Are there patterns of high or low inappropriate rejection rates amongst the custody officers?**

**Question Five: Are there any known characteristics of the officers that correlate with rates of appropriateness of rejection?**

In order to investigate these questions, a multi-method design was used, combining existing data from the turning-point experiment, an officer survey and a targeted questionnaire of particular officers. A multi-method approach is useful because it allows the problem to be considered from different viewpoints: the combination of data gives a richer picture of both the outcomes and the possible mechanisms of causality.

All questions were examined using the data from the Cambridge Gateway; question five was examined by linking data from the Gateway with the survey and training questionnaire. After introducing the materials used to answer these questions the specific methodology for tackling each one shall be described.

## Materials

There are three main sources of material: data from the Cambridge Gateway, survey data, and a small training questionnaire. These data help identify the volume and reasons for rejection of randomisation as they changed over the life of the experiment, assisting us to determine if the algorithmic triage model we adopted was an effective one.  I shall detail each of these in turn:

## Cambridge Gateway Data

The Gateway was an on-line tool available 24/7 to custody officers at their terminals in the custody suites. It performed three functions:

Firstly it took officers through the key eligibility criteria to help them assess if the case was ostensibly eligible, recording their rationale.

Secondly, if the case was ostensibly eligible it asked if there was any exceptional reason as to why the case should not proceed to random allocation, and recorded their decision and rationale.

Thirdly it then randomly allocated eligible cases to either prosecution as normal or to make the offer of diversion to the offender. Eventually the system was developed to record the offenders' response and automatically generate a notification E-mail to the team that would be dealing with them.

This study utilises the following four variables from the Gateway:

i. Date
ii. Gateway allocated URN
iii. Custody Officer collar number (their identification)
iv. Narrative rationale for rejection of any ostensibly eligible case given by officer

The dataset used in this study commenced in March 2012, 4 months into the experiment, and runs through to the very end of the experiment in June 2014. It was chosen to commence at this point as the first four months of the experiment was a familiarisation and training period for custody staff to adapt to the new ways of working. This 32 month period contains three periods of missing data: 27 – 31 May 2012 (5 days); 20-23 Jan 2013 (3 days); and 21 March – 16 April 2013 (26 days). The missing data periods are 2.7% of the time span under study, so the impact of their loss is limited. This dataset has 14,681 entries made over that 32 month period.

Only the date and URN were automatically assigned during the process of inputting cases into the Gateway. Officers input their own personal identification (collar number) and errors in inputting have led to being unable to identify some decision-makers. 132 unique entries have been input into the Gateway collar number field to creating those 14,681 entries. On closer inspection 36 of these appeared to be errors in inputting as they are not valid collar numbers or relate to officers not connected to the experiment: each of these only have one or two entries each which supports the conclusion that they are typing errors. Of the remaining 96 identifiable officers six could be identified as having retired, leaving 90 officers identified as making Gateway decisions and still currently serving.

Likewise, the narrative account or rationale given by the officer as to why they have excluded a case that was ostensibly eligible is recorded by the officer themselves, in the operational custody suite environment. There are 244 records where this has occurred. These records are made 24/7, often during busy and pressurised periods, and this is reflected in the quality and level of detail recorded. Even so, only seven of these records have too little information on them to understand and assess the reason why they case was excluded.

It is possible that either inadvertently or deliberately, ostensibly eligible cases were rejected from random allocation (i.e. false negatives), but this fact was hidden due to it being recorded in the incorrect field. It is beyond the scope of this study to discover if this problem exists, or the scale of it. As project staff did not come across instances of this whilst managing the experiment it is felt that if it has occurred it will be in very small numbers, and therefore unlikely to impact on this study's findings.

An initial review of the reasons officers gave for rejecting ostensibly eligible cases, and an understanding of previous literature on police officers use of discretion, led the author to suspect there may be certain motivations behind the decisions. Suspicions were that these decisions related to how the officer valued rehabilitation over punishment; how effective at reducing reoffending they believed prosecutions were; whether they had a cynical or tragic world view (Muir 1977); or whether they made these decisions intuitively or more deliberately (Kahneman 2011). Hypothesising that officers views in these areas may correlate to the quality of their decision-making in the Turning-Point experiment led to the utilisation of a survey of the decision-makers.

## Custody Officer Survey

Between February and May 2014 a survey of all the custody officers in West Midlands Police was conducted as part of a larger piece of work, unrelated to this thesis. This gave an opportunity to probe officers' positions on issues that could have had an impact on the experiment. This has the potential to add more explanatory value and to give a deeper understanding of why decisions were taken, and how we may be able to influence this in the future (Foddy, 1993). This is in line with the desire for this study to be a piece of real world research (Robson 2002) that has an impact on future operational behaviour.

Questions were developed in three broad categories: questions relating to officer attributes (such as age, gender, educational level); questions relating to officers views on offenders, prosecution and decision-making; and questions relating to their understanding and support for Turning-Point (only answered by those staff that had worked on the project). These questions are detailed in Appendix B.

### *Sample*

West Midlands Police operate 11 custody suites across the force area, with 120 – 140 officers assigned to them at any one time (135 at the time the survey commenced). 43% of these are in one of the five Birmingham custody suites that were running the experiment. As the experiment was live for 36 months there was a regular turnover of staff.  The five custody suites involved in the experiment had approximately 58 custody officers assigned to them at any one time.

In total 136 responses to the survey were returned, 119 from officers still serving in custody, giving a response rate of 88.1% for serving custody officers. 17 custody officers that had left the custody department by the time the survey was conducted also completed the survey. 90 identifiable officers had entered cases into the Cambridge Gateway whilst working on Turning-Point. 76 officers identified in the survey that they had at some time worked on Turning-Point (see Figure 4). 70 of these identified themselves individually in the survey, 6 declined to.  Therefore there was an 84% return rate for the officers who had worked on Turning-Point and used the Gateway. If we take into account that six of these officers had retired or left the force before the survey was conducted, we have a response rate of 90%. Such a high response rate was achieved because the project manager personally followed up requests for completion with personal contact. It is noted that this approach could have had an inadvertent effect on the responses given by staff due to the managers lack of independence from the workplace and the issues discussed in the survey (Foddy, 1993),but it was felt that the benefit of increased reliability achieved from a high response rate outweighed this threat.

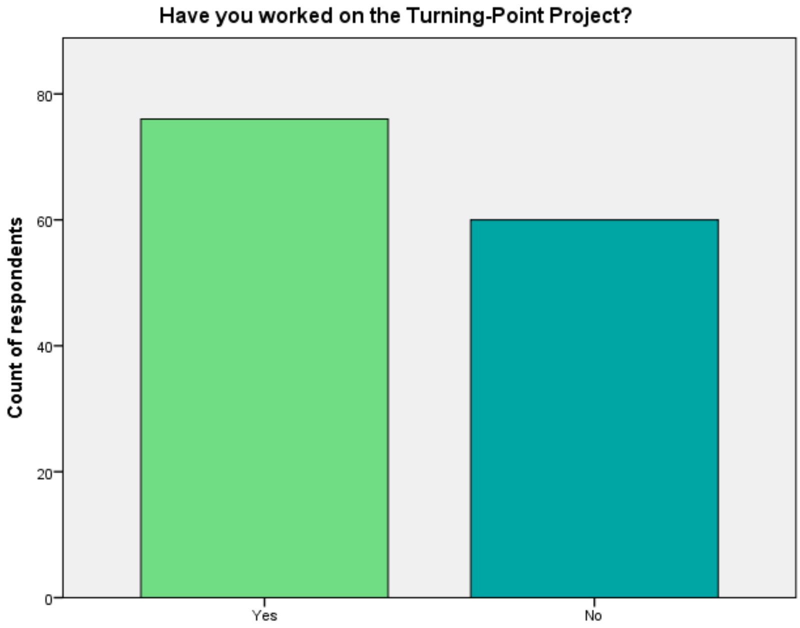


Figure 4 Survey Responses Regarding Turning-Point Experience

### *Procedure*

This survey was conducted during 10 scheduled training days between February and May 2014. All custody staff force-wide should attend one of these 10 training days, and so theoretically all should have a request made of them to complete the survey. The Project Manager or a member of the project team attended 9/10 of these training days to present, explain and run the survey. This was deliberately done to ensure a high response rate, and to reassure the participants as to the confidentiality of their responses, an important factor for any survey (Foddy, 1993). The first two sessions were run on paper;

the subsequent eight sessions were completed on-line. The survey evolved over the first three deliveries, based on practioner and academic feedback; therefore not every respondent answered exactly the same questions. This allowed questions participants' found ambiguous to be removed or reworded in a circular process of sampling and refinement (Creswell, 2007). Once finalised the survey was placed on-line, for two reasons. Firstly this reduced the resources required to collect the data in a digital format thus reducing the researchers' workload; secondly it allowed the survey to be completed by officers who weren't attending the training days as they had left the department.

## Training Questionnaire

The research identified a small cohort of officers (N=8) with high rates of inappropriate rejection. The possibility of this being due to these staff having had less or even no training in relation to the experiment was worthy of exploration. No records had been kept of attendance at training events and this issue had not been part of the survey. As numbers were small (eight officers, two of which had retired) each officer was contacted individually by the project manager (via telephone) and asked to recall what training they had participated in and whether they had had any additional, bespoke training or development from the project team. This was recorded on a brief questionnaire, the results of which are recorded in Appendix F.

## Research Questions

In seeking to answer the main enquiry of whether the algorithmic triage model of case selection can lead to high levels of consistency and validity in decision-making, this research sought to answer the five specific questions with the following methodology:

### Question One: What is the rate of rejection of randomisation for cases that were ostensibly eligible?

This question is central to judging the overall external validity of the Turning-Point Experiment. Data is available From the Cambridge Gateway on the total no of cases evaluated, 14,681, and the total number of cases eligible but where randomisation was rejected: 244. The same dataset can also identify the number of cases deemed eligible for random allocation and then assigned: 681 cases. Cases ostensibly eligible for random assignment consist of these 681 plus the 244 where random allocation was rejected: a total of 925 eligible cases. These cases passed all the screening questions, and so were theoretically eligible. The rate of rejection as a percentage of all eligible cases is a more useful measure of rejection than the simple numbers rejected as it puts the issue into context. This is also a figure that can be taken and compared to other experiments, and therefore used as a benchmark in this research.

It is also important to bear in mind that the experiment was built to allow rejections in appropriate circumstances as this was seen as a necessary safeguard, and allowed staff to protect the force's reputation. Situations in which such rejections were appropriate were for instance where particularly vulnerable victims had been targeted, or offending was related to ongoing prosecutions. Therefore as well as looking at the overall rejection rate the research also distinguished between appropriate and inappropriate rejections. This allowed for more nuanced information about the quality of officer decision-making, and focused on the issue of potential improvement for future experiments.

### Question Two: Does the overall rate of rejection of randomisation vary over the life of the experiment?

This question investigated changes in rejection rates over time, and made it possible to examine any possible causes for such variation. Would any changes correlate with improvements made to the decision-making process, the involvement of staff or training approaches? With an experimental lifespan of nearly three years the most appropriate scale to investigate rejection rates was monthly, absorbing daily and weekly fluctuations and giving a more consistent output. Both overall rejection and inappropriate rejection rates were calculated over time to give the reader a better understanding of the issue.

Monthly rejection rates were calculated by calculating the percentage of rejected or inappropriately rejected cases of all eligible cases, month by month. Standard deviation was calculated to measure the variability in these figures.

### Question Three: Can we differentiate between rejections that were appropriate and inappropriate?

This is a particularly important question as Turning-Point deliberately built in the facility for custody staff to reject cases in order to ensure the forces reputation was protected. A proportion of these rejections will have been for sound reasons and will therefore reflect good decision-making. We need to be able to separate these from the inappropriate decisions: it would be a false assumption to think that all rejections are a bad thing. The only data that we can use to determine this is the rationale recorded in the Cambridge Gateway for the 244 rejected cases.  Whilst a small number of rationales are recorded poorly (N=7, 2.5%), the vast majority, 97.5%, have enough detail to answer this question. Of course, an assessment can only be made of what is recorded, so if there were valid reasons for rejection, but they are not recorded on the Gateway, then they will be assessed as inappropriate.

There is an issue in measuring the 'appropriateness' of rejections, in that this was a subjective decision. Most of the eligibility criteria were objective and clear cut, such as: number of previous convictions; whether the offence required the Director of Public Prosecutions (DPP) authority to charge; or it being a drink-driving offence. However, other criteria were less objective: the likely sentence on conviction for example (see Appendix One for the definitive list). Most subjective of all was our advice to officers to reject ostensibly eligible cases if taking them forward to randomisation, for whatever reason, was likely to damage the reputation of the force. It is impossible to give prescriptive guidance on this question, so we were highly reliant on 3 factors: providing guidance, examples and training that illustrated to officers the threshold of acceptability that we were looking to set; following this up with feedback, coaching and support, and having intelligent, experienced and confident decision-makers able to assimilate this advice and apply it consistently.

To ensure that the benchmark used to assess 'appropriateness' had some level of validity itself the author devised a measurement framework utilising 4 assessors rating all 244 of the rejected cases, independently of each other. After careful analysis of the rationales for these cases, a five point grading measure was constructed that would capture all responses: Table 1. Rationales were scored in this way to facilitate later analysis of appropriateness. Whilst there was little to differentiate appropriate decisions from one another, inappropriate decisions could be categorised as in appropriate or clearly (strongly) inappropriate. This allowed weighting to be built into the scoring and officer performance to be better differentiated.

Table 1 Appropriateness Assessment Categories and Weighted Scoring

| | |
|---|---|
| 1) Too little information to assess | 0 |
| 2) Error – case should have been rejected for standard reason | 0 |
| 3) Appropriate Rejection | -1 |
| 4) Inappropriate Rejection | +1 |
| 5) Clearly Inappropriate Rejection | +2 |

Four assessors where selected to carry out the task: the project manager (the author), the field researcher, the project custody lead and a custody officer that had been closely involved in developing the eligibility criteria during the early stages of the experiment. Assessors had access to all the information provided during training, all advice and guidance on the matter, and an additional set of 'Assessors Guidance Notes' (Appendix C). These had been produced by the author to assist in consistent interpretation of the recorded rationales.

The four sets of responses were then compared for inter-rater reliability: measurements used were Cronbach's Alpha and Cohen's' Kappa: "these statistics are quality indicators of the measurement reproducibility" (Gwet 2012, p.vii).

Each of the four assessors' scores was calculated for each of the 244 decisions to reject randomisation, and a mean calculated for each decision across all 4 assessors (by summing up all 4 scores for each case and dividing it by four): giving a 'mean Inappropriate Rejection (IR) score'.

Once each decision had a mean IR score of between -1 and +2, a threshold was determined to allow the classification of decisions as either appropriate or inappropriate. Decisions with a mean IR score of more than 0.5 were classed as inappropriate, the rationale being that their score was closer to inappropriate than zero. Decisions with a mean IR score of less than zero were classed as appropriate (an appropriate decision could only score -1). Scores of 0 to 0.5 were removed as these reflected errors or decisions that could not be scored, neither being a measure of appropriateness. The mean IR measure is more reliable than taking any one assessor's scores as it is free of individual bias.

## Question Four: Are there patterns of high or low inappropriate rejection rates amongst the custody officers?

If we can understand any patterns of rejection of randomisation by officer then we are more likely to be able to identify reasons or drivers for inappropriate rejection. Given the earlier coding of the assessments (the creation of an IR score) it is possible to identify how inappropriate each decision is, on a scale of -1 to +2. We are able to identify the officer making each decision and are therefore also able to aggregate all the scores for all of their inappropriate decisions, creating a 'composite IR score'. For example, this means that if an officer dealt with three cases, and two of them were coded as clearly inappropriate (with a score of 2 each) and one was coded as appropriate (with a score of -1), this officer's composite IR score would be 3.

However, this summing up of IR scores has the danger of penalising those officers that have put in the most cases: numbers of cases input to the Cambridge Gateway per officer vary greatly, from 2 to 651. The mean is 168 (standard deviation: 140.3)

Therefore, if we divide each officer's composite IR score by the number of cases they submitted to the Gateway we get a better comparison of the performance of each individual officer. This is called the officer's 'Inappropriate Rejection (IR) Rate'. By plotting this on a histogram we can identify if there are a number of officers worth investigating further due to their high IR rates. This can be done by introducing information from the survey and may lead to conclusions relevant to issues of implementation or training.

## Question Five: Are there any known characteristics of the officers that correlate with rates of appropriateness of rejection?

Training and managing staff involved in experiments can be resource intensive, costly and problematic. If we are able to identify characteristics of staff or training programmes that correlate with better performance then we have a starting point for discovering if they actually cause better performance. Such information may help researchers to run more efficient experiments.

Analysis of the rationales recorded in the Cambridge Gateway may identify recurring themes or patterns. The Gateway data was augmented with data from both the survey and the targeted questionnaires of the high IR officers. The survey data is in the main ordinal level data, consisting of Likert scale responses. There is also one scale level data response (Q34), and three nominal level data responses (Q2, Q15 and Q29). These responses can be compared between two groups of officers: those that made inappropriate rejections, and those that did not. The Mann-Whitney U Test is the best measure of differences between these two groups on the ordinal and scale level responses within the survey (Field 2013). For the three survey questions with nominal level responses the best test will be the Chi Squared Test (Field 2013).

We are also able to see if there is any significant difference between the group of officers involved in Turning-Point (most of which will have received training and coaching in the relevant decision-making), and those from the rest of the force. We will look to see if this training appears to have had any impact of officers' views as recorded in 15 of the survey questions, again using the Mann-Whitney U Test. 15 slightly different questions will used, none of which utilise nominal level responses, therefore the one test will be sufficient.

Finally we will investigate the overlap between the two groups of officers: the IR and AR groups. Understanding if these are separate groups of staff, or if in fact we find the same officers making both types of decisions will help us determine if performance is more likely to be related to personal or task factors. Again this will be done using the Mann-Whitney U-Test and the Chi-Square Test.

# Results

The results of the analysis are presented question by question in a natural progression moving from rejection rates through measures of appropriateness of decisions to using the survey data to try and understand why officers reject cases. This data helps determine the effectiveness of the algorithmic triage model utilised in Turning-Point, and how it may be further improved.

## Question One: What is the rate of rejection of randomisation for cases that were ostensibly eligible?

A total of 14,681 were assessed for eligibility and entered into the Cambridge Gateway: 925 were deemed to be ostensibly eligible. Of these 244 were rejected by officers and 681 assigned. This gives an overall rejection rate of 27.16% of all eligible cases, or 1.66% of the total number of cases assessed.

However, the experiment wanted staff to reject some cases, so it helps to look at the proportion of *inappropriate* rejections. There were 110 inappropriate rejections: 45.08% of all rejections. These were 11.90% of all eligible cases, averaged over the experiments life span. There were also 7 cases (2.94% of rejections) where it was impossible to tell whether the decision was appropriate or not. If all of these were in fact inappropriate rejections then our IR rate rises slightly to 12.66% of eligible cases, or 0.79% of the total number of cases assessed.

## Question Two: Does the overall rate of rejection of randomisation vary over the life of the experiment?

The overall throughput of cases rises during the early stages of the experiment, as the geography of the project increases, then drops slightly during the latter part of the project, approximately in line with reductions in custody throughput experienced force-wide (see Figure 5).

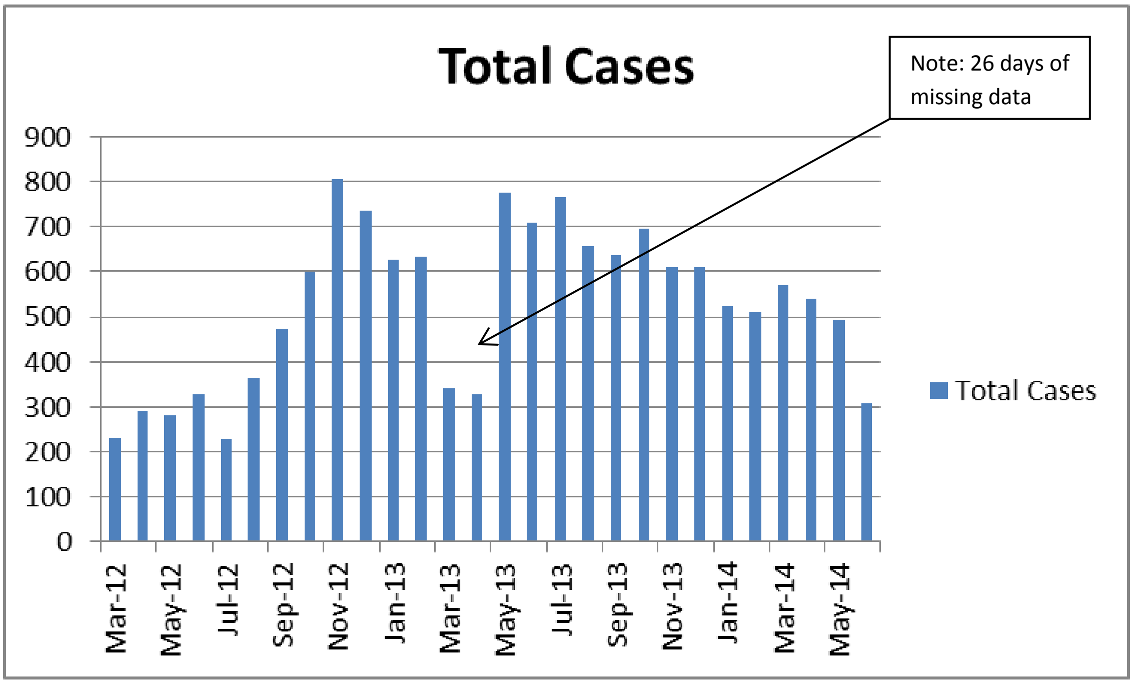


Figure 5 Monthly Throughput of Total Cases into the Cambridge Gateway

Rejections vary over the 32 months of data from one to 21 rejections. The average number of rejections per month is 8.71. This figure is very variable (standard deviation: 5.12). The rate of rejections as a proportion of all ostensibly eligible cases is a more informative figure: Figure 6.

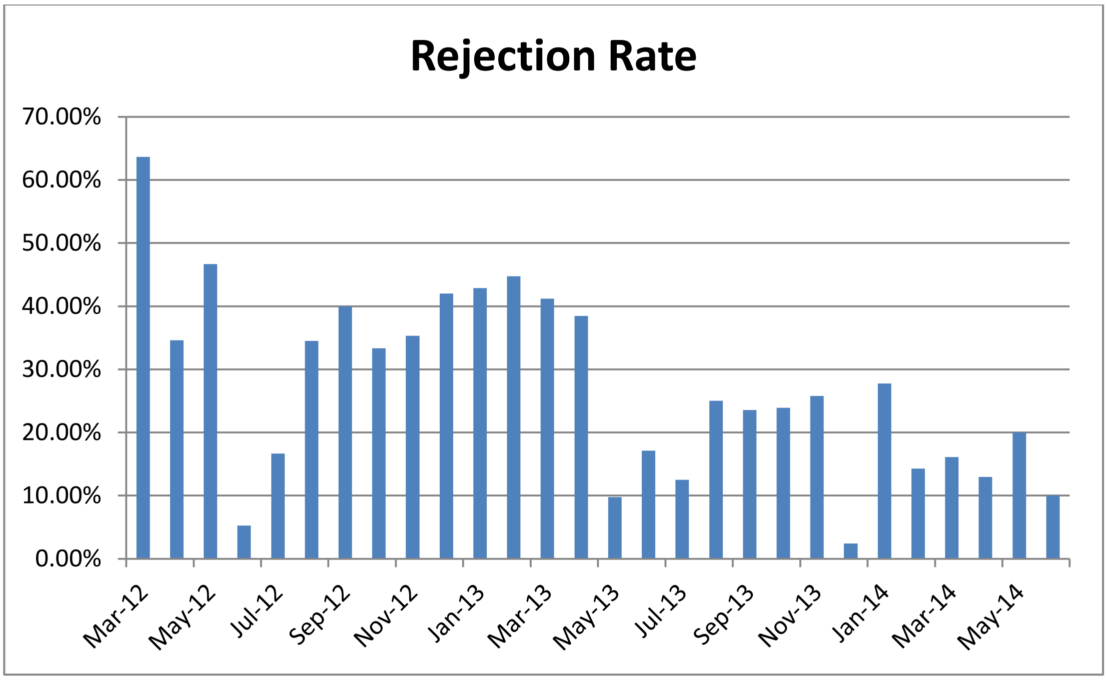


**Figure 6 Monthly Rejection Rate (%Rejected of Eligible Cases)**

However, we are aware that only a proportion of the rejections are inappropriate. It is therefore of interest to plot the rate of inappropriate rejections. This speaks better to the issue at the centre of this study – the quality of decision-making pre random assignment. Figure 7 illustrates this. It will be noted that there are three months when no inappropriate rejections are made, and a peak of inappropriate rejections (42%) of all eligible cases early on in the experiment, in March 2012. The average rate of inappropriate rejections over the entire time-span studied is 14.73% (standard deviation: 11.74%). There is a clear downward trend in inappropriate rejections over the life of the experiment, as evidenced by the linear trend line.

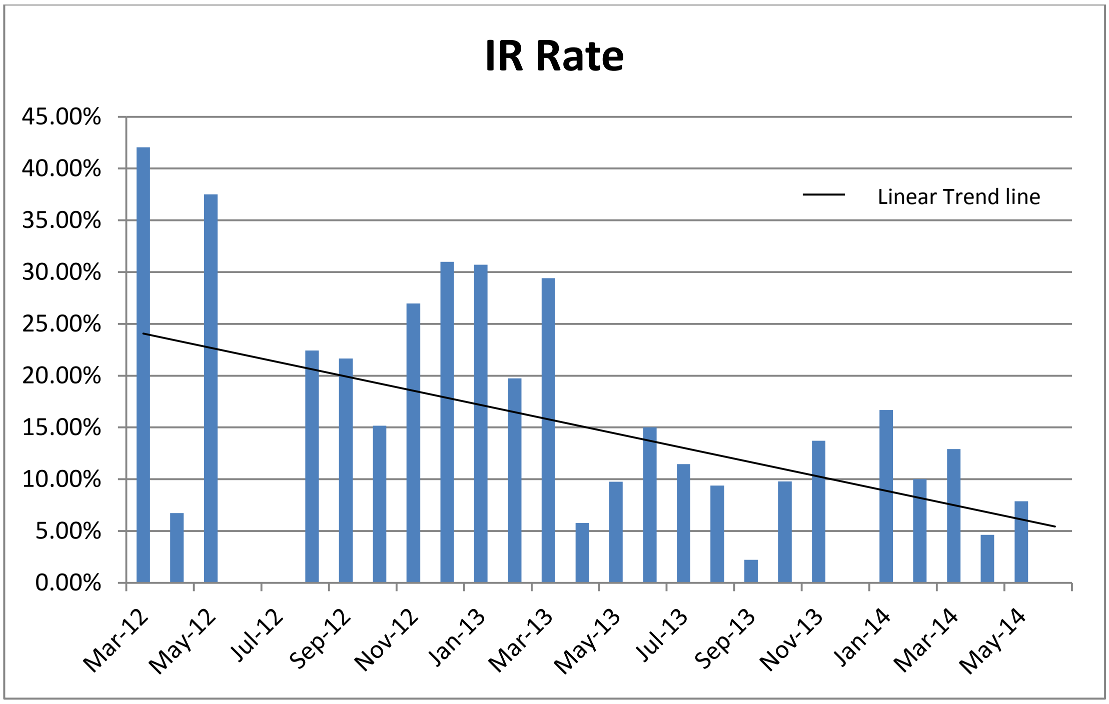


**Figure 7 Inappropriate Rejection Rate (IR/Eligible Cases)**

In summary, there is a clear pattern of reductions in both overall rejections and in the inappropriate rejection rate during the life of the experiment. However this against a background of high levels of standard deviation. This shows an improving picture over the experiments life-span. In the discussion we will explain how this could be influenced by the training and development regime implemented for this project.

## Question Three: Can we differentiate between rejections that were appropriate and inappropriate?

It is essential to break down the complete list of rejections to determine which were broadly in line with the objectives of the experiment and necessary to protect the force reputation, and which weren't. We are using this measure as an indicator of the quality of decision-making, in seeking evidence on the impact of using the algorithmic triage model.

By the use of 4 independent assessors scoring each decision against a 5-point scoring matrix we were able to determine the proportions of decisions in each category: Table 2.

Table 2 Proportions of all Rejections in each Assessment Category

| | | |
|---|---|---|
| Too little information to assess | 7 | 2.86% |
| Error – case should have been rejected for standard reason | 87 | 35.66% |
| Appropriate Rejections | 40 | 16.39% |
| All Inappropriate Rejections | 110 | 45.08% |
| Total | 244 | 100.00% |

### Comparing the Four Independent Assessments

Four assessors independently assessed each of the 244 rejections into one of the five categories. Each assessor's results are contained in Appendix D. The use of multiple assessors was necessary for two reasons: firstly it enables a comparison of the 4 results to see how well they correlate. The decision is subjective, so to see if key members of the project team agree is of relevance. This may indicate how consistent message operational decision-makers were getting from the leadership team, helping one judge how consistent we can realistically expect our decision-makers to be. Secondly it allows the development of a score for appropriateness of each decision that is not based on just one assessors view, but on multiple views. Such a measure will have more validity and be less open to criticism of bias (Gwet 2012).

In order to calculate correlations between the four assessors, the scores for just the three measures of appropriateness were utilised. A bipolar scale was used (with 'Inappropriate' and 'Clearly Inappropriate' being grouped together) to ensure that the comparison is of appropriateness assessments alone. The results of the Spearman's rho test are shown in Table 3.

Correlation coefficients are good between all assessors: ranging from 0.593 to 0.837 (Scores 0.40-0.59 are described as showing a moderate correlation; 0.60-0.79 are described as strong and scores above 0.80 are described as showing a very strong correlation, Field 2013). We can therefore see that the correlations are generally 'strong'. Significance is very good, consistently at p<.001. The assessments of Rater A1 and Rater A4 show a particularly strong correlation: 0.837. These assessors were the two staff most closely involved in the experiment: the project manager (the author), and the field researcher.

**Table 3 Correlations between Four Independent Assessments of Appropriateness**

**Correlations**

| | | Rater-A1 | Rater_A2 | Rater_A3 | Rater_A4 |
|---|---|---|---|---|---|
| Rater-A1 | Correlation Coefficient | 1.000 | .657** | .660** | .837** |
| | Sig. (2-tailed) | . | .000 | .000 | .000 |
| | N | 244 | 244 | 244 | 244 |
| Rater_A2 | Correlation Coefficient | .657** | 1.000 | .593** | .633** |
| | Sig. (2-tailed) | .000 | . | .000 | .000 |
| | N | 244 | 244 | 244 | 244 |
| Rater_A3 | Correlation Coefficient | .660** | .593** | 1.000 | .638** |
| | Sig. (2-tailed) | .000 | .000 | . | .000 |
| | N | 244 | 244 | 244 | 244 |
| Rater_A4 | Correlation Coefficient | .837** | .633** | .638** | 1.000 |
| | Sig. (2-tailed) | .000 | .000 | .000 | . |
| | N | 244 | 244 | 244 | 244 |

**. Correlation is significant at the 0.01 level (2-tailed).

Table 4 Cronbach’s Alpha across Four Assessors

**Reliability Statistics**

| Cronbach's Alpha | N of Items |
|---|---|
| .890 | 4 |

There are better tests of the relationship between assessors’ scores. Cronbach’s Alpha is a measure of internal consistency and is 0.890 for these four assessments – that is on the cusp of a good to excellent grade (see Table 4). Cohen’s Kappa is a statistical measure of inter-rater reliability and will therefore be the best measure of consistency in this situation. The inter-rater reliability for the assessors here is shown in Table 5. In interpreting Kappa, .41 - .60 is moderate, .61 – .80 is considered substantial agreement and .81 – 1.0 is considered almost perfect agreement (Field 2013). Significance is again p<.001, showing that this is highly unlikely to be a coincidence. We have generally moderate agreement between our four assessors in their assessment of appropriateness of the 244 decisions to reject randomisation (and substantial agreement between two assessors again).

Table 5 Kappa for Four Assessors

| KAPPA | Rater 1 | Rater 2 | Rater 3 |
|---|---|---|---|
| Rater 2<br>Sig. | .551<br>.000 | | |
| Rater 3<br>Sig. | .551<br>.000 | .431<br>.000 | |
| Rater 4<br>Sig | .787<br>.000 | .556<br>.000 | .498<br>.000 |

In summary, we can show generally moderate levels of consistency between our four independent assessments of whether the reasons recorded on the Cambridge Gateway to justify rejecting an otherwise eligible case are appropriate reasons or not. This led to the development of a measure of appropriateness for decisions based on a combination of the scores from the four assessors, superior to relying on any one individual's scores.

## Assessing 'Appropriateness' of Rejections

Having established that we have moderate consistency in assessing for appropriateness we can now combine the four assessments to produce a mean score for each decision. By allocating values as per Table 1 we are able to provide a greater range of scores to better differentiate between decisions. This discounts the categories that are not measuring appropriateness, whilst also reflecting the degree or strength of appropriateness. A mean of the four assessors scores for each rejection are then used to give an overall measure of appropriateness to each decision: the 'mean IR score': Table 6.

Table 6 Frequency of Decisions Appropriateness Score

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | -1.00 | 4 | 1.6 | 1.6 | 1.6 |
| | -.75 | 7 | 2.9 | 2.9 | 4.5 |
| | -.50 | 13 | 5.3 | 5.3 | 9.8 |
| | -.25 | 16 | 6.6 | 6.6 | 16.4 |
| | .00 | 73 | 29.9 | 29.9 | 46.3 |
| | .25 | 10 | 4.1 | 4.1 | 50.4 |
| | .50 | 11 | 4.5 | 4.5 | 54.9 |
| | .75 | 21 | 8.6 | 8.6 | 63.5 |
| | 1.00 | 28 | 11.5 | 11.5 | 75.0 |
| | 1.25 | 20 | 8.2 | 8.2 | 83.2 |
| | 1.50 | 13 | 5.3 | 5.3 | 88.5 |
| | 1.75 | 12 | 4.9 | 4.9 | 93.4 |
| | 2.00 | 16 | 6.6 | 6.6 | 100.0 |
| | Total | 244 | 100.0 | 100.0 | |

Each decision now has a score which lies on a continuum of ‘appropriateness’ which can be used for categorisation. Appropriate decisions are those with scores below zero, and inappropriate decisions are those with a score over 0.5. This results in 110 of the 244 decisions to reject randomisation as being categorised as ‘Inappropriate’ (45.1%):

highlighted in yellow on Table 6; and 40 cases being categorised as ‘Appropriate’ (16.4%): highlighted in blue.

94 cases (38.5%) are neither, being classed as errors as they should have been excluded for other reasons earlier in the decision-making process, or there just being too little information to make an assessment on. In these cases the decision to exclude is correct; it is just recorded in the incorrect field.  To categorise these I have reverted to the assessment of Assessor A_1 (the Author/Project Manager), who identified 7 cases (2.9%) as ‘Too Little Information’, and 87 (35.7%) as ‘Error’. Figure 8 shows the spread of cases by their scores.

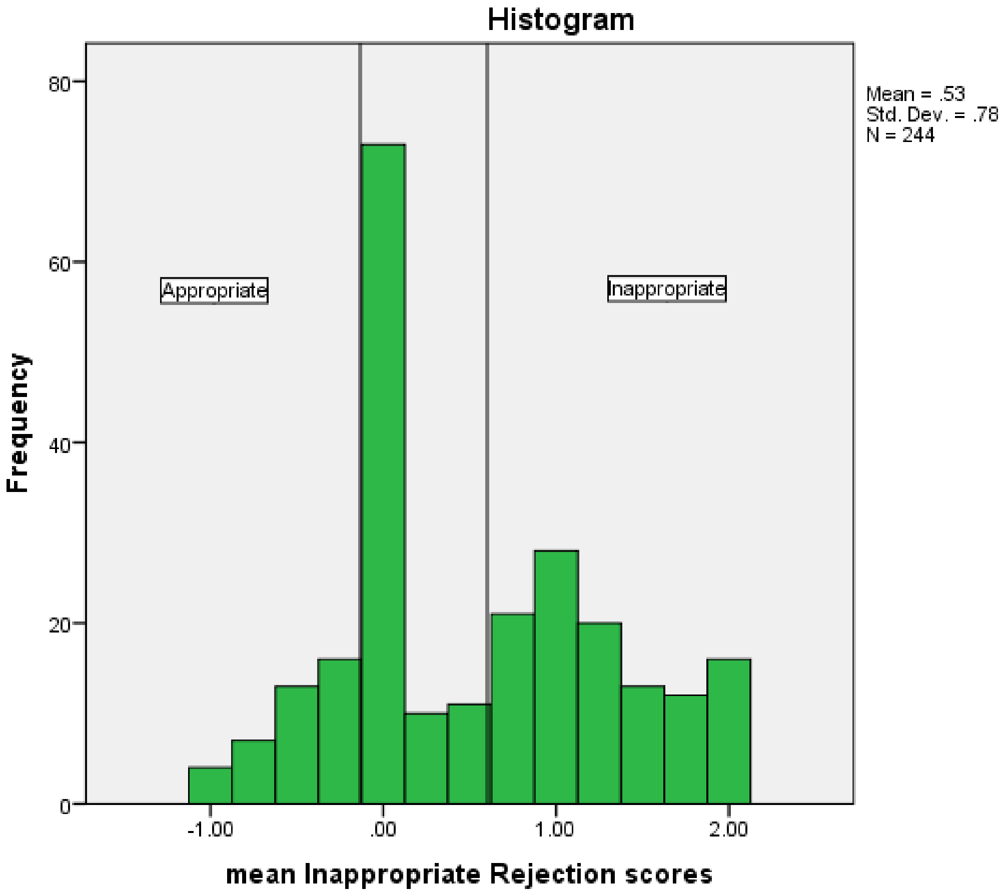


Figure 8	Spread of Appropriateness Scores across all 244 Rejections

The ratio of 'appropriate to inappropriate decisions, is 1:2.75. It is a valid question to ask if this ratio is acceptable in the experimental field – is the benefit of having more subjective selection criteria worth the loss of nearly three times as many eligible cases as those excluded appropriately. We will return to this question in the discussion.

## The Appropriate Rejections

Analysis of the 40 cases assessed as appropriate rejections shows some key themes emerging: issues of safeguarding others, usually the victim, dealing with an abuse of trust, and offences against the vulnerable crop up regularly. Safeguarding issues are cited in six cases, with four of these safeguarding children and two safeguarding adults. This overlaps with cases of specific vulnerability: five cases involved elderly victims, often deliberately targeted. One rationale was due to the vulnerability of a child. Often in these cases the abuse of trust by the offender was mentioned as a key part of the rationale: this occurred in six of the 40 cases. Other issues include where deportation was required independent of the outcome of any court case (i.e. the offender was already liable to be deported for other reasons); three cases where the offender was already on the Turning-Point Project and so this further offence not only deemed them illegible for the project but also led to prosecution in the original case; three cases involved serving or past police officers/employees and the decision-maker felt that diversion was both inappropriate but also open to criticism for lacking in independence; finally there were two cases where the instant offence was connected to other court cases and it was felt appropriate to allow the court to deal with the entirety of the offending (including one case of witness intimidation).

Some cases of abuse of trust appear in this group, as appropriate rejections – for each the rationale makes a good case for exceptionally excluding them even though the custodial threshold for exclusion was not met.

## Themes of 'Inappropriateness'

Analysis of the 110 rejections identified as being inappropriate identified certain themes or recurring issues: Table 7.

Table 7 Themes of Inappropriate Rejection (IR) Rationales

| Theme | No of cases mentioned | Percentage of all IR cases (N = 110) |
|---|---|---|
| Lack of Remorse/denial/attitude | 34 | 30.9% |
| Wider offending | 18 | 16.4% |
| Assault Against Police | 12 | 10.9% |
| Too serious | 11 | 10% |
| Breach of Trust | 10 | 9.1% |
| Vulnerable victim | 9 | 8.2% |
| Preplanning/multiple offenders | 7 | 6.4% |
| CPS decision | 4 | 3.6% |

These themes sometimes contain a variety of different, but similar or related reasons. Officers' rationales also tend to be short and so some interpretation is required. Many rationales contained multiple themes. Eight themes have been identified and we will briefly explore these in turn:

### *Lack of Remorse/Denial/Attitude*

The largest theme that appears to stand out from the data is a theme around the attitude or response of the offender when being dealt with by police: being mentioned in nearly a third of the cases inappropriately rejected (N=34). This includes where the offender shows no remorse for their offending; hasn't made any admissions; frustrates the enquiry (e.g. giving false details); the officer suspects the offender wouldn't comply with the alternative sanction for some reason; or there is something else disagreeable about their 'attitude' (such as being aggressive or unreasonable). This includes cases where the justification is that the offender has previously not complied with alternatives to court. It is obviously important to keep cases such as these in the experiment as we need to know how both treatments work with this group of offenders for a fair comparison. None of these reasons were part of the selection criteria. It appears that a value judgment is being made here as to whether the offender 'deserves' the alternative to court, probably based on the belief that the court will be more punitive that the alternative, and that punishment is what is needed (or deserved) in this case – possible evidence of Muirs' cynicism.

### *Wider Offending*

16.4% of cases inappropriately rejected included the wider offending of the offender as part of the rationale for rejection (N=18). This category does not include proven offending of the individual, as anything more than one previous conviction would automatically exclude them, but refers to officers' suspicions or the fact that there are multiple current offences. So, for example, cases of suspected drug dealers only being charged with possession offences; harassment charges that reflect multiple incidents; and multiple thefts being dealt with simultaneously are included. This includes a small number of cases where the offender has ongoing drink or drug problems that are identified as underpinning their offending – cases that appear to have been ideal for the intervention being tested under the project.

### *Assault against Police*

12 cases (10.9% of the 110 inappropriate rejections) were rejected outright due to being an assault against a police officer. This was a controversial matter with operational staff and much thought had been put into whether to automatically exclude these cases. Considering the wide variety of 'seriousness' of these offences, the Project Board had decided that it was essential to know whether this approach could help reduce assaults against officers, so they were deliberately included. Serious cases would be excluded by the high likelihood of a custodial sentence. This was an unpopular policy decision with operational staff and whilst some cases were taken forward to randomisation, we can see that discretion was also used to avoid randomisation.

### *Too Serious*

11 cases (10% of all inappropriate rejections) were seen as too serious to include – often because of the value involved (be it theft, damage or of drugs involved), or because of the surrounding circumstances of the offence. Each of these cases had been identified as *not* meeting the exclusionary threshold of likely to result in a custodial sentence, yet the officer still decided to exclude them, feeling they were so serious only a prosecution would be acceptable to the public. Based on the rationale recorded on the Gateway, the

assessors did not agree with that assessment: it is possible the assessors place a higher value on protecting the experiments integrity.

### *Breach of Trust*

10 cases (9.1% of all inappropriate exclusions) had a breach of trust mentioned as part of the rationale for rejecting randomisation. Although this is in effect a subset of the previous 'seriousness' category it is worthy of mention in its own right, as it was often raised as a concern by operational staff.  Such cases were much more likely to reach a custodial threshold than similar cases without this particular aggravating factor. In these 10 cases however, all had been assessed as not reaching the custodial threshold and none had an adequate explanation as to why they were unsuitable, and so were assessed as inappropriate rejections.

### *Vulnerable Victim or Organised Offending*

Nine cases (8.2%) were rejected inappropriately with a rationale explaining that part of the rationale was the vulnerability of the victim, usually because of their young age. There were no claims that a prosecution would give necessary protection (e.g. through bail conditions), and again the custodial threshold had not been met. These decisions could be more of a moral judgement on the offending rather than a substantial reason for rejection. Seven cases (6.4%) were excluded where rationales' mentioned elements of preplanning or offenders acting together as the justification. Again these are aggravating factors, but the cases were still assessed by the officers as not reaching the custodial threshold. It is therefore difficult to view these factors as appropriate reasons for rejection.

### *CPS Decision*

Only a small number of cases (four: 3.6%) gave the rationale of a CPS authority to charge as the reason for rejection. The correct process was for the selection criteria to be assessed *before* putting the case to CPS, and where this had inadvertently been missed then to go back to CPS seeking the decision be overturned: there was no evidence this had occurred in these cases.

### *Quotes*

Appendix E lists a selection of direct quotes from the rationales recorded on the Cambridge Gateway, focussed on those that show a lack of understanding of the objectives and rationale for the experiment. These evidence cases that the alternative approach being tested (heavily reliant on rehabilitation, reparation and restoration) may have been more successful at dealing with. These include ongoing problems or disputes; offenders not recognising the impact of their actions; offending within care or children's homes by residents; and offenders with substance abuse issues.

Analysis has uncovered eight key themes behind the inappropriate rejections. These appear to indicate some moral judgements may be being made by officers, and that these officers may, at times, be failing to discern the difference between the criteria required for the experiment, and those for the application of standard out of court disposals. However, the proportion of cases excluded in this way is not so high as to cause serious validity problems. The scale of the problem also diminished significantly over the life of the experiment.

## Question Four: Are there Patterns of High or Low Inappropriate Rejection Rates Amongst Custody Officers?

The mean number of entries in the Cambridge Gateway for custody officers is 168 (Standard deviation: 142.8). There is a high degree of variation due to the long length of service in the relevant custody suites by a handful of officers, and the occasional shifts in the project area by custody staff permanently posted elsewhere in the force.

### Table 8 Breakdown of Decision Types by Numbers of Officers Making Them

| Type of Decision | N - Decisions | N - Officers |
|---|---|---|
| Total cases assessed | 14,681 | 96 |
| Rejections | 244 | 60 |
| Inappropriate Rejections | 110 | 41 |
| Appropriate Rejections | 40 | 27 |

Table 8 shows that 62.5% of the officers involved in Turning-Point made the 1.6% of decisions that rejected ostensibly eligible cases. 42.7% of the staff involved made the 0.75% of inappropriate rejection decisions. The 4.3% of all eligible cases that we assessed as appropriate rejections were made by 28.1% of the officers involved. Overall the small proportions of rejected cases are not concentrated in an equally small number of officers: there is a wider spread.

## Patterns of Inappropriate Rejection by Officer

The 110 IR decisions were made by 41 officers. Each IR decision was graded on a scale of 0.50 – 2.00, where 2.00 is the most inappropriate. It is useful to plot the sum of these scores by officer responsible to observe the totals: Figure 9. This illustrates the 'total inappropriateness' score for each officer, but is independent of the volume of cases they input in total.

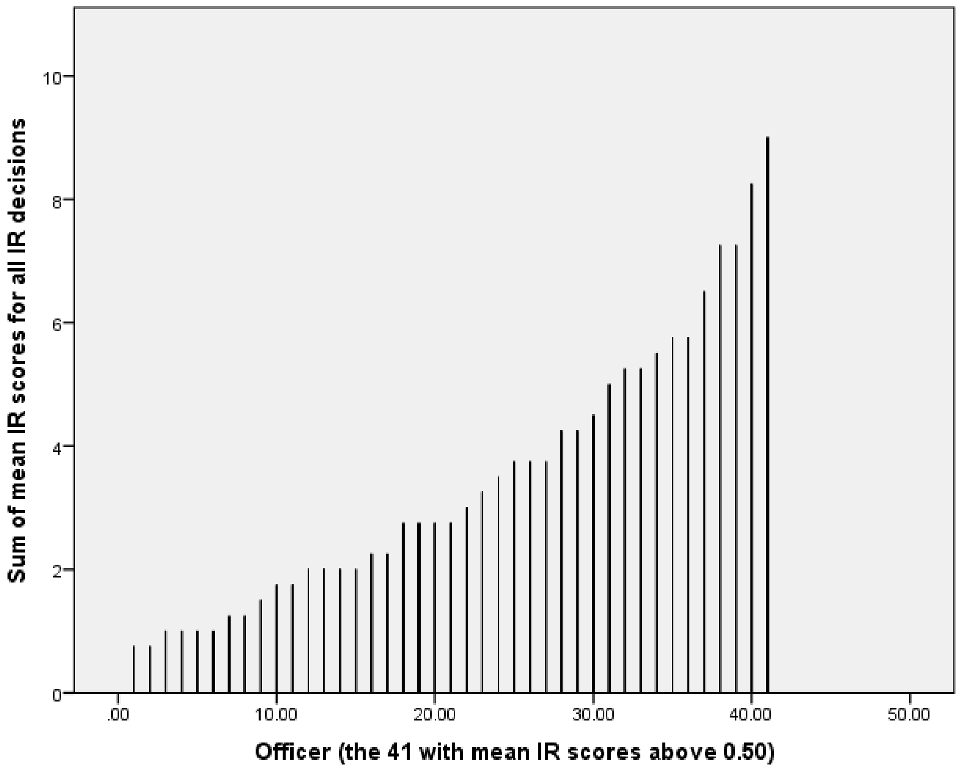


Figure 9 Officers Composite Inappropriate Rejection (IR) Scores

Figure 10 shows the same information, now adjusted to take into account the volume of cases submitted by each officer. The Y axis is each officers IR rate: the total inappropriateness score divided by the number of cases they dealt with in total. 33 of the 41 officers have a score of 0.02 or lower. The mean is 0.04 (standard deviation: 0.14).

These 41 officers have a higher mean number of entries into the Gateway – 236 compared to the total populations of 168: 40% higher. The officers with high IR rates tend to have significantly lower numbers of overall entries. For example the officer to the far right entered only two cases in total, one with a mean IR score of 1.75. Only the two top scoring officers have an IR rate above 0.10. We can see that amongst the inappropriate rejecters, overall rejection levels are low and the only real reason for officers to stand out is more due to low levels of overall involvement, rather than high numbers of poor decisions. The overall rare nature of rejection decisions, combined with some officers having very limited involvement in the experiment, explains this pattern better than individual poor performance.

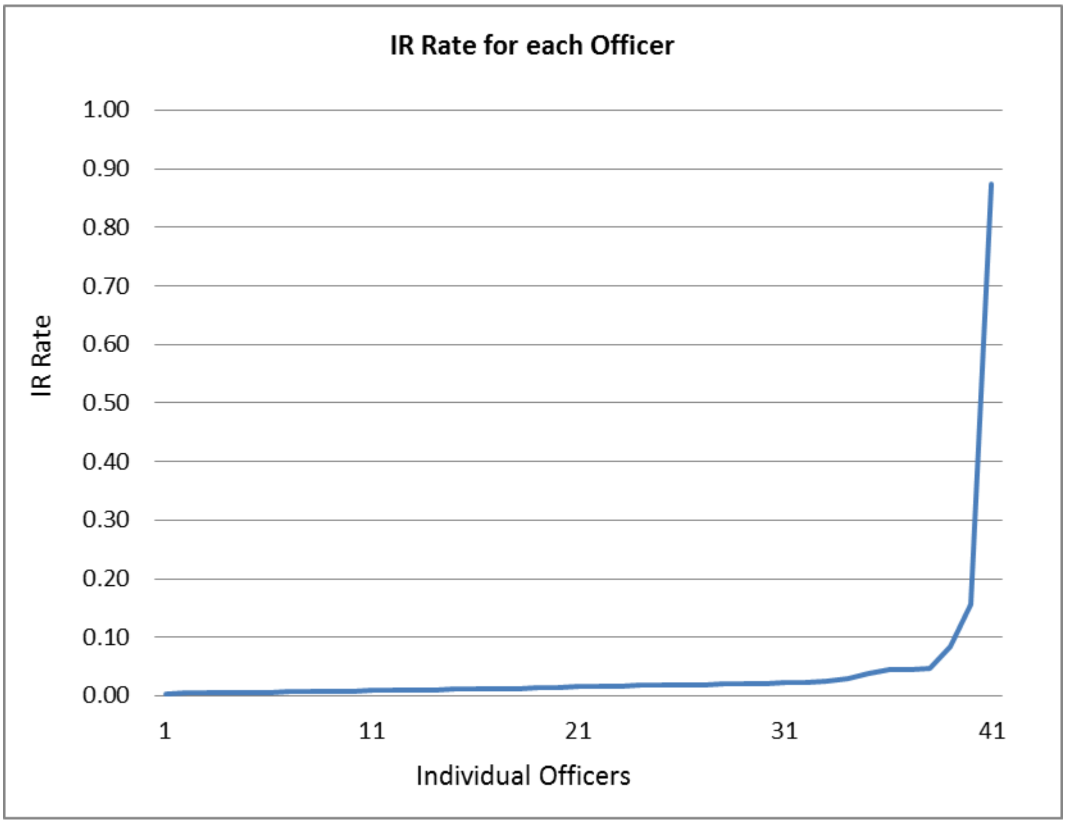


**Figure 10 Inappropriate Rejection (IR) Rate by Officer**

## Patterns of Appropriate Rejection

40 appropriate decisions were made by 27 officers: i.e. decisions on this 4.3% of all eligible cases were made by 28.1% of the officers involved in the project. The pattern of decision making is shown in Figure 11. Only seven officers made more than one appropriate rejection.

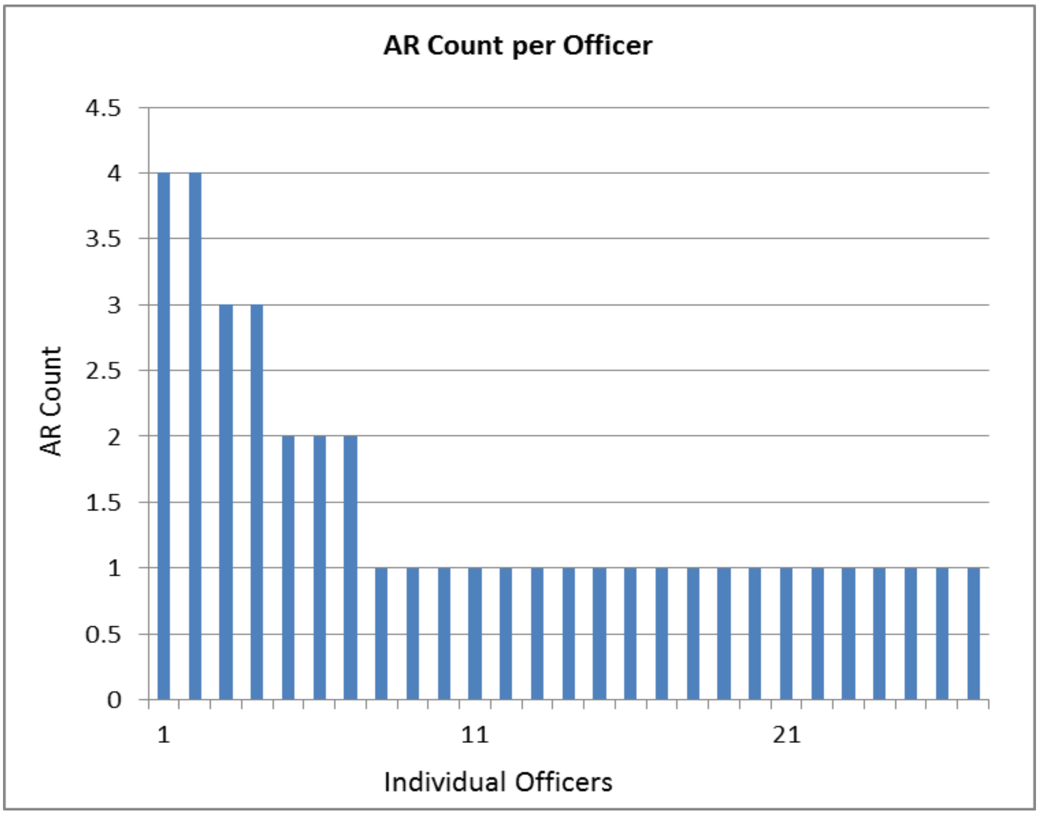


Figure 11 Count of Appropriate Rejections (AR) per Officer

As with the inappropriate rejections, this rate can also be shown in relation to the number of total entries each officer made into the Gateway: Figure 12. The fact that the score allocated to an appropriate decision was in the negative, and each decision could only score a maximum on -1, is reflected in the appearance of this chart.

It is of note that the two officers with the highest number of appropriate decisions (AR) also featured in the high IR cohort. The two officers that made four appropriate rejections made three and eight inappropriate decisions also. Four of the top 7 officers by volume of ARs also featured in the IR cohort. This would tend to suggest that we do not have separate populations of good or poor decision makers, *per se*.

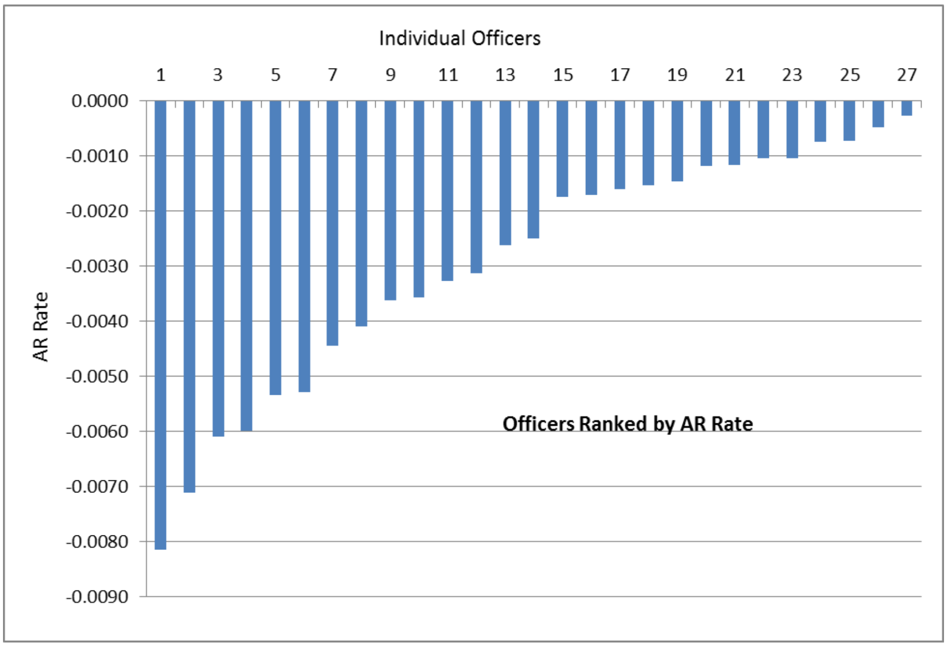
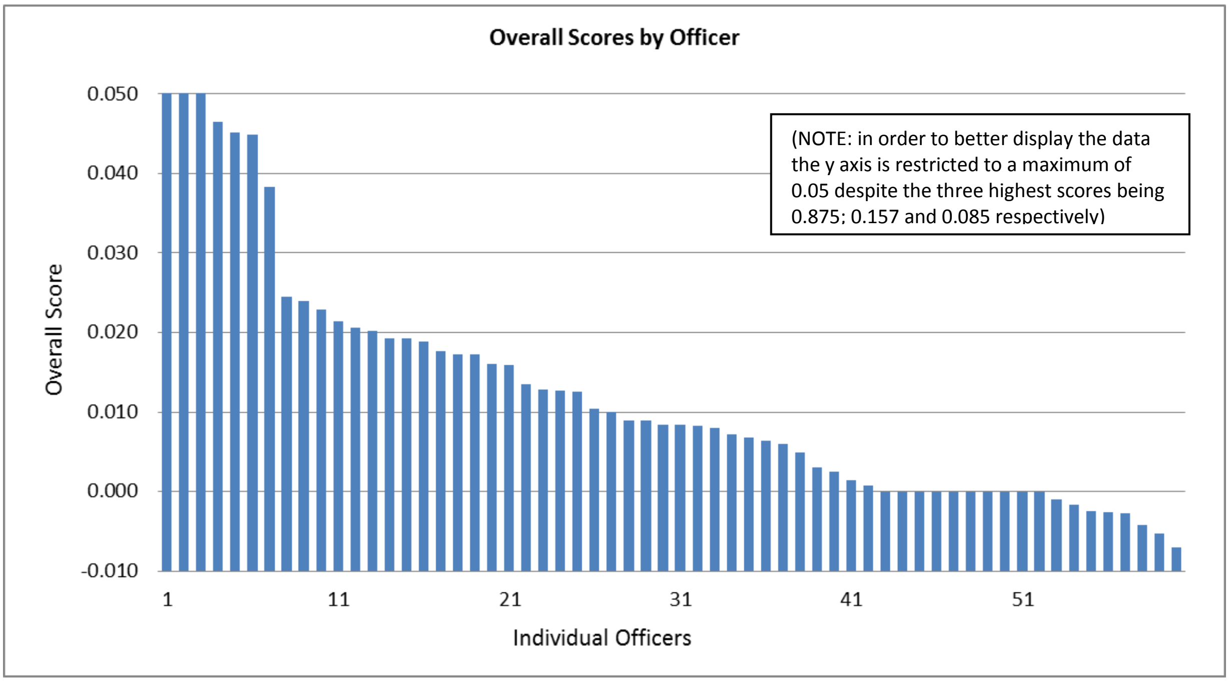


Figure 12 Appropriate Rejection (AR) Rate, by Officer

## Officers Overall Scores

Rather than viewing officers' performance through the 110 inappropriate decisions, or the 40 appropriate ones, it is possible to score the 60 officers that made rejections on the sum of both the scores, combining all of their rejection decisions. The overall appropriateness of all their rejections can then be calculated by dividing this score by the total number of entries they have made in the Gateway. The results of this analysis are shown in Figure 13. Note the zero point of the Y axis which denotes that overall the officers rejection scores are neutral (i.e. neither appropriate nor inappropriate). The eight officers to the right of the graph all have overall scores in the negative (i.e. appropriate outweighing inappropriate decisions).



(NOTE: in order to better display the data the y axis is restricted to a maximum of 0.05 despite the three highest scores being 0.875; 0.157 and 0.085 respectively)

Figure 13 Overall Scores for all Officers Making Rejections

Both Figures 10 and 13 identify the same seven officers at the extreme end of the inappropriateness scale. Hypothesising that these scores may relate to the level of training that these officers received they were contacted and completed a very short questionnaire. Two of the seven have since retired, leaving only five questionnaires completed, the results of which can be found in Appendix F.  In brief, this shows that all but one of the five officers attended the standard full days training in relation to the experiment: one even being involved in the extended training at Cambridge University at the commencement of the project. Also three of the five received an hour's personal update for the field researcher in 2013. Both of these sessions were the standard training that we attempted to get all officers to participate in. There was not significant extra training over and above this that anyone received. From this brief snapshot it would appear that the high levels of inappropriate decision-making were not due to a deficit in training or development inputs for those officers.

## Officers Making Both Appropriate and Inappropriate Rejections

Figure 14 brings the IR and AR rates for officers together into the same chart, thus assisting in the understanding of how much officers make both AR and IR decisions. The data clearly shows that these are not two distinctly separate groups, but that many of the officers making AR's also made IR's. However, there are a significant body of IR decision-makers that have not made AR's.

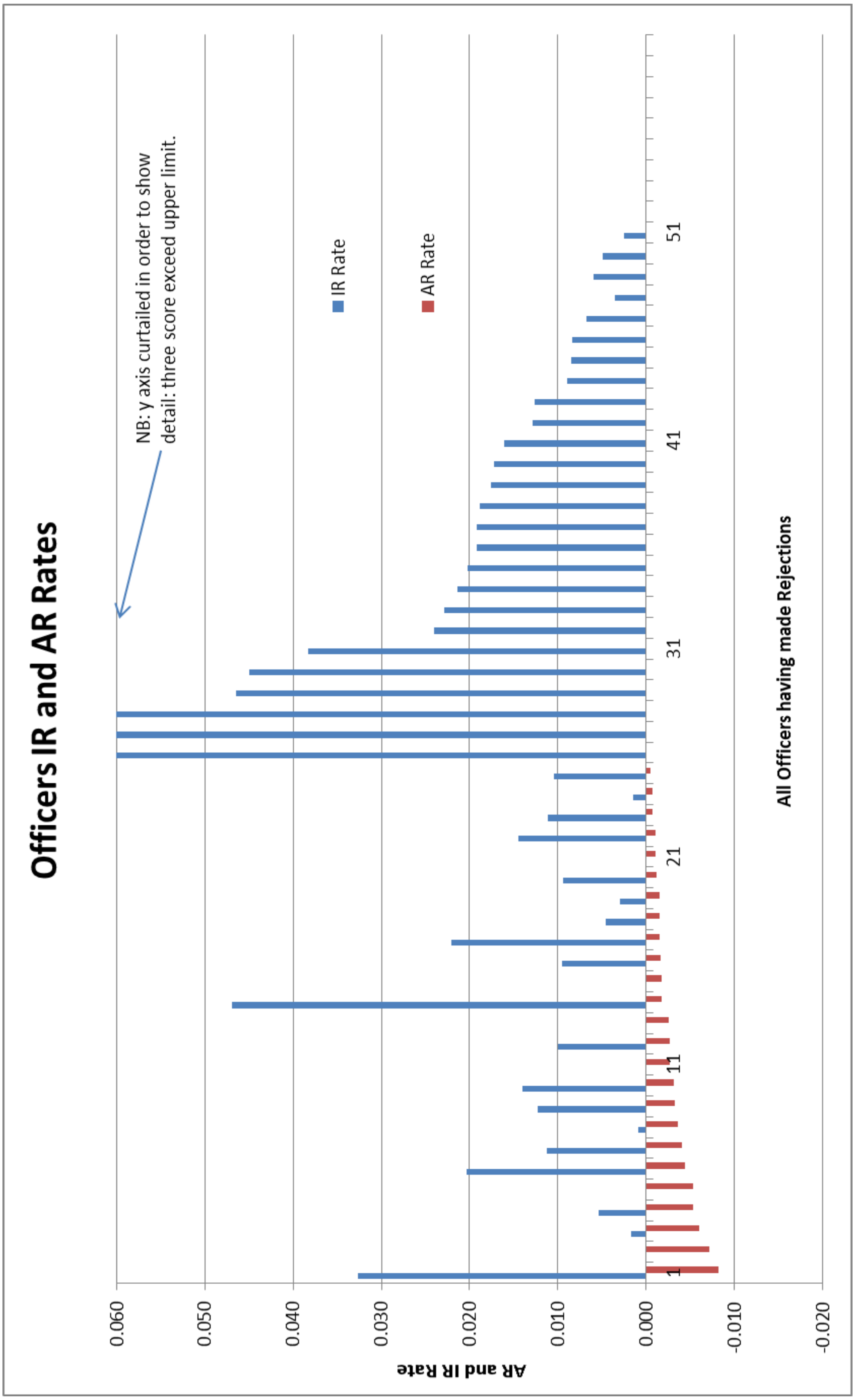


**Figure 14** IR and AR Rates for each of the Officers having made Rejections

As we have identified, there is some overlap between the officers that have made inappropriate rejections and those that have made appropriate ones: Figure 15.

62.9% of officers making an AR decision also made an IR decision.

42.5% of officers making an IR decision also made an AR decision.

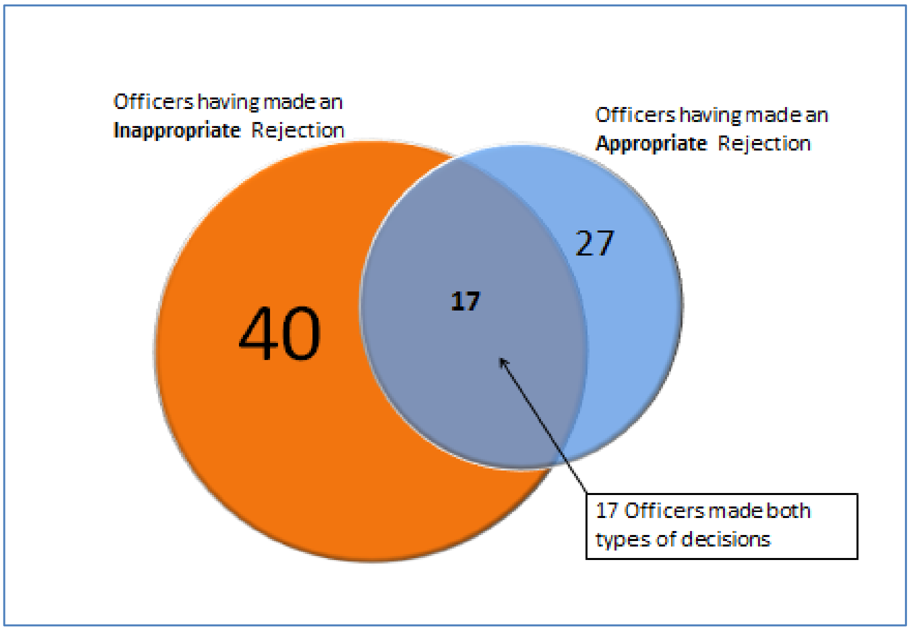

Officers having made an **Inappropriate** Rejection
Officers having made an **Appropriate** Rejection
40
17
27
17 Officers made both types of decisions

Figure 15 Overlap of Officers Making Both Types of Decisions

## Question Five: Are there any known characteristics of the officers that correlate with rates of appropriateness of rejection?

The data from the officer survey can be combined with the data on appropriateness of decisions from the Cambridge Gateway, allowing us to investigate any possible correlations that could help explain officers' rejection of randomisation. The following has been tested for any relationships:

- IR officers against the survey questions;
- Officers having worked on Turning-Point and those that haven't, against the survey questions.

### Survey Responses for Officers making IR Decisions

Of those 41 officers who made the IR decisions, 4 have since retired, leaving 37 asked to complete the survey: 31 of which have done so and identified themselves (an additional 6 responses to the survey have been obtained from officers involved in the project who have not identified themselves, so it is not possible to identify if they form part of the 41 Rejecters or not). We therefore have a survey response rate for the Inappropriate Rejecters of between 83.8% and 100.0%.

It was hypothesised that there may be relationships between officers' responses to 15 questions relating to their attitudes to offenders and prosecution/diversion; their support for experimentation and this experiment in particular; or some of their personal factors. The results of Mann-Whitney U and Chi Squared tests did not support these hypotheses in any area but educational level. There was a small relationship between a higher level of educational achievement and officers being in the group that had not made any inappropriate rejections: Table 9:

Table 9 Relationships between IR Decision-Makers and Survey Questions

| Survey Question | IR Median | non-IR Median | Mann Whitney U Test Significance |
|---|---|---|---|
| Q30 – Age | 4 | 4 | 0.343 |
| Q31 - Educational level | 5 | 6 | 0.033 |
| Q33 - Length of service | 4 | 4 | 0.412 |
| Q34 - Length of Custody service | 4 | 4 | 0.210 |
| Q3 - % of decisions made fast | 3 | 3 | 0.759 |
| Q7 - Reduce reoffending a priority | 2 | 2 | 0.963 |
| Q9 - Prosecution effectiveness | 3 | 3 | 0.678 |
| Q13 - % deserve rehabilitative alternative | 3 | 3 | 0.742 |
| Q20 - Understand project | 2 | 2 | 0.950 |
| Q21 - Supportive of project | 2 | 2 | 0.302 |
| Q23 - Supportive of experimentation | 2 | 2 | 0.795 |
| Q24 - Comfortable with randomisation | 2 | 2 | 0.874 |
|  | % | % | Chi Square |
| Q29 – % Male | 96.8 | 77.8 | 0.055 |
| Q2 – % Preference for fast decision-making | 67.7 | 62.2 | 0.621 |
| Q15 - % Bad decisions, not people | 71 | 75.6 | 0.655 |

## Survey Responses for Officers on the Experiment Compared to the Rest

Due to the nature of the experiment officers involved had training and developmental inputs relating to the effectiveness of different responses in dealing with offenders. It could be hypothesised that this training would have led to differing response from this group to certain questions on the survey, specifically those relating to their attitudes to punitive, rehabilitative and restorative options. The results of Mann-Whitney U and Chi Squared tests did not support these hypotheses: Table 10. We can take from this that being involved in Turning-Point has not significantly changed officer’s views of the value or importance of these options. The project team had seen changing officer views in these areas as a necessary step to increasing compliance. These results would not support that assumption.

Table 10 Relationship between Survey Responses and Involved in Experiment

| Survey Question | Worked on Turning-Point | Not worked on Turning-Point | Mann-Whitney U Test: Significance |
|---|---|---|---|
| Q30 – Age | 4 | 4 | 0.253 |
| Q31 - Education | 5 | 5 | 0.356 |
| Q33 - Length of service | 4 | 4 | 0.160 |
| Q34 - Length of custody service | 4 | 4 | 0.712 |
| Q3 - % Decisions made fast | 3 | 3 | 0.600 |
| Q7 - Reduce reoffending a priority | 2 | 2 | 0.559 |
| Q9 - Prosecution effectiveness | 3 | 3 | 0.434 |
| Q13 - % deserve rehab alternative | 3 | 3 | 0.783 |
| Q4a - Amount of discretion | 1 | 1 | 0.255 |
| Q5 - Clear on the 'right thing' | 2 | 2 | 0.993 |
| Q6a - Importance of compensation | 2 | 2 | 0.106 |
| Q6b - Importance of punishment | 2 | 2 | 0.124 |
| Q6c - Importance of RJ | 2 | 2 | 0.794 |
| Q6d - Importance of rehabilitation | 2 | 2 | 0.258 |
| Q6e Importance of Education | 2 | 2 | 0.258 |

## Discussion and Implications

We now have a considerable body of data relevant to our question: can using an algorithmic triage model to support eligibility decisions in an RCT lead to high levels of consistency and validity in police officer decision-making, especially for more subjective selection criteria? The discussion begins with considering the ratio of appropriate to inappropriate rejections of randomisation and considering if this was acceptable. We then consider at the consistency, the 'measurement reproducibility' (Gwet 2012) between our four test assessors and what that means for consistency in operational officer decision-making. The themes of inappropriate rejection and the impact these may have had on external validity are considered. We ask if this would have been different if researchers had taken on this role, and consider if so, could they have made the same appropriate rejections and protected the experiment to the same degree as the officers?

### Differentiating between Appropriate and Inappropriate Rejections

We have seen that the ratio of appropriate to inappropriate decisions is 1:2.75. It is an opportune and sensible question to ask if this ratio is acceptable – is the benefit of having more subjective selection criteria worth the loss of nearly three times as many eligible cases as those excluded appropriately?

The 40 cases appropriately rejected are those that the public are most likely to baulk at being diverted from court. The analysis shows that this is not so rare an event – 40 incidences out of 734 eligible cases (5.4%). But of course it is not the number of cases that is important – but the potential damage that could be done to the forces reputation if these cases were diverted. Whilst it is hard to second guess the press and public's reaction if any of the cases appropriately rejected had been included, these are the type of cases that would have been more difficult to justify. An adverse reaction from the public against just one case may well have put the whole experiment in jeopardy. This is a perfect example of the importance of taking ‘influence factors’ into account (Fixsen et al. 2005). In this sense the use of officer discretion worked well to exclude ostensibly eligible cases that could have caused reputational difficulties if included. It is also possible that this same mechanism worked internally to a degree: the inclusion of these more challenging cases could have lost the project significant internal support. This justifies the use of a subjective ‘catch all’ opportunity for officers to exclude ostensibly eligible cases.

The second point to pull out of the results on individual decisions is the variability in assessing ‘appropriateness’ of the rejections amongst the four assessors. Kappa scores show a moderate correlation between most assessors, with the two most closely aligned to the project showing a substantial correlation. This evidences the difficulty in achieving consistency in one element of the selection criteria – the ‘any other reason to exclude’ question. This is the most subjective part of the decision making process, and only comes into operation in a small proportion of total cases. With these moderate levels of consistency in core project staff it is reasonable to assume that consistency would have been lower across the total population of officers involved.

However this is just one small, albeit important, element of the selection process. The algorithmic triage model, operationalised through the Cambridge Gateway, supported 14,681 decisions, of which only 244 involved the rejection of random allocation for ostensibly eligible cases. The advantage of the model is that it facilitated the capture of data to evidence the scale of this problem, and captured officers’ decisions which allowed subsequent feedback and analysis. Thus the model facilitated a much deeper understanding of the decision-making and was an essential precursor to the process of individual feedback and coaching that underpinned the steady improvements in performance. As Figure 7 illustrates, this supported the improvement from an IR Rate of over 40% at the start of the project, to one of less than 10% by the end. The fact that this is an evolutionary process which takes time is relevant to future experiments which should plan to run in a test and developmental mode for a period of time to allow the coaching and improvements to take place.

## Recurring Themes of Rejection

The themes within the rationales of appropriate rejections fall into two broad categories – those that had practical reasons for being unsuitable (e.g. deportation in progress); and those where the risk to reputation was considered too great.  These evidence the appropriate use of discretion and make an argument that such flexibility is necessary.

The themes of rationales for inappropriate rejections may give some clues towards the motivation behind them:

Firstly there were a proportion that cite reasons that are not only unsuitable specifically for Turning-Point, but also more generally for all prosecution decisions, e.g. taking intelligence or suspected activity into account. There are also examples of the offenders' attitude to the police or the investigation swaying the decision towards a prosecution (presumably as it is seen as being 'harsher').  This could lead to interpretations of a 'cynical' outlook (Muir, W, 1977); or as giving credence to the 'police interest' interpretation (Sanders, et al., 2010), although the overall numbers are such that it appears to be a problem on a minor rather than major scale.

Secondly there are a number of cases that show the continued application of the traditional diversion criteria to Turning-Point, where it is not applicable, e.g. a lack of an admission or the presence of multiple offences. This evidences the difficulty of running two very similar interventions alongside each other – the easy confusion of separate, but similar criteria. This may also be evidence of one of the issues identified in implementing improvements in medicine: the compatibility of the of the recommendation with existing practioner values (Grol & Grimshaw 2003). The standard service and the set of values that entails appear resistant to change in our experiment.

Thirdly, there are cases, mainly those relating to breach of trust and vulnerable victims, which possibly reflect a moral judgment. In 1993 research showed 'has strong moral values' came low down the list of attributes custody staff felt they needed for their job (Waddon & Baker 1993); however this does not negate the potential for moral values to be playing a part in decision-making here. Early scholars in police culture would argue that morality plays a significant role in officers' actions (Muir 1977; Skolnick 1966; Wilson 1968). They may also be examples of Lipskys' (1980) 'worker bias': cases which invoke hostility and reflect a general assessment of the offenders 'worthiness' to receive a diversionary intervention. Lipsky also describes a process he calls 'creaming', where cases most likely to succeed are prioritised (Lipsky, 1980). What we see here could be interpreted as creaming, with cases most deserving prioritised. We could also interpret these findings in the framework Skinns proposes, seeing the impact of an adversarial system on the officers' decision-making, or the tension between due process and crime control (Skinns 2011).

Whilst the numbers are low, and internal validity is still high as the cases are excluded before random assignment, this may still affect the external validity of the experiment. To some degree random assignment of the full eligible population has not occurred: a small group who were deemed by the selectors to be undeserving of the opportunity to participate were excluded.

Sherman (2009) argues that researchers are the best agents of randomisation due to their independence and rigorous focus on experimental fidelity. The reasons above evidence the reasons for his concerns in using practioners to make these key decisions. One may look at the reasons for the inappropriate rejections and conclude that an *independent* researcher would have been less likely to have come to the same conclusion.

We cannot say whether a researcher could have used this discretion appropriately to keep out the appropriately rejected cases and protect the force and project to the same degree that officers did. They may have found it difficult to elicit the detail and depth of information on each case that would have been necessary to identify such cases. Therefore we may conclude that the use of officers to make these selection decisions reduced the validity of the experiment, but was necessary to protect both its existence and the reputation of the force.

So, whilst the structured, guided nature of the Cambridge Gateway was able to assist with decision-making on all of the objective selection criteria, it was less able to guide the decision maker in making this final subjective assessment. However, what it was able to do was to ensure that officers recorded their rationale for this decision, which opened it up to retrospective scrutiny. During the project this scrutiny allowed project leaders to assess these decisions in real time and give feedback, both on individual decisions, but also to all staff on recurrent themes or issues. The impact of this is seen in the following section, where we will discuss the improvement in rejection rates during the experiments life-span.

## The Improving Rejection Rate

High level results for the experiment show that over the 32 months for which we have data there was an average inappropriate rejection rate, pre-randomisation, of 12%. We have no benchmark to compare this to as the literature on experiments in policing to date fails to describe this particular aspect in detail: a wider reporting of rejection levels would be advantageous to allow comparison. Given the nature of the study, the diffuse nature of the field station, the numbers of staff involved and the requirement to include a subjective selection criteria, 12% inappropriate rejection rate could be viewed as acceptable. When taken in the context of 14,681 individual decisions by officers this appears even more favourably.

Of greater interest is the drop in the overall rejection rate and inappropriate rejection rate over the life of the experiment (Figure 7). This shows a steady reduction in the error rate, reflecting an improvement in the quality of decision-making:

- Phase One IR Rate: 28.8%
- Phase Two IR Rate: 21.7%
- Phase Three IR Rate: 8.9%

When we map this trend against key events in the life of the project some possible explanations become evident: Figure 16. These explanations relate to the development of the staff involved in making these decisions; and the decision-making tool, the 'Cambridge Gateway', which shall be discussed next.

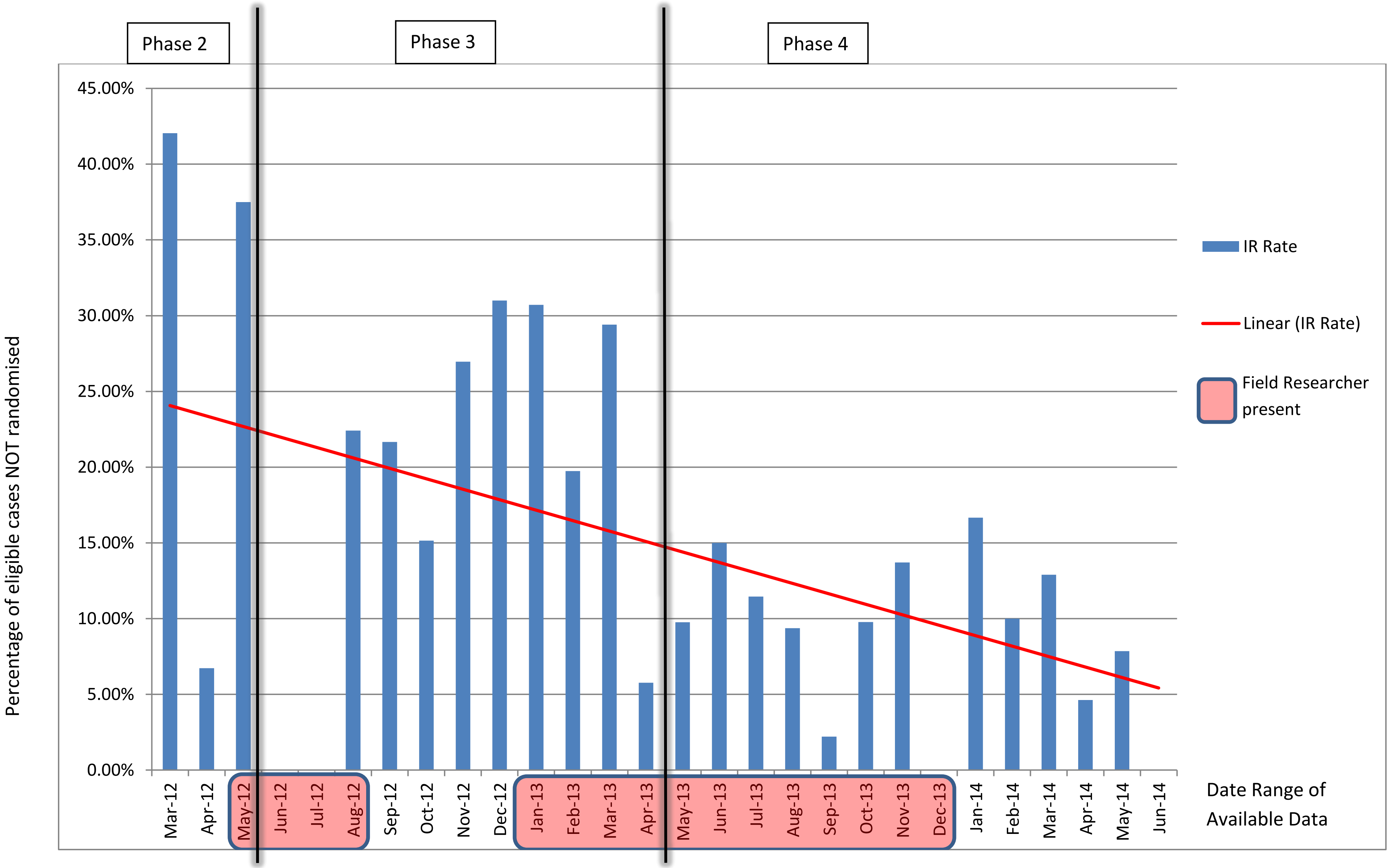


**Figure 16 Experimental Timeline Showing Key Events and Phases**

## Developing the Individual

It was recognised by the research team that the use of a devolved and distributed model for determining case eligibility would be challenging, and could pose a threat to external validity. Therefore the training and development of staff skills in this area was crucial. The field researcher was only present for 16 of the 37 months the experiment was actively selecting cases for (dates highlighted in pink on Figure 16).

In order to develop skills the leadership team opted for a mixed methods approach. Officer time available for training was limited. Initially this involved a two-day retreat to Cambridge University, as recommended by Sherman 2009, but later this had to be reduced to six hours in-force for practical reasons. Guidance notes for each role, process charts and explanations of the theory were prepared and made available via the force intranet, to give a definitive reference point at all times of the day or night.

This was then supported by a regime of audit and feedback/coaching. For custody staff this was in three key domains. The first two of these were proactive audits: checking of custody records to ensure cases weren't being omitted from the Gateway altogether, and dip sampling of cases to test the integrity of the assessment of likelihood of a custodial sentence: both of which showed very high compliance and accuracy rates. Neither of these two systems operated for the entire life of the project: both were much more heavily utilised in the earlier stages, as once standards were found to be high there was not sufficient resources to continue frequent audits. Thirdly, project staff also monitored and responded to those rejections that were recorded, particularly those that appeared suspect. This information was used by the project team to give feedback to operational staff, very much in a supportive and developmental tone. In addition the project team made themselves available to contact to discuss cases or issues, and were also the resolving body for disputes over eligibility.

Monitoring of a number of measures during the first 12-16 months of the projects life, and the arrival of the field researcher in 2013, led to a big push to speak with all custody staff in person during March –July 2013. This consisted of approximately an hour of face to face contact (usually one to one, occasionally in very small groups of 2-3 officers) and was in effect a coaching session. This aimed to hit four objectives: to brief staff on changes to the Gateway tool that came into effect at the start of Phase Four; to give general feedback on common issues and errors that were occurring across the board; to give specific feedback to officers on their performance and to answer any specific concerns they had; and to show organisational interest and attention to their work to remind them it was still being closely observed by the project team and senior managers. This coaching was conducted almost exclusively by the field researcher due to lack of capacity from other (police) team members.  It is felt that this wave of coaching and organisational attention contributed significantly towards the step change in IR Rate between Phases Three and Four shown in Figure 16, and therefore was an effective practice.

Finally attempts were made to share general lessons learnt via global E-mails, and to highlight success stories and ongoing issues via a newsletter. The overall strategy was attempting to maximise the influence of the small, mainly part-time project team across the wide audience involved, using a variety of different methods.

The experiments leadership was acutely aware of the demands on officer time and so attempted to minimise demands on their time and attention. As the only dedicated worker on the project much of this responsibility for personal feedback and coaching fell on the field researchers' shoulders. This approach had much in common with the developmental elements of Fixsens' core implementation components described earlier: Figure 17.

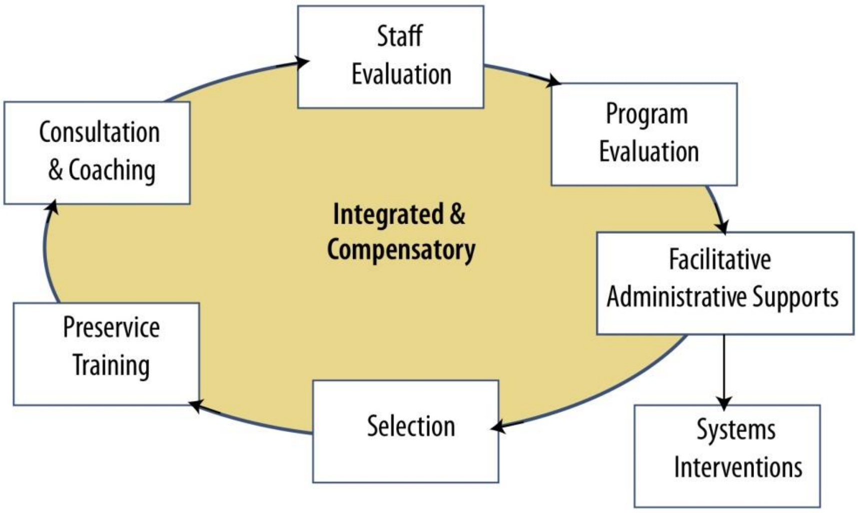
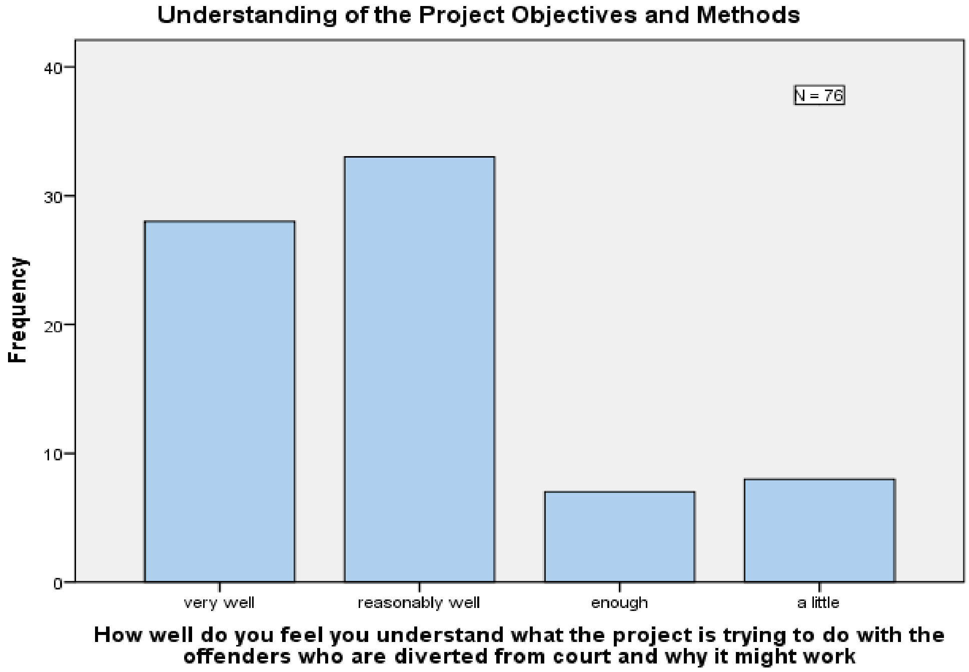


Figure 17 Core Implementation Components (Fixsen et al, 2005)

It is suggested that this multi-level approach that included Fixsens core implementation components contributed towards the incremental improvement in decison-making performance over the project life-span. Particularly, the presence in-field of the lead researcher, and her activity in coaching individual officers is seen as a very influential factor in this improvement, supplementing the building blocks of training and audit already delivered.

It is worth noting that unlike many previous experiments, Turning-Point did not select staff to take on the custody officer (case selection) role in any way whatsoever. This role fell on whichever staff happened to be working in the project area at the time, and suffered from the same rotation of staff, performance and motivation issues that any workforce would normally experience. Bearing in mind the potential benefit staff selection could have had, the performance achieved appears all the more remarkable. Further evidence that the approach taken to train and develop decision-making skills in staff was broadly effective comes from responses to a question regarding officers understanding of the project in the survey, Figure 18:



## Figure 18 Survey Responses – Understanding the Project

80% of the 76 respondents involved in Turning-point state they understand the project reasonably or very well; over 90% felt they understood it ‘enough’ or better. These results would tend to support the finding that the training and development regime was effective.

We would therefore suggest that the multi-level and mixed method approach is an effective one for a large scale experiment involving devolved decision-making under a algorithmic triage model by operational officers, such as in Turning-Point. Now that we know and understand much more about the nature of the barriers and facilitators of improved practice in this scenario, future researchers will be better placed to tailor their implementation regime, as recommended by Grol and Grimshaw, 2003.

## Developing the 'Cambridge Gateway'

The principal task of the Gateway was the ability for operational staff to enter in key case details and for a computer to randomly allocate it to one treatment or the other. The case details and the allocation would be automatically recorded, allowing research staff to monitor fidelity and track cases through the pipeline. Thus the independence of random allocation was preserved, without the problematic need for academics to be involved on a 24/7 basis (Ariel et al. 2012).

Once the platform was in existence it also became obvious that its support could be extended beyond simple random assignment, to include helping staff work through the eligibility criteria for the experiment. It evolved further to include capability, post randomisation, to inform the necessary teams of the impending case enroute to their team, via automated E-mail. This evolutionary approach to project implementation fits with the recommendations by Pressman and Wildavsky (1973).

The Gateway tool had two significant upgrades during the life of the experiment. Initially this was at the end of the testing and piloting phase, Phase Two. During this period feedback had been elicited from staff and improvements were made to the eligibility criteria. A great deal of emphasis had been placed on involving the operational staff in the construction of the experiment during this phase, and this upgrade reflected improvements borne out of their experiences. For example, the exclusions of drink-driving and any other offences that required a driving ban were added in, with the rationale that we could not provide any similar intervention outside of the prosecution process. One selection question which was worded as a double negative was reworded for clarity. Additionally the layout and appearance of the tool was improved to make it more user-friendly.

The second upgrade was at the start of Phase Four, when improvements meant that officers could not re-randomise cases (a small but persistent problem in earlier phases, leading to' treatment as assigned' issues); that the common reasons for ineligibility led to exclusion more swiftly; and officer's down-stream of random assignment were automatically notified of the work coming their way. This upgrade significantly improved the experience of using the tool for operational staff, especially for the vast majority of cases which were excluded on the grounds of previous criminal history or likelihood of a custodial sentence. This reduced the amount of time and work required by operational staff as well as automating some of the workflow, thus reducing errors or omissions.

It is hypothesised that both of these upgrades to the Gateway tool contributed to the improvements in performance shown in Figure 16, with the second one being more impactive. Whilst these upgrades didn't directly affect the one subjective decision to exclude an otherwise ostensibly eligible case, they did evidence the organisation listening to staff, taking their concerns seriously and making improvements to their advantage. Interpreted in the light of organisational justice theory (Bradford et al. 2013) this approach has increased staff engagement and support for the project, leading to better performance. It is also possible that the improvements to certain elements of the Gateway had beneficial effects on decision-making more broadly.

The implications of these findings are that the tool itself is a very beneficial and important link in the decision-making process, that it can be adapted to raise performance, and that it can complement a transformational and inclusive leadership approach to experimentation.

## Patterns of Rejection amongst Officers and Possible Causes

We have seen that 42.7% of the decision-makers made one or more inappropriate rejections, and that many of these also made appropriate rejections. There is a gradual spread of IR scores between those officers with the highest and lowest. It seems probable that the issue for inappropriate decision-making is more likely to be one pertaining to the decision than the individual. The few officers that stand out with high IR rates tend to have low rates of decision-making overall, but appear to have received much the same formal training as their better performing counterparts. Officers with high *counts* of inappropriate decisions correspond with those making the highest volume of decisions overall, so their actual *rates* of inappropriate decision-making tend to be low.

We could take from this that familiarity with decision-making tends to relate to better (lower) rates of inappropriate decisions. This again supports other findings, that high skill levels are encouraged by a combination of training, practice, feedback and support (Fixsen et al. 2005). In turn this then tends to suggest that the coaching and feedback elements of the development regime used were the most effective, as officers with low levels of Gateway use would have received less of this, as both will have been directly related to time spent on the experiment.

The lack of any correlation between officer's views of rehabilitation and diversion, and high IR rates suggests that this was not playing out particularly in the decision-making. It could also indicate that a values-based approach to training staff, one that emphasised the benefits of rehabilitation and diversion, may be less important than we originally envisaged. This would suggest a more technical approach, i.e. one which focuses on experimental validity and fidelity, rather than cultural values, will be superior in reducing inappropriate rejection of randomisation: one focussed on experimental compliance over developing a greater value on rehabilitative or restorative approaches.

We must consider the *lack* of correlations between factors highlighted in our findings, too. Whilst the lack of correlation between performance and personal factors such as gender, age, length of service is not a surprise, other results are more noteworthy. Any hypothesis that there would be a relationship between officers with Kahnemans' intuitive or deliberate decision-making styles and performance has not been supported by this analysis (Kahneman, D, 2011). Likewise with officers having a more tragic or cynical outlook (Muir, W, 1977). We could consider that negating the influence of these factors on our decision-making makes the argument that the alternatives we have proposed are all the stronger through lack of competing explanations.

That the one significant correlation we found was with educational accomplishment of officers is interesting. Figure 15 shows the survey responses on this issue. This is a weak relationship, but is present none the less. It is possible that this could support our emerging conclusion that the decison-making around whether to exclude an ostensibly eligible case was complex, such that more highly educated officers were better able to cope. Has the structure and discipline to deal with complexity, to think systematically and critically, come with higher levels of education? Are more educated officers able to better reflect on their decisions and identify their own use of inappropriate factors, or better assimilate fedback that points this out? If so then this is likely to have far reaching consequenses, as policing in general becomes increasingly complex and the need for thoughful and considered use of discretion increases. Such findings would support the many comentators that argue the need to raise educational levels in the service, and the development of the new College of Policing which has education of officers as one of its primary objectives (College of Policing, 2014).

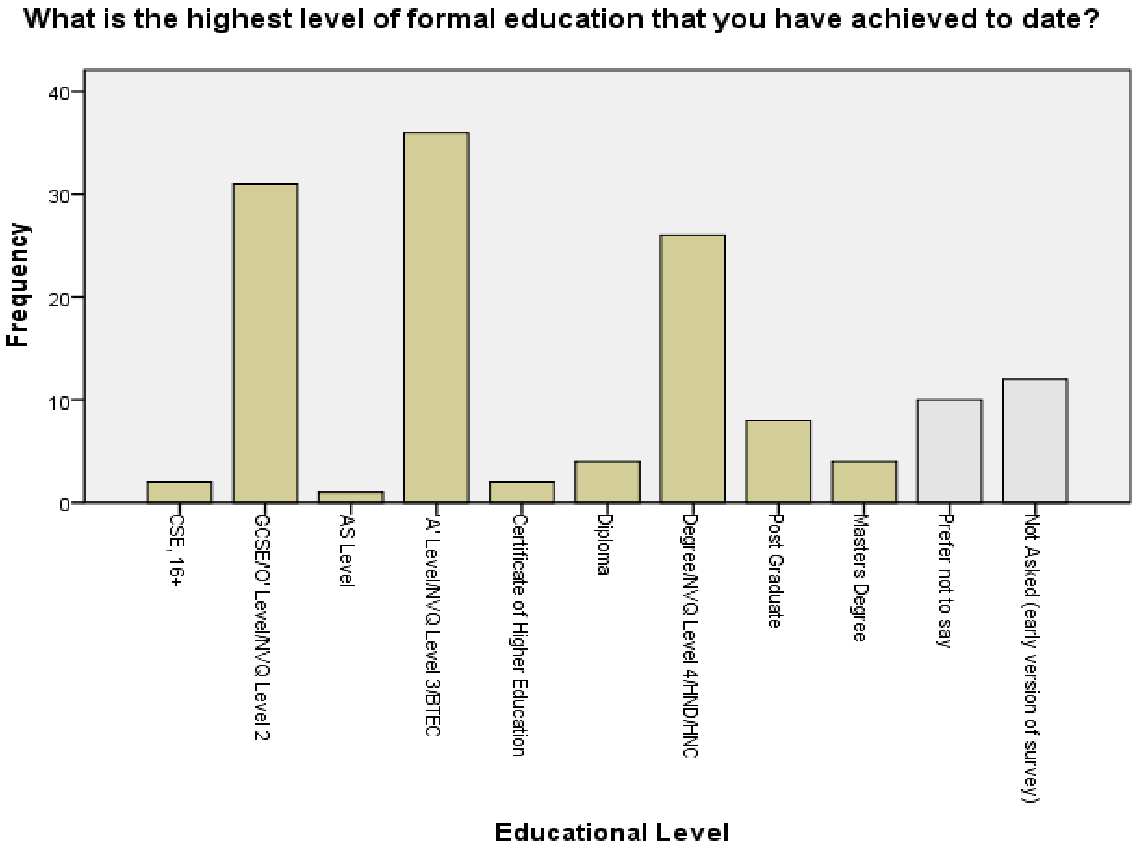


**Figure 19 Officers Educational Levels – Survey Responses**

We have concluded that the analysis tends to suggest there isn’t a population of ‘rejecters’ *per se*, which  leads to a conclusion that it may be the nature of the decision itself, rather than the decision-maker, that fosters inappropriate decision-making. The decision required an assessment of all factors to identify if inclusion of the case was just too risky for the force. This was a subjective and complex decision: even the four assessors completing the inter-rater reliability assessments only had a moderate level of agreement. Also, officers making inappropriate decisions also often made appropriate ones. Therefore it should come as no surprise that we found little difference in the survey responses of those making inappropriate decisions, and the rest.

## Conclusion and Recommendations

We now return to the central question of this study: to determine if using an algorithmic triage model to support eligibility decisions in a Randomised Control Trial (RCT) can lead to high levels of consistency and validity in police officer decision-making, especially in relation to a subjective decision.

We have shown that custody officers, working in busy and pressurised custody suites in a demanding urban policing environment, can make selection decisions for an RCT, with high levels of consistency and validity. High performance is supported by a bespoke selection and random allocation tool (the Cambridge Gateway in this instance); and by a multi-level and mixed method approach to implementation.

The facility given to custody officers to reject ostensibly eligible cases was used appropriately and fulfilled its function of protecting the force and the continuation of the experiment itself. However, this discretion was used inappropriately at times, but this misuse declined significantly as individual performance was monitored and staff coached, and as the Gateway tool was improved to both simplify and streamline the process.

We would recommend the use of an algorithmic triage model, supported by suitable IT such as the Cambridge Gateway, by police officers in future RCTs to fulfil the case selection and randomisation roles. Adaptation of the tool to the specific needs of the experiment will increase consistency and fidelity of the decision-making, thus ensuring the experiment has high levels of external validity. In addition to considering the use of this model, experimenters must keep in mind the complexity and subjectivity of the decisions they are expecting staff to make, and keep these to a minimum.

We would also recommend that any force looking to implement such an algorithmic triage model adopt a multi-level and mixed method approach to implementation. Such an approach would include initial training, on-line advice, audit and feedback and individual coaching in the field. Our advice would be that the audit, feedback and coaching is essential to ensure that practice changes in the desired manner, and that the use of a dedicated in-field researcher is invaluable in delivering this.

To any readers that are considering replicating the Turning-Point experiment we are able to offer specific advice, learning from our experience in Birmingham. Key recommendations have been synthesised into a one page checklist for future researchers and lead officers, contained in Appendix G.

Whilst RCTs have traditionally been seen as the 'gold standard' of evaluations, they have also been viewed with caution and trepidation; difficult to implement well and requiring considerable resources. By adopting the algorithmic triage model and utilising police officers in roles previously retained for researchers RCTs are now much more achievable. It is hoped this will accelerate the use of RCTs in policing to better understand how to reduce crime and harm in our communities. Through increased experimentation the service will become more open to scientific enquiry and more receptive to relevant research.

It is hoped that this study will encourage the use of an algorithmic triage approach to experimentation in policing and that this will help officers across the world provide a better service to the communities we serve.

# Bibliography

Abramowicz, M., Ayres, I. & Listokin, Y. (2011) Randomising Law, *University of Pennsylvania Law Review*, 1 – 64.

Ariel, B., Vila, J. & Sherman, L. (2012) Random Assignment Without Tears: How to Stop Worrying and Love the Cambridge Randomizer, *Journal of Experimental Criminology*, 8(2), 193–208.

Birmingham City Council, (2014) *Population and Census.* Retrieved 12th September 2014 from: http://www.birmingham.gov.uk/cs/Satellite?c=Page&childpagename=Planning-and-Regeneration%2FPageLayout&cid=1223096353755&pagename=BCC%2FCommon%2FWrapper%2FWrapper

Boruch, R., Snyder, B. & DeMoya, D. (2000) The Importance of Randomized Field Trials. *Crime & Delinquency*, 46(2), 156–180.

Bradford, B. Quinton, P. Myhill A. & Porter, G. (2013) Why do "The Law" Comply? Procedural Justice, Group Identification and Officer Motivation in Police Organizations, *European Journal of Criminology*, 11(1), 110–131.

Braga A. Welsh, B. Papachristos, A. Schnell, C & Grossman L. (2014) The Growth of Randomized Experiments in Policing: the Vital Few and the Salience of Mentoring, *Journal of Experimental Criminology*, 10, 1 – 28.

Bronitt, S. & Stenning, P. (2011) Understanding Discretion in Modern Policing, *Criminal Law Journal*, 35, 319 – 332.

Brown, D. (1997) *PACE Ten Years On - A Review of the Research (Home Office Research Study 155),* London: Home Office Research and Statistics Directorate.

College of Policing, (2014) *'What We Do'.* Retrieved 19 September 2014 from http://www.college.police.uk/en/19789.htm]

Cordray, D.S. (2000) Enhancing the Scope of Experimental Inquiry in Intervention Studies. *Crime & Delinquency*, 46(3), 401–424.

Creswell, J. W. (2007) *Qualitative Inquiry and Research Design - Choosing Among Five Approaches,*(Second Edition), Thousand Oaks, CA: Sage

Criminal Justice Joint Inspection (2009) *Exercising Discretion: The Gateway to Justice*, London, HMSO.

Dunford, F. Huizinga, D. & Elliot, D. (1990) The Role of Arrest in Domestic Assault: The Omaha Police Experiment, *Criminology*, 28, 183-206.

Fixsen, D. Naoom, S. Blasé, K. Friedman, R. & Burns, B. (2005) *Implementation Research: A Synthesis of the Literature*, Tampa: Florida.

Foddy, W. (1993) *Constructing Questions for Interviews and Questionnaires - Theory and Practice,* Hong Kong: Cambridge University Press.

Greene, J.R. (2014) New Directions in Policing: Balancing Prediction and Meaning in Police Research, *Justice Quarterly*, 31(2), 193 – 228.

Grol, R. & Grimshaw, J. (2003) Research into Practice: From Best Evidence to Best Practice : Effective Implementation of Change in Patients' Care, *The Lancet*, 362, 1225–1230.

Gwet, K.L. (2012) *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters,* Gaithersburg, MD: Advanced Analytics, LLC.

Hawken, A. & Kleiman, M. (2009) Managing Drug Involved Probationers with Swift and Certain Sanctions: Evaluating Hawaii's HOPE, Washington, DC: National Institute of Justice, Office of Justice Programs.

Kahnemen, D. (2011) *Thinking Fast and Slow*, London: Allen Lane.

Kelling, G. (1999) *Broken Windows and Police Discretion*, Washingon: National Institute of Justice.

Kilburn, M. R. (2012) Lessons from the Randomized Trial Evaluation of a New Parent Program: When the Evaluators See the Glass as Half Full, and the Community Sees the Glass as Half Empty, *Journal of Experimental Criminology*, 8(3), 255–270.

Killias, M. Aebi, M. F. & Ribeaud, D. (2000) Learning Through Controlled Experiments: Community Service and Heroin Prescription in Switzerland. *Crime & Delinquency*, *46*(2), 233–251.

Laub, J. H., & Sampson, R. J. (1993) Turning Points in the Life Course: Why Cange Matters to the Study of Crime, *Criminology*, 31(3), 301–325.

Lipsky, M. (1980) *Street-Level Bureaucracy: Dilemma of the Individual in Public Service,.* New York: Russell Sage Foundation.

Lunn, C. et al. (2012) Receptivity to Research in Policing. *Justice Research and Policy*, 14(1), 61 – 95.

Muir, W.K. (1977) *Police - Streetcorner Politicians*, Chicago: University of Chicago Press.

Neyroud, P. W. (2011) Crim-PORT 1.0: Operation Turning Point: an Experiment in “Offender Desistance Policing” West Midlands Police and Cambridge University. Retrieved 12 November 2014 from:

http://www.crim.cam.ac.uk/research/experiments/rex-post/operation_turning_point.pdf

Neyroud, P. W. (2013) 'An Interim Report into the Turning-Point Experiment' (unpublished report to West Midlands Police, July 2013)

Oakley, A. (2000).A Historical Perspective on the Use of Randomized Trials in Social Science Settings, *Crime & Delinquency*, 46(3), 315–329.

Petrosino, A. Turpin-Petrosino, C. & Guckenburg, S. (2010) *Formal System Processing of Juveniles: Effects on Delinquency*, Campbell Systematic Reviews.

Pressman, J. L. & Wildavsky A. B. (1973) *How Great Expectations in Washington are Dashed in Oakland*, Berkeley, CA, University of California Press.

Robson, C. (2002) Real World Enquiry; Approaches to Social Research. In *Real World Research*. Blackwell Publishing, 3-18.

Roman, C. G., Fontaine, J., Fallon, J., Anderson, J., & Rearer, C. (2012) Defending and Managing the Pipeline: Lessons for Running a Randomized Experiment in a Correctional Institution, *Journal of Experimental Criminology*, 8(3), 307–329.

Rose, G. & Hamilton, R., 1970. Effects of a Juvenile Liaison Scheme, *Brit. J. Criminology*, 10(1), 2 – 20.

Ruane, J. (2005) *Essentials of Research Methods*, Oxford, Blackwell Publishing.

Sanders, A., Young, R. & Burton, M. (2010) *Criminal Justice*(Fourth Edition), Oxford: Oxford University Press.

Sherman, L. (1998) *Evidence-Based Policing*, Police Foundation: Washington.

Sherman, L. Schmidt J. Rogan, D. Smith, D. Gartin P. Collins E. & Bacich A. (1992) The Variable Effects of Arrest on Criminal Careers: The Milwaukee Domestic Violence Experiment, *The Journal of Criminal Law and Criminology*, 83(1), 137.

Sherman, L.W. (2007) The Power Few: Experimental Criminology and the Reduction of Harm, *Journal of Experimental Criminology*, 3(4), 299–321.

Sherman, L & Strang, H. (2007) *Restorative Justice: The Evidence*, London: The Smith Institute.

Sherman, L. (2009) An Introduction to Experimental Criminology. In Piquero, A. R. & Weisburd, D. (eds) *Handbook of Quantitative Criminology,* 399 - 436, New York: Springer.

Sherman, L. W. & Neyroud, P. W. (2012) *Offender-Desistance Policing and the Sword of Damocles,* London: Civitas.

Skinns, L. (2011) *Police Custody - Governance, Legitimacy and Reform in the Criminal Justice Process,* New York: Willam Publishing.

Skolnick, J.H. (1966) *Justice Without Trial: Law Enforcement in Democratic Society*, New York: John Wiley and Sons.

Strang, H. & Sherman, L. (2012) Experimental Criminology and Restorative Justice: Principles of Developing and Testing Innovations in Crime Policy. In Gadd, D. Karstedt, S. and Messner S. (eds) *The Sage Handbook of Criminological Research Methods*, 395 – 424, London: Sage.

Waddon, A. & Baker, C. (1993) *Keeping Pace -A Survey of Custody Officers*, Bangor: University of Wales.

West Midlands Police (2014) West Midlands Police - About Us. Retrieved 12 September 2014 from http://www.west-midlands.police.uk/keeping-you-safe/about-us/our-structure/index.aspx

Weisburd, D. (2000) Randomized Experiments in Criminal Justice Policy: Prospects and Problems. *Crime & Delinquency*, 46(2), 181–193.

Wilcox, A. F. (1972) *The Decision to Prosecute* London: Butterworths.

Wildavsky, A. & Pressman, J. L. (1973) *Implementation: How Great Expectaions in Washington are Dashed In Oakland,* Berkeley: University of California Press.

Wilson, J. Q. (1968) *Varieties of Police Behavior,* Cambridge, MA: Harvard University Press.

# Appendix A – Prosecution Position

**The Turning-Point Project: Stage II**

Prosecution Decision Making

Project Position and Advice

This document is to be read in conjunction with other reports giving a wider explanation of the experiments objectives and methods.

It is confirmed that this project can be conducted in full compliance with the 4th Edition of the Directors Guidance on Charging. We see no conflict between either the spirit or the detail of this document, and the proposals that follow.

It is also our understanding is that the initial decision whether to prosecute or to offer participation in this project can in all cases be a police decision. This allows us to develop a simple and swift decision making model, and negates any need to train CPS prosecutors. This decision is based on the fact that police have an over-riding duty to consider the public interest test in all prosecution decisions and have the authority to divert from prosecution to an alternative disposal all categories of offences. This is consistent with recent advice from the Prosecution Team at CPS HQ, London, and is confirmed by the local CPS representative Mr Mark Paul who is an integral part of our Project Board

The Full Code Test: The Evidential Test

The Project only concerns itself with offenders whose cases have passed the first stage of the full code test – the evidential test. A key element of the project is the prosecution of non-compliance, and therefore all cases must be evidentially ready. The Project does not infringe upon the investigation process, and is first considered at the prosecution decision stage.

### The Full Code Test: The Public Interest Test

The second stage of the full code test is the public interest test. This is the arena within which key decision making for this project sits.

There is clear precedent for similar criminal cases leading to different outcomes, dependent upon the local availability of options. For example, some forces trialled the use of Youth Restorative Disposals, whilst others did not. This does lead to the likelihood that identical cases could lead to different disposals; dependant on what was available at the time at each specific location. Another example is current variation across the country in the nature and availability of local or community resolutions. So long as the outcome achieved locally is appropriate, proportionate and selected according to national prosecution policy then this diversity is both legal and ethical.

Likewise, it is accepted that similar or even identical cases will be dealt with differently within the scope of this project. In fact, that is a necessary part of the experiment, to allow the direct comparison of two different outcomes on two identical (or as close to identical as we can get) groups of cases. There is a clear public interest in the use of such experiments to test the effectiveness of prosecution policy in the real world, thus contributing to the development of the best possible policy. This situation is similar to the use of Randomised Control Trials in medical trials. The key element that makes such action ethically justifiable is that we genuinely do not know which of the two options are the best – such a test is required to obtain the data as to which is best, and so to inform future policy.

### Working within the Public Interest Test

Our position is to conduct a two-step public interest test in the relevant area during the implementation of this project. The first stage is to consider the test as it is considered now, under current policy. Cases we would currently divert to non-court disposals will be diverted as normal. Where we come to the conclusion that there currently isn't a suitable police disposal available that would allow us to divert a case from prosecution, officers would normally resort to a prosecution as best fulfilling the public interest. At this stage we will then move onto the second step: to reconsider this question in the light of the additional opportunity available: the Turning-Point Project.

The second step is to then apply a series of filters, followed by a random allocation of cases to either the experimental group, or the control group. The experimental group will go ahead to be offered participation in the entirely voluntary experiment, and the control group will continue to be prosecuted according to current policy.

Decision makers must always bear in mind that a proportion of the selected cases will be prosecuted, and therefore this must be a proportionate disposal for every case taken further to step two.

The purpose of the filters is to exclude unsuitable cases from participation in the project. Our target group is offenders whom we assess are both:

- Low risk of causing serious harm
- Likely to respond positively to the programme

Within the experiment we are developing a 'harm forecasting' tool to assist with the assessment of risk of harm to the public, but until that tool is ready to use we are having to create a series of 'filters' to weed out unsuitable cases.

The filters to select cases out of the experiment are as follows:

1) the offender is a young person (aged U18) at the time of the decision
2) The offender is judged on reliable information held to pose a medium or high risk of causing serious harm to another person *(professional judgement now, replaced by CHI asap*);
3) where, if found guilty, the sentence the court is likely to impose in this case, for this offender, will be a high level community order or custodial sentence (*use Mags Sentencing Guidelines*);

4) where the offender has previous convictions, they are assessed as being unlikely to complete their Turning-Point programme (*professional judgement again, but will require evidence to support, such as previous FTA etc*);

5) where an order of the court is required, and a similar outcome cannot be achieved within Turning-Point (e.g. Sex Offenders Order);

6) where, if found guilty, the sentence the court will impose includes an obligatory driving disqualification or licence endorsement (note: we can split cases to prosecute driving offences whilst including criminal offences in the project);

7) all drink/drug-driving offences

8) offences involving the use or threatened use of a firearm, imitation firearm, knife or an offensive weapon '*per s*e' (note this does not exclude cases of possession only)

9) where the consent of the DPP or a Law Officer is required to prosecute;

10) that involves causing a death;

11) connected with terrorism or official secrets;

12) sexual offences involving offenders or victims aged under 18;

13) any case where the custody officer believes it is necessary to refuse bail due to the threat posed by the offender to witnesses or the community;

14) any offender who does not have an address suitable for the service of summons;

15) any offender who does not live close enough to the relevant police area (Birmingham) to facilitate contact with police officers as might reasonably be required by the project;

16) any offender who is currently on bail to court for an offence, on licence or serving a court-imposed sentence in the community.

Also, the following cases are initially excluded whist further consideration is given to their possible inclusion:

17) domestic abuse cases according to CPS policy

18) hate crime according to CPS policies.

In all cases, the victims views on the matter of prosecution or otherwise of the offender will be taken into account, but they will not generally be overriding.

In two categories of offences the victims desire for an immediate prosecution will be dominant:

- any sexual offence;
- those where more than minor or passing harm has been caused (physical or psychiatric).

In the most sensitive and personal cases, the explicit consent of the victim will be required. We will not divert these cases from court without the victims support. Cases will be judged on a sliding scale, with the more sensitive and personal the offence, the more dominant the victims views on prosecution.

Where a case passes all of these filters then we will, in addition, ask the police decision maker (custody officer) to exclude any further cases which they feel have an overwhelming public interest in favour of a prosecution. This is intended to give the discretion to the decision maker to exclude unsuitable cases which we have been unable to foresee at the time of writing this guidance. It is anticipated this discretion will be rarely used, and each case thus excluded will be justified by the decision maker.

Cont…

CPS Definitions:

Domestic Abuse:

"Any incident of threatening behaviour, violence or abuse (psychological, physical, sexual, financial or emotional) between adults who are or have been intimate partners or family members, regardless of gender or sexuality".

An adult is defined as any person aged 18 years or over. Family members are defined as mother, father, son, daughter, brother, sister and grandparents whether directly related, in-laws or step-family.

Hate Crime:

“A Hate Crime is any criminal offence which is perceived by the victim or any other person, to be motivated by a hostility or prejudice based on a person’s race or perceived race; religion or perceived religion; sexual orientation or perceived sexual orientation; or against a person who is transgender or perceived to be transgender; or a person’s disability or perceived disability”.

## Randomisation

Once a case has passed all the filters it will then be randomly allocated to either the experiment or control group. Randomisation will be completed by computer programme designed by Cambridge University, ensuring independence and impartiality. The split between the proportion of cases diverted and prosecuted will be adjusted to ensure that the number of diverted cases is at a volume that the resources allocated to deal with them can manage. This will never drop below 10% of the entire sample.

## Voluntary Participation

If a case progresses though all of the filters and is randomly assigned to the experiment group, this will result in an offer being made to the offender to enter into a voluntary agreement with the police. This will involve them taking or desisting from certain actions, as recorded in a ‘Turning-Point Plan’. The offer will be that if they both complete the agreed plan and do not reoffend further within the agreed time period; we will not prosecute them in the instant case.

It will be made clear that participation is entirely voluntary, and that they are free to choose not to participate, in which case the prosecution will go ahead as normal. They will be entitled to and offered legal advice to assist them in making this decision.

### Holding the Prosecution Decision

If the offender agrees to participate they will be reported for consideration of the question of raising summons, and released from custody without charge. The prosecution decision will then be 'held' to see if they observe the agreed plan. If they do not, or if they reoffend, then the decision will be reviewed in the light of this, crucial, new information. As the effectiveness of the threat of prosecution is largely dependent upon our ability and willingness to prosecute in the event of non-compliance, then the general response to non-compliance must be prosecution. This will be via the raising of a summons at the earliest opportunity.

However, the decision to prosecute still needs to be made by applying the full code test. There may be circumstances when even after a failure to keep to the plan, or after reoffending, when a prosecution will not be in the public interest. There is an additional risk of the victim or witnesses being unwilling to support a prosecution at a later date; however it is hoped that with good witness care this possibility will be minimalised. Any requirement for CPS to authorise the prosecution, under the Directors Guidance, will also need to be complied with.

### The Crime Report: Disposal

As this project is in effect trialling a new type of out of court disposal, our force crime recording system (CRIMES) did not have a specific Clear Up Code (CUC) to use for cases finalised in this way. We have now created CUC 42 specifically for this project. This is to be used in cases where we decide not to prosecute immediately, and the offender has agreed to involvement in this project. It will also help track relevant cases and monitor any impact on crime detection performance.

Extent of the Project

The project will take place across the city of Birmingham, involving offences occurring on that area and investigated by Investigation Teams owned by Local Command Units

Produced and agreed with CPS March 2014

Author: Insp Jamie Hobday

CJS

## Appendix B – Survey Questions

1) Questions relating to officer attributes:

a) Q29 – Gender of officer

b) Q30 – Age of officer

c) Q31 - Educational level – on a 12 point scale

d) Q33 - Length of police service

e) Q34 - Length of police service in the custody environment

2) Questions relating to officers views on relevant parts of their role in custody:

f) Q2 –" Please select which one of the following options better describes how you made this decision: 'Instinctive natural and swift', or: 'Required time, concentration and effort"

g) Q3 – "When making these decisions generally, roughly what percentage of decisions do you make instinctively, naturally and swiftly?"

h) Q4a – "When making these decisions, do you feel the amount of discretion custody officers have is appropriate?"

i) Q5 – "The Force wants you to 'do the right thing' in these situations and allows you some freedom to use your discretion to make the best decisions. How clear are you about what the Force's view of the 'right thing' is in these situations?"

j) Q6 – "Out of court disposals could require more from the offender in order to avoid court. Please rate how important it would be to you to have the following options available:" Five options to grade: compensation; punishment; RJ; rehabilitation, and education.

k) Q7 - "When making these prosecution/diversion decisions how much of a priority in your mind is reducing reoffending?"

l) Q9 - "For low level offending, how effective do you think prosecution and court sentences are at reducing reoffending?"

m) Q13 - "What proportion of offenders passing through your custody suite do you think deserve a chance to put things right or engage in rehabilitative activities as an alternative to court?"

n) Q15 - "Which one of the following two statements do you feel is closest to reality as you experience it: Most of the offenders I have contact with are usually people who have made bad decisions or are in bad situations;or Most of the offenders I have contact with are usually bad people whom the public need protection from."

3) Questions relating to their understanding and support for Turning-Point, and experimentation generally (only answered by those staff that had worked on the project):

a) Q20 - "How well do you feel you understand what the project is trying to do with the offenders who are diverted from court and why it might work?"

b) Q21 - "How supportive do you feel about trying to work with this group of offenders in this way?"

c) Q23 - "How do you feel about the police running a live experiment like this generally?"

d) Q24 - "How comfortable are you with eligible cases being randomised to either once course of action or another, for the purposes of this experiment, even if this means that some cases end up not prosecuted?"

## Appendix C – Assessors Guidance Notes

### Assessing the 'Appropriateness' of Randomiser Rejections:

### Assessor Guidance Notes

This exercise is designed to investigate the rationales officers gave for excluding *otherwise eligible* cases from diversion into Turning-Point. Data is taken from the records submitted into the on-line 'Randomiser'. The focus of our interest is only those cases that were marked as eligible for randomisation, but with the officer decided to exclude for an additional reason. All the cases on the attached spread sheet have been marked as *eligible on all other criteria*. Custody officers had been trained that they could exercise their discretion in this way according to the following guidance:

"*Where a case passes all of these filters then we will, in addition, ask the police decision maker (custody officer) to exclude any further cases which they feel have an overwhelming public interest in favour of a prosecution. This is intended to give the discretion to the decision maker to exclude unsuitable cases which we have been unable to foresee at the time of writing this guidance. It is anticipated this discretion will be rarely used."* (TPP Prosecution Position, May 2012)

In the design of the experiment it was felt necessary to retain this discretion for decision makers as it was impossible to predict the nature of all possible cases coming in front of them, and a rigid adherence to the eligibility criteria could have led to cases that would damage the reputation of the force and the confidence of the public.

Assessment

You will find that in many cases there is limited information to make your assessment from. Please do you best to determine what the relevant circumstances were likely to be from the rationale given. Use these rules to give consistency across assessors:

- ❖ Don't ever assume that a custodial sentence would be likely from the rationale given. In every case included on the spread sheet the officer has indicated that a custodial sentence would NOT be likely, so make your assessment assuming that they were right in this.

- If multiple reasons are given, one of which is a clear eligibility criteria then mark the cases as an error entry as it is clear that it should have been excluded earlier on those grounds.

- It is a judgement whether a rationale for exclusion is appropriate or not. Below is some additional guidance to help assessors make this judgement consistently. It is important for assessors to bear in mind the reasons why the experiment was created: to test if this way of dealing with some offenders out of court would be more effective than prosecuting. Therefore the eligibility criteria were deliberately inclusive of cases that would never usually be diverted form court. We asked officers to stick to the criteria wherever possible.

- The difference between an ‘inappropriate’ and ‘clearly inappropriate’ assessment is one of degrees of scale. The ‘clearly inappropriate’ category is designed to catch those decisions that undermined the experiment by excluding cases that clearly should have been included.

Assessment Categories

There are 5 categories in to which we will categorise each decision/rationale:

- Error Entry
- Too little information to assess
- Appropriate
- Inappropriate
- Clearly Inappropriate

Assessors will need to be conversant with the eligibility criteria, which is reproduced at the end of this guidance. In addition there are further notes to help assessors interpret the criteria in a consistent manner:

**Error Entry:**

To be used where the case should have been excluded earlier in the decision making process. i.e. the error is in marking this as eligible and in need of special exclusion. This will include where any of the eligibility criteria have not been met, or there has been an error in the processing of the case. This includes cases suitable for any type of out of court disposal, where the powers of the court are required or bail is being refused, the PIC lives too far away from the operational area to participate, has multiple convictions, is being dealt with by a team not included in the experiment, or is of no fixed address. It includes cases where the PIC has declined or refused to participate in Turning-Point (Note: it would not include cases where the PIC hasn't been given the choice, and the officer has made that decision). It also includes cases where the entry in the randomiser is late and the PIC has already been charged. Many cases fall into this category.

**Too little Information to Assess:**

The rationale given by the officer gives no indication of whether the decision was appropriate or not, for example if they were directed to come to this decision by a senior officer, or if just the crime type is given. Please use this category sparingly.

**Appropriate:**

Where from the rationale given it appears that the decision to exclude the case was appropriate and in line with the decision making the experiment designers had in mind. It is not that the cases misses one of the standard eligibility criteria, but is something additional. This will generally be where there is some reason that makes the case more serious, such that the public interest can only be met by continuing with a prosecution. This could include some public or child protection issues (but won't include all these automatically), special vulnerability of the victim, where there are multiple on-going investigations, recent charges, and some abuse of trust cases (only where additional information is given to explain why this should be excluded). It will include cases where the PIC was already on Turning-Point and this subsequent case is a breach of that plan. It will also include cases where the PIC is likely to be deported (differentiating from those where they live too far away or have no fixed address, which should be excluded earlier and so should be marked as 'error entry').

**Inappropriate:**

Where the rationale appears to contradict both the eligibility criteria and is outside of the experimental objectives, but is understandable considering normal decision making outside of this experiment. This includes cases where a reason is given that is not part of the eligibility criteria and it is not clear that on that ground alone it is unsuitable. For example if 'breach of trust' is given with no further explanation (remember, all these cases have been marked as not likely to receive a custodial sentence). Other inappropriate reasons would include: the value of stolen/damaged property alone, delays in processing the PIC, involvement of young victims, multiple charges, recent out of court disposals issued, the crime type alone is used to justify exclusion, PIC has frustrated investigation or behaved badly to officers, or it is felt that a conviction is necessary (without mention of court orders). Any rationale that explains the CPS has already made a decision to charge will fall into this category too, as in these cases the advice was to go back to CPS seeking their support to divert.

Remember that officers were asked to be bold and to try and include as many cases as possible, so if the case hits all the eligibility criteria and unless there is a strong rationale for exclusion, it should be included. If not, then the assessor should classify it as 'inappropriate'.

**Clearly Inappropriate:**

Where the rationale given clearly shows a breach of eligibility criteria with no justifiable reasoning, and the decision is in contradiction to the experimental objectives. These will often evidence a lack of understanding of how the experiment seeks to work at tackling underlying issues and changing offending behaviour. This will always include cases of police assault as all selected cases have been identified as not likely to receive a custodial sentence and the policy decision was to include these. Other rationales that rely on lack of remorse, admission, or failure to name co-offenders will sit in this category. Cases involving on-going disputes, substance addiction, and mental health or learning difficulties will sit in this category where no other reasons to exclude are given. Cases where the officer believes the PIC won't engage without offering them to opportunity to decide will also sit in this category.

Recording and Returning your Assessment

Please record the category you assess each case as in the column on the right hand side of the spread sheet. Do not move the lines or change their order in any way. If you could get your results back to me ASAP I'd be really appreciative as I want to start analysing them next weekend.

Many thanks for your help with this – it's really appreciated. I will share my results with you all as you'll no doubt be interested in how our assessments compared to each other!

Remember: this is not an assessment of if the case was suitable for diversion, but an assessment of if the decision making was appropriate and in line with the experiments objectives.

Many thanks,

Jamie

## Appendix D – Assessments of ‘Appropriateness’

**RaterA1**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Clearly Inappropriate | 42 | 17.2 | 17.2 | 17.2 |
| | Inappropriate | 79 | 32.4 | 32.4 | 49.6 |
| | Appropriate | 35 | 14.3 | 14.3 | 63.9 |
| | error entry | 81 | 33.2 | 33.2 | 97.1 |
| | too little info to assess | 7 | 2.9 | 2.9 | 100.0 |
| | Total | 244 | 100.0 | 100.0 | |

**RaterA2**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Clearly Inappropriate | 78 | 32.0 | 32.0 | 32.0 |
| | Inappropriate | 39 | 16.0 | 16.0 | 48.0 |
| | Appropriate | 11 | 4.5 | 4.5 | 52.5 |
| | error entry | 107 | 43.9 | 43.9 | 96.3 |
| | too little info to assess | 9 | 3.7 | 3.7 | 100.0 |
| | Total | 244 | 100.0 | 100.0 | |

**RaterA3**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Clearly Inappropriate | 23 | 9.4 | 9.4 | 9.4 |
| | Inappropriate | 64 | 26.2 | 26.2 | 35.7 |
| | Appropriate | 70 | 28.7 | 28.7 | 64.3 |
| | error entry | 78 | 32.0 | 32.0 | 96.3 |
| | too little info to assess | 9 | 3.7 | 3.7 | 100.0 |
| | Total | 244 | 100.0 | 100.0 | |

**RaterA4**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Clearly Inappropriate | 72 | 29.5 | 29.5 | 29.5 |
| | Inappropriate | 51 | 20.9 | 20.9 | 50.4 |
| | Appropriate | 28 | 11.5 | 11.5 | 61.9 |
| | error entry | 88 | 36.1 | 36.1 | 98.0 |
| | too little info to assess | 5 | 2.0 | 2.0 | 100.0 |
| | Total | 244 | 100.0 | 100.0 | |

A1:A2

**Symmetric Measures**

| | | Value | Asymp. Std. Error[a] | Approx. T[b] | Approx. Sig. |
|---|---|---|---|---|---|
| Measure of Agreement | Kappa | .551 | .044 | 10.923 | .000 |
| N of Valid Cases | | 244 | | | |

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

A1:A3

**Symmetric Measures**

| | | Value | Asymp. Std. Error[a] | Approx. T[b] | Approx. Sig. |
|---|---|---|---|---|---|
| Measure of Agreement | Kappa | .551 | .044 | 10.923 | .000 |
| N of Valid Cases | | 244 | | | |

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

A1:A4

**Symmetric Measures**

| | | Value | Asymp. Std. Error[a] | Approx. T[b] | Approx. Sig. |
|---|---|---|---|---|---|
| Measure of Agreement | Kappa | .787 | .035 | 15.956 | .000 |
| N of Valid Cases | | 244 | | | |

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

A2:A3

**Symmetric Measures**

| | | Value | Asymp. Std. Error[a] | Approx. T[b] | Approx. Sig. |
|---|---|---|---|---|---|
| Measure of Agreement | Kappa | .431 | .041 | 10.036 | .000 |
| N of Valid Cases | | 244 | | | |

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

A2:A4

**Symmetric Measures**

| | | Value | Asymp. Std. Error[a] | Approx. T[b] | Approx. Sig. |
|---|---|---|---|---|---|
| Measure of Agreement | Kappa | .556 | .046 | 10.643 | .000 |
| N of Valid Cases | | 244 | | | |

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

A3:A4

**Symmetric Measures**

| | | Value | Asymp. Std. Error[a] | Approx. T[b] | Approx. Sig. |
|---|---|---|---|---|---|
| Measure of Agreement | Kappa | .498 | .043 | 11.300 | .000 |
| N of Valid Cases | | 244 | | | |

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

## Appendix E – IR Rationale Quotes

Quotes from Officer Rationales – specific, selected examples of poor reasons for rejection from the 100 IR cases.

| Rationale | Case Ref. |
|---|---|
| This is a violent offence against police aggravated by alcohol & Class A use - court disposal required | (49596) |
| PIC show little regard for his actions | (77124) |
| PIC is an alcoholic who has been causing nuemrous problems for neighbourhood officers at a problem location | (77385) |
| to protect police officer in the course of his duty | (89175) |
| There are multiple offences that are linked to alcohol… | (21981) |
| This is ongoing dispute involving violence and damage to property - court disposal necessary to prevent escalation | (25932) |
| and has a long history of being involved in road rage type incidents | (35009) |
| The vitim and PIC are know to each other and may lead to interference and possibly committing further offences. | (47458) |
| there are concerns the dispute may escalate. | (27412) |
| Not considered that project would have sufficient effect on him | (85909) |
| This offence was conduucted against his carers at a childrens home and is the equivalent of a domestic indicedent. PIC is already being given other support by social service | (94714) |
| Alcohol treatment requested RE sentence plus compensation for damage to store property | (12706) |
| This is a longstanding neighbour dispute, multiple diversionary measures have been tried and failed | (15932) |
| there are concerns the PIC will follow through with his threats against the IP, and will not be ordered into drug treatment | (76998) |

| | |
|---|---|
| This is a violent pre-meditated attack as a result of an ongoing dispute between two families and repurcussions are feared | (79777) |
| Offence against careers and property at the care home where he lives | (97690) |
| THIS DEFENDANT WAS RESIDENT IN A CARE HOME AND THREATENED ONE OF THE STAFF WITH A PAIR OF NAIL CLIPPERS | missing |
| although TPP can help with restoritive justice, there are a number of outstanding offeders not named by suspect, no remorse… | (12553) |
| Susepct … and the male she had an isue with last night has learnign difficulties … and I fear that we must act to safeguard this lcoaiton adn persons and not via turning point as the courts may not make an order regards her MH issues, but she has a care worker and suport from MH Team adn they are aware of arrest and charge an will increase thier supoport psot charge | (35829) |
| Cps haVE ALREADY AUTHORISED THEM PIC TO BE CHARGED ON THE GROUNDS THAT HE HAS ADMITTED TAKING THE ITEMS BUT DID NOT BELIEVE THIS TO BE A THEFT | (85479) |

## Appendix F – Training Questionnaire Results

| Officer (Ranked) | Training Day Attended? | Cambridge Training Day? | Researcher 1:1? |
|---|---|---|---|
| 1 | Yes | Yes | No |
| 2 | Retired | - | - |
| 3 | Retired | - | - |
| 4 | Yes | No | No |
| 5 | No | No | Yes |
| 6 | Yes | No | Yes (small group) |
| 7 | Yes | No | Yes |

## Appendix G – Checklist for Replication of Turning-Point

The authors’ suggestions to consider if replicating the experiment:

- Engage with the CPS early and get their support
- Attempt to get the CPS to simplify the selection criteria from the original version and to agree to diversion even if they have given a charge authority first
- Consider if you can reduce the upper seriousness limit from ‘a likely custodial outcome’ without making the decision-making too complex
- Consider excluding all breach of trust and assault public service workers on duty cases
- Ensure visible support from command level to reassure staff
- Build strong trust with the field researcher and develop the project together
- Engage with operational staff early and involve them in design
- Ensure you have a test phase during which you can adapt and improve the design
- Take as long as you need during the test phase before going live
- Overtly tackle issues of denial/attitude/remorse; wider offending; other seriousness and vulnerability as inappropriate rejections from an early stage
- Develop a bespoke version of the Gateway
- If using separate teams then identify key leads from each and bring them together to discuss performance and  progress regularly
- Carefully audit and watch the case flow in real time, responding to issues as they arise
- Share learning across the whole team, and share good news stories even wider
- Ensure you have capacity to coach individuals in the field