

**A Thesis Submitted for the Degree of PhD at the University of Warwick**

**Permanent WRAP URL:**

<http://wrap.warwick.ac.uk/146860>

**Copyright and reuse:**

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: [wrap@warwick.ac.uk](mailto:wrap@warwick.ac.uk)

# **Data Display as a Catalyst for Dysfunction in Police Performance Management Systems**

**Simon Guilfoyle**

**Warwick Business School**

**A thesis submitted to the University of Warwick for the Degree of**

**Doctor of Philosophy**

*“The police as a general rule hasn’t got a clue how to use statistics”.*

(Survey respondent 3644)

# Table of Contents

<b>Title Page</b>	i
<b>Table of Contents</b>	iii
List of Chapters	iv
List of Figures	x
List of Tables	xii
List of Appendices	xiv
<b>Acknowledgments</b>	xv
<b>Declaration</b>	xv
<b>Abstract</b>	xvi
<b>Glossary</b>	xvii
<b>Reference List</b>	254
<b>Appendices</b>	Attached

## List of Chapters

<b>Chapter 1: Introduction</b>	1
1.1 Background	1
1.2 Prior Research and Current Unanswered Questions	3
1.3 Aims, Objectives and Methods	6
1.4 Contributions	7
1.5 Thesis Structure	8
<b>Chapter 2: Literature Review</b>	10
2.1 Part One: Performance Management and the Policing Context	10
2.1.1 Introduction	10
2.1.2 Performance Measurement, Management and Information: An Overview	10
2.1.3 UK Police Performance Management: A Recent History	14
2.1.3.1 Post 1997: New Labour, New Momentum	17
2.1.3.2 A Change of Emphasis	19
2.1.3.3 UK Police Performance Management: Key Traits and Future Challenges	22
2.1.4 Binary Comparisons	26
2.1.4.1 Binary Comparisons and Behavioural Change	30
2.1.4.2 Binary Comparisons: A Summary	33
2.1.5 League Tables	33
2.1.5.1 League Tables and Behavioural Change	40
2.1.5.2 League Tables: A Summary	43
2.1.6 Numerical Targets	44
2.1.6.1 Numerical Targets: A Brief Timeline	44
2.1.6.2 Numerical Targets and Goal Setting Theory	45
2.1.6.3 Setting Numerical Targets	49
2.1.6.4 Targets and Behavioural Change	56
2.1.6.5 Targets in Policing	59
2.1.6.6 Numerical Targets: A Summary	64

2.2 Part Two: Data Display	65
2.2.1 Introduction	65
2.2.2 Data Display: An Overview	65
2.2.3 Reference Dependence	72
2.2.4 Statistical Process Control	77
2.3 Chapter Two: Summary and The Imperative for Further Research	81
<b>Chapter 3: Philosophical and Methodological Foundations</b>	<b>83</b>
3.1 Introduction	83
3.2 Critical Realism: Background and Key Characteristics	83
3.2.1 Reality and Knowledge	84
3.2.2 Entities, Structures and Mechanisms	85
3.2.3 Causation	87
3.2.4 Retroduction	89
3.2.5 Critical Realism and Statistical Analysis	90
3.2.6 The Benefits of a Critical Realist Account	91
3.3 Methodology and Methods	93
3.3.1 Methodological Framework	93
3.3.2 Research Design Considerations	94
3.3.3 Overarching Analytical Strategy	96
3.3.4 Survey Instrument: Background and Design	98
3.3.4.1 Section One: Experimental Testing	99
3.3.4.2 Section Two: Data Capture of Personal Experiences	101
3.3.4.3 Section Three: Respondents' Rank / Force Data	102
3.3.4.4 Section Four: Qualifying Questions	102
3.3.5 Survey Instrument: Pilot Phase	103
3.3.6 Survey Instrument: Live Phase	104
3.3.7 Robustness of the Survey Instrument	105
3.3.8 Reliability	107
3.3.9 Sampling and Biases	109
3.4 Summary	113

<b>Chapter 4: Qualitative Analysis and Findings</b>	114
4.1 Introduction	114
4.2 Sentiment Analysis	115
4.2.1 Inter-rater Reliability	116
4.2.2 Main Data Set	118
4.2.3 Permutations	119
4.2.4 Sentiment Analysis: Summary and Conclusions	120
4.3 Thematic Analysis	120
4.3.1 Introduction	120
4.3.2 Thematic Analysis: The Gioia Method	121
4.3.3 Application of the Gioia Method	122
4.3.4 Data Structure and Discussion	124
4.3.5 Aggregate Dimension: Engagement with Performance Information	127
4.3.5.1 Introduction	127
4.3.5.2 Second Order Theme: Data Display Format	127
4.3.5.3 Second Order Theme: Catalyst for Assumptions	131
4.3.5.4 Second Order Theme: Decision-making	135
4.3.5.5. Second Order Theme: Statistical Literacy	138
4.3.6 Aggregate Dimension: Behavioural Phenomena	142
4.3.6.1 Introduction	142
4.3.6.2 Second Order Theme: Catalyst for Inquiry	142
4.3.6.3. Second Order Theme: Basis for Operational Change	146
4.3.6.4 Second Order Theme: Perverse / Dysfunctional Responses	150
4.3.7 Aggregate Dimension: Organisational Climate / Cultural Context	156
4.3.7.1 Introduction	156
4.3.7.2 Second Order Theme: Psychological Impact	156
4.3.7.3 Second Order Theme: Influence of Senior Managers	158
4.3.7.4 Second Order Theme: Attitudes towards Performance Information	163
4.4 Theoretical Model	167

4.4.1 Theory Generation and Objectives of the Model	167
4.4.2 Commentary and Explication	171
4.4.3 Critical Realist Perspectives	175
4.5 Summary and Interim Conclusions	176
<b>Chapter 5: Quantitative Analysis and Findings</b>	178
5.1 Introduction	178
5.2 Question 1: Interpretation of the Stimulus	179
5.2.1 Binary Comparisons	179
5.2.2 League Tables	181
5.2.3 Numerical Targets	182
5.2.4 SPC Charts	183
5.2.5 Contextualised Peer Comparisons	185
5.2.6 Question 1: Summary of Findings	186
5.3 Question 2: Effect of Interpretation on Levels of Concern	187
5.3.1 Binary Comparisons	188
5.3.2 League Tables	189
5.3.3 Numerical Targets	189
5.3.4 SPC Charts	190
5.3.5 Contextualised Peer Comparisons	191
5.3.6 Question 2: Summary of Findings	192
5.4 Question 3: Behavioural Responses	193
5.4.1 Binary Comparisons	193
5.4.2 League Tables	195
5.4.3 Numerical Targets	196
5.4.4 SPC Charts	198
5.4.5 Contextualised Peer Comparisons	199
5.4.6 Question 3: Summary of Findings	201
5.5 Regression Analysis	201
5.5.1 Binary Comparisons	202
5.5.2 League Tables	204

5.5.3 Numerical Targets	205
5.5.4 SPC Charts	206
5.5.5 Contextualised Peer Comparisons	207
5.5.6 Regression Analysis: Summary of Findings	208
5.6 Additional Analysis	208
5.6.1 The ‘Don’t Know’ Group	208
5.6.2 Organisational Influence	210
5.6.3 Familiarity with Statistical Process Control	213
5.6.4 Familiarity with the Author’s Prior Work	216
5.6.5 Stratification and Inter-Block Analysis	217
5.7 Multiple Variants: Exploratory Analysis	218
5.8 Quantitative Analysis: Summary and Key Findings	220
5.8.1 Binary Comparisons	220
5.8.2 League Tables	221
5.8.3 Numerical Targets	221
5.8.4 SPC Charts	222
5.8.5 Contextualised Peer Comparisons	222
5.9 Quantitative Analysis: Overall Conclusions	223
<b>Chapter 6: Discussion</b>	225
6.1 Introduction	225
6.2 Theoretical Process Model	225
6.2.1 Antecedents to Data Display	227
6.2.1.1 Statistical Literacy	227
6.2.1.2 Influence of Senior Managers	230
6.2.1.3 Attitudes towards Performance Information	232
6.2.2 Core Concepts	234
6.2.2.1 Data Display	234
6.2.2.2 Decision Making	237
6.2.2.3 Behavioural Dysfunction	239
6.3 Summary	242

<b>Chapter 7: Conclusions</b>	244
7.1 Introduction	244
7.2 Summary of Findings	244
7.3 Theoretical and Practical Implications	247
7.4 Contribution	249
7.5 Further Research and Considerations	251

## List of Figures

Figure 2.1: Durham Police – Antisocial behaviour (ASB) table	27
Figure 2.2: Strathclyde Police - Violent crime table	28
Figure 2.3: Binary comparison chart	29
Figure 2.4: Sanction detection rate (i.e. solved crimes)	31
Figure 2.5: Violence with Injury (VWI) satisfaction rates	31
Figure 2.6: Missing people - performance data	32
Figure 2.7: iQuanta Snapshot – Leicestershire Police	34
Figure 2.8: Metropolitan Police crime data dashboard	35
Figure 2.9: iQuanta - Robbery Comparative Bar Chart	35
Figure 2.10: Police Forces Comparison Table	36
Figure 2.11 League Table – Teenage Conception Rates	39
Figure 2.12: Strathclyde Police – Performance report	51
Figure 2.13: West Yorkshire Police - Key Performance Indicators and Targets	52
Figure 2.14: British Transport Police – Sickness data table	52
Figure 2.15: Metropolitan Police – Stop and search performance data	53
Figure 2.16: British Transport Police – Sickness target chart	54
Figure 2.17: British Transport Police – Railway disruption target chart	55
Figure 2.18: British Transport Police – Railway disruption progress chart	56
Figure 2.19: Example of a SPC chart	78
Figure 2.20: Contextualised peer comparison chart 1	80
Figure 2.21: Contextualised peer comparison chart 2	80
Figure 3.1: Critical realist view of causation	88
Figure 3.2: Crime figures table	99
Figure 4.1: Data Structure: Binary Comparisons / League Tables / Numerical Targets	125
Figure 4.2: Data Structure: SPC charts	126
Figure 4.3: Theoretical Process Model – Influence of Data Display on Behavioural Dysfunction	170
Figure 5.1: Binary Comparison stimulus	180
Figure 5.2: League table stimulus	181

Figure 5.3: Numerical target stimulus	182
Figure 5.4: Statistical process control chart stimulus	184
Figure 5.5: Contextualised peer comparison chart stimulus	185
Figure 5.6: Theoretical Process Model – Statistical Testing of Variables	223
Figure 6.1: Theoretical Process Model	226
Figure 6.2: Antecedents to Data Display and Core Concepts	237

## List of Tables

Table 2.1: Studies relating to data display and performance information use	68
Table 3.1: Indicative statistics	105
Table 3.2: Gender frequencies	112
Table 4.1: Frequency table (all categories)	118
Table 4.2: Additional statistics (all categories)	118
Table 4.3 Second Order Theme Data Display Format	130
Table 4.4: Second Order Theme: Catalyst for Assumptions	134
Table 4.5: Second Order Theme: Decision-making	137
Table 4.6: Second Order Theme: Statistical Literacy	141
Table 4.7: Second Order Theme: Catalyst for Inquiry	145
Table 4.8: Second Order Theme: Basis for Operational Change	149
Table 4.9: Second Order Theme: Perverse / Dysfunctional Responses	155
Table 4.10: Second Order Theme: Psychological Impact	158
Table 4.11: Second Order Theme: Influence of Senior Managers	162
Table 4.12: Second Order Theme: Attitudes towards Performance Information	166
Table 5.1: Binary comparison stimulus (Question 1)	180
Table 5.2: League table stimulus (Question 1)	182
Table 5.3: Numerical target stimulus (Question 1)	183
Table 5.4: SPC chart stimulus (Question 1)	185
Table 5.5: Contextualised peer comparison stimulus (Question 1)	186
Table 5.6: Binary comparison stimulus (Question 2)	188
Table 5.7: League table stimulus (Question 2)	189
Table 5.8: Numerical target stimulus (Question 2)	190
Table 5.9: SPC chart stimulus (Question 2)	190
Table 5.10: Contextualised peer comparison stimulus (Question 2)	191
Table 5.11: Question 3 variables data table (Binary Comparisons)	194
Table 5.12: Question 3 aggregated variable statistics table (Binary Comparisons)	194
Table 5.13: Question 3 variables data table (League tables)	195
Table 5.14: Question 3 aggregated variable statistics table (League tables)	196

Table 5.15: Question 3 variables data table (Numerical targets)	197
Table 5.16: Question 3 aggregated variable statistics table (Numerical targets)	197
Table 5.17: Question 3 variables data table (SPC charts)	198
Table 5.18: Question 3 aggregated variable statistics table (SPC charts)	199
Table 5.19: Question 3 variables data table (Contextualised peer comparisons)	199
Table 5.20: Question 3 aggregated variable statistics table (Contextualised peer comparisons)	200
Table 5.21: League Tables Regression Analysis Data Table	204
Table 5.22: Numerical Targets Regression Analysis Data Table	205
Table 5.23: SPC Charts Regression Analysis Data Table	206
Table 5.24: Contextualised Peer Comparisons Regression Analysis Data Table	207
Table 5.25: Mann-Whitney U-tests ('Don't know' group)	209
Table 5.26: Cross-tabulation (Frequency variable against output from Questions 1, 2 and 3)	211
Table 5.27: Mann-Whitney U-tests (Frequency variable against output from Questions 2 and 3)	212
Table 5.28: Cross-tabulation ('SPC familiarity' variable against output from Questions 1, 2 and 3)	214
Table 5.29: Mann-Whitney U-tests ('SPC familiarity' variable against output from Questions 2 and 3)	215
Table 5.30: Cross-tabulation ('SG familiarity' variable against output from Questions 1, 2 and 3)	216
Table 5.31: Mann-Whitney U-tests ('SG familiarity' variable against output from Questions 2 and 3)	217

## **Appendices**

Appendix 'A': Survey instrument – Pilot phase contemporaneous document

Appendix 'B': Survey instrument - Correspondence

Appendix 'C': Survey instrument – MS Word export

Appendix 'D': Sentiment analysis – Coding template document

Appendix 'E': Sentiment analysis – Permutations matrix

Appendix 'F': Free text responses – Master sheet

Appendix 'G': Stratification and Inter-Block Analysis

Appendix 'H': Multiple Variants – Exploratory Analysis

## **Acknowledgments**

I would like to express thanks to my supervisors, Mark Johnson and Pietro Micheli, along with all the police officers who took the time to partake in the survey (I read every single one of your 13,135 comments and I sincerely hope the wisdom contained within them is absorbed and acted upon by the UK police service). I also thank all who have supported me in this 7-year endeavour – you know who you are.

## **Declaration**

This thesis is my own work and has not been submitted for a degree at another university. Excerpts from this original research were published within two articles in *Policing: A Journal of Policy and Practice* during 2015 and 2016 respectively; this content is highlighted accordingly within relevant sections of the thesis.

## Abstract

UK police performance management systems have long been blighted by recurrent forms of behavioural dysfunction, yet little is truly known about the root causes of this phenomenon. However, strong associations between dysfunction and certain types of performance information (such as league tables) are well-documented within the literature and it is believed such incidences chiefly arise as a consequence of adverse organisational conditions.

However, the prospect that the *data display* format used to present performance information may be partly responsible for inducing dysfunction remains unexplored. Whilst prior studies indicate data display can influence decision-making, none have evaluated whether it might also be a direct antecedent to behavioural dysfunction. Therefore, this study explores the possibility that the visual appearance of police performance information could itself be a factor capable of triggering dysfunction.

Adopting a critical realist stance, this research employs thematic analysis and experimental psychometric testing within a mixed methods framework, to produce a robust theoretical model and original explanatory account for the phenomenon of interest: *behavioural dysfunction in police performance management systems*.

The thesis speaks most directly to the burgeoning body of research evaluating *performance information use*, but is specifically positioned in an overlap between this domain and other studies that explore implications of data display; essentially, it synthesises and contributes to both fields of literature by, for the first time, establishing firm links between data display and behavioural dysfunction.

The original contribution generated by this study substantially extends findings of prior research by identifying data display as a potent ‘standalone’ catalyst, singularly capable of triggering dysfunction; in doing so, it postulates the existence and operation of a mechanism that systematically influences performance information users’ cognitive and behavioural responses, thereby producing the phenomenon of interest.

## **Glossary**

ACPO - Association of Chief Police Officers

APACS - Assessments of Policing and Community Safety

APCC – Association of Police and Crime Commissioners

APP - Authorised Professional Practice

BCU – Basic Command Unit

BVPIs - Best Value Performance Indicators

CDRP - Crime and Disorder Partnership

CSP – Crime Safety Partnership

CSR - Comprehensive Spending Review

HMIC – Her Majesty’s Inspectorate of Constabulary

IPCC – Independent Police Complaints Commission

MSF – Most Similar Force Group (See also MSG)

MSG – Most Similar Force Group (See also MSF)

NPM - New Public Management

PASC - Public Administration Select Committee

PCC - Police and Crime Commissioner

PMDU – Prime Minister’s Delivery Unit

PPAF – Police Performance Assessment Framework

PSAs - Public Service Agreements

PSU - Police Standards Unit

# Chapter One

## Introduction

### 1.1 Background

Behavioural dysfunction within the UK police service has long been associated with the way in which performance information is used (College of Policing, 2013; Independent Police Complaints Commission, 2013; Metropolitan Police Federation, 2014; Kent Police, 2013a; 2013b; Public Administration Select Committee, 2013a; 2013b; Wiltshire Police, 2014a; 2014b; Home Office, 2015). This phenomenon is also confirmed as being a highly-impactive and recurrent issue from the author's personal perspective, a serving police officer since 1995.

Such dysfunction often involves 'gaming', namely "*...making performance on the measured performance dimension appear better, when in fact it is not*" (Kelman and Friedman, 2009, p.917) however, cases also involve other counterproductive or even unethical practices. Examples include officers focusing on low level misdemeanours to the detriment of more serious crimes (Cumming, 2014), falsifying performance data (Whitehead, 2010), misrecording offences (Patrick, 2013) and even engaging in unlawful activity (Laville, 2012).

Although there is limited research into this phenomenon within police organisations, parallel instances of dysfunction are widely-documented in other public sector domains where comparable forms of performance information are used; this is particularly notable within the fields of healthcare (Bevan and Hood, 2006; Francis, 2013a; 2013b; Longman, 2013) and education (Hood, 2006; Rothstein, 2008; Heinrich and Marschke, 2010; Shorrock and Licu, 2013).

Instances from within healthcare settings include hospital records being manipulated (Healthcare Commission, 2009), patients being unnecessarily admitted to wards (Smith, 1993) and a variety of gaming tactics designed to avoid low positions in league tables (Marshall *et al*, 2000; Dawson *et al*, 2005). Similarly, in the education sector, there have been examples of 'teaching to the test' (Heinrich and Marschke, 2010), exam grade inflation (Koretz, 2002) and falsification of school attendance records to meet targets (Rothstein, 2008).

Examples of behavioural dysfunction were cited in the performance management literature as long ago as the 1950s (Granick, 1954; Berliner, 1956; Ridgway, 1956), with the consensus being that causes can generally be traced to various organisational conditions (Moynihan, 2016a). This literature is extensive and well-established; however, more recently, a branch of research into *performance information use* has grown into a vibrant and fascinating subset (see Moynihan, 2008; 2009; Micheli and Neely, 2010; Moynihan and Pandey, 2010; Moynihan *et al*, 2012; Nielsen, 2013; Kroll, 2013; 2015; Kroll and Moynihan, 2017).

Studies within this area confirm multiple factors affect performance information use, and in particular, can aid or impair what Moynihan terms ‘purposeful’ use of performance information (Moynihan, 2008); this is characterised by data-driven decision-making, leading to improved efficiency and effectiveness (Moynihan and Lavertu, 2011; Kroll, 2013; 2014). However, there is scant research that specifically evaluates underlying drivers for behavioural dysfunction, or conduct that Moynihan terms ‘perverse’ performance information use (see Moynihan, 2009). Similarly, there is next to nothing that assesses whether the presentation of performance information itself may be a contributory factor towards dysfunction (Moynihan, 2016a).

During the last 35 years, UK police performance management has become defined by the use of numeric measures and throughout this time, instances of dysfunction have also become widely-associated with certain types of performance information (Home Office, 2015). Three formats in particular are most commonplace and it is therefore these that attract the closest scrutiny in this research. They are:

1. *Binary comparisons* (i.e. comparing one isolated numeric value to another, then interpreting any difference between them as though it were a trend or trajectory; for example, ‘crime has risen by 6% compared to last month’).
2. *League tables* (i.e. peer comparison / ranking systems).
3. *Numerical targets* (i.e. precise aspirational numeric values, such as ‘reduce crime by 5%’).

Whilst the literature indicates dysfunction is often associated with league tables and numerical targets (Rothstein, 2008), it is acknowledged this could be attributed to a

number of underlying causes. However, the prospect that the type of *data display*<sup>1</sup> used to present numeric performance information may be partially responsible for inducing dysfunction remains unexplored. Similarly, it is not known whether some data display formats may actively moderate such dysfunction.

Therefore, whilst acknowledging wider factors known to affect behaviour within performance management systems, this research delves into the hitherto-overlooked prospect that data display could be an influencing condition in its own right. In doing so, it explores its influence on the cognitive and decision-making processes activated during engagement with performance information; it also questions whether its influence could also ultimately affect officers' behaviour.

## **1.2 Prior Research and Current Unanswered Questions**

Previous studies have explored relationships between performance information use and a variety of themes, including:

- Organisational culture (Franklin, 2000);
- Professionalism (Moynihan and Ingraham, 2004);
- Leadership (Dull, 2009);
- Organisational learning (Moynihan and Landuyt, 2009);
- Public service motivation (Moynihan and Pandey, 2010);
- Innovation (Moynihan *et al*, 2012);
- Leadership (Kroll and Vogel, 2013);
- Measurement system maturity (Kroll, 2014);
- Organisational routines (Kroll, 2015);
- Organisational change (Andersen and Moynihan, 2016a);
- Performance assessment routines (Moynihan and Kroll, 2016);
- Social equity (Kroll, 2017);
- Policy and governance (Moynihan *et al*, 2017);
- Programme evaluation (Kroll and Moynihan, 2017);
- Participation in setting goals (Choi and Moynihan, 2019);
- Citizen participation (Kroll *et al*, 2019), and;
- Problem-solving (Moynihan *et al*, 2019).

---

<sup>1</sup> *Data Display*: i.e. the way in which data are presented (see Tufte, 1990; 2001; 2013; Olsen, 2013a; 2015; James and Van Ryzin, 2015; Isett and Hicks, 2018).

These studies inform a range of overlapping fields; however, they do not consider the implications of, or for, perverse performance information use to any great extent. Meanwhile, separate research has identified a range of perversity and unintended consequences in performance management systems and this is enlightening when it comes to classifying types of dysfunction or providing case studies; examples include data distortion (Longman, 2013), unhealthy competition (Heinrich and Marschke, 2010), ‘tunnel vision’ (Pidd, 2005), unethical working practices (Patrick, 2009), as well as myriad forms of ‘gaming’ (Bevan and Hood, 2006).

These (and other) studies highlight consistent forms of dysfunction and also identify an apparent association between them and certain types of performance information. In particular, the use of numerical targets and league tables has long been associated with perverse outcomes, with several well-documented cases being recorded in the UK public sector (Hood, 2006; Neyroud and Disley, 2007). However, it remains a mystery as to whether these longstanding problems are a foreseeable by-product of the use of such formats (and if so, why this might be), or if dysfunction arises only as a consequence of improper application.

Conventional wisdom generally assumes dysfunction occurs as a consequence of improper application, or organisational factors (Latham, 2004; Franco-Santos and Bourne, 2011; Mannion and Braithwaite, 2012; Mountford and Wakefield, 2012; Moynihan, 2016a). For example, cases of gaming associated with league tables are attributed to them being used to ‘name and shame’ poor performers (Le Grand, 2010). However, whilst such diagnoses provide plausible explanations for dysfunction, they overlook the possibility that the data display format used to present performance information itself may also somehow influence dysfunction.

There are also very few studies examining the effects of data display on performance information use, resulting in a relative ‘blind spot’ when it comes to our understanding of this area. However, research exploring the general impact of data display provides findings that are potentially relevant to this field of study, as they demonstrate the way in which data are presented can influence how visual cues are perceived (Tufte, 2013; Olsen, 2013a; 2013b).

Therefore, as a primary objective of performance information is to inform decision-making (Moynihan, 2008), research in the field of data display could open up

implications for performance information use. Studies within healthcare also suggest the way in which data are presented can affect interpretation (Brewer *et al*, 2012; Kurtzman and Greene, 2016; Schmidtke *et al*, 2017a) and influence decision-making (Elbel *et al*, 2014; NHS, 2018).

These comparatively recent studies establish a link between data display and decision-making; however, although prior research is clear in this respect, there is nothing that directly examines any potential relationship between such decisions and subsequent behavioural dysfunction. It therefore seems vital to explore these links to more fully understand underlying factors responsible for triggering dysfunction, as this could not only produce fresh theoretical and practical contributions, but also solve longstanding problems through improved design of performance information.

In respect of the broader policing context, prior research has explored a range of disparate subject areas; for example, Van Maanen (1973; 1974; 1975) studied police behaviour, culture and socialisation, Wilson (1968) examined organisational influence on the actions of street-level officers and Holdaway (1983) studied race relations. Meanwhile, Manning (1977) produced a comprehensive analysis of police legitimacy, whilst Waddington (1999) studied the ‘canteen subculture’, and more recently, Marks (2004) researched organisational transformation, whilst Ramshaw (2012) explored typologies of patrolling styles.

These studies examine various elements of police culture, organisation and behaviour; meanwhile, few have focused on police performance information use and none have specifically researched causes of behavioural dysfunction. Consequently, although there is an extensive body of literature on performance management in general, as well as a broad spectrum of policing studies, synthesis of the two fields is rare. Furthermore, whilst there is a wide expanse of research into purposeful performance information use, there is comparatively little on associated perversity (Moynihan, 2016a). This doctoral study seeks to address this shortfall.

Finally, although this research sits primarily beneath the umbrella of the performance management literature, and specifically within the field of performance information use, its focal point lies in an overlap between this domain and that of data display; fundamentally, this thesis synthesises and contributes to both fields of literature by establishing firm links between data display and behavioural dysfunction.

### 1.3 Aims, Objectives and Methods

The primary aim of this study lies in establishing whether data display is a standalone factor capable of influencing the likelihood, nature, or extent of behavioural dysfunction in police performance management systems. It assesses the effects of common forms of data display used to present police performance information, with the intention of identifying any notable relationships, tendencies or mechanisms that may either heighten or moderate dysfunction.

The research therefore examines the experiences of UK police officers to identify recurrent themes and patterns associated with dysfunction and / or data display. It also deploys experimental stimuli to assess various methods of presenting numeric police performance information, in order to establish whether data display directly affects assumptions about performance and if such assumptions ultimately act as a catalyst for particular behavioural outcomes. The overarching research question is:

*“Does data display influence the likelihood, nature or extent of behavioural dysfunction in police performance management systems, and if so, why?”*

In other words, in respect of dysfunction:

1. How likely is it to happen?
2. What happens?
3. How much of it happens?
4. Why does it happen?

To answer the research question, this study analyses qualitative and quantitative data, gathered through a large-scale survey instrument deployed throughout all UK police forces, which incorporates embedded psychometric micro-experiments. Mixed methods (see Creswell, 2008; 2009; 2013) are employed: firstly, *sentiment analysis* (see Turney, 2002; Wilson *et al*, 2005) is conducted, followed by Gioia’s (2012) thematic analysis method; this identifies key concepts and themes that are then arranged into *data structures*. The analysis identifies notable relationships, enabling the generation of an empirically-grounded theoretical model capable of demonstrating the influence of data display upon behavioural dysfunction.

Next, statistical analysis is undertaken to uncover any strong patterns or tendencies present within the experimental component of the study. This is followed by an assessment of the strength and direction of effects observed when users interact with various performance information formats, along with further development of the theoretical model. This approach is novel, as the bulk of research into performance information use tends to involve observational studies, with relatively few studies incorporating experimental designs (Kroll, 2014; Moynihan *et al*, 2016).

#### **1.4 Contributions**

This research extends the findings of prior studies that indicate data display affects decision-making, by identifying new relationships and recurrent patterns of cognitive and behavioural responses typically activated during engagement with certain data display formats. Centrally, the theoretical model explicates a sequence of events that theorises links originating from data display, through assumptions and decisions, to behavioural phenomena; moreover, it postulates the existence and operation of a mechanism that consistently sets users on a path towards dysfunction when particular data display formats are used.

Furthermore, the study enriches the literature on performance information use and contributes to management research by firmly establishing the prospect that data display is indeed a factor singularly capable of triggering behavioural dysfunction. It introduces the notion that the characteristics of some formats habitually impair engagement with performance information and consequently produce remarkably consistent types of dysfunction; it also establishes that alternative forms of data display promote effective interpretation and decision-making, thereby leading to reduced levels of dysfunction.

The study is therefore unique in demonstrating that particular formats consistently influence decision-making in a particular way, directly leading to particular forms of behavioural dysfunction, even in otherwise ‘healthy’ organisational environments and where all other conditions remain constant. In summary, it comprehensively establishes that the data display format chosen to present numeric police performance information has a substantial impact upon behavioural dysfunction; an assertion that is not made in any prior research.

## 1.5 Thesis Structure

Following this introduction, Chapter Two sets out the context for the research by reviewing performance management literature, along with primary documentation obtained directly from UK police forces and additional secondary material. It discusses performance management, performance measurement and the use of performance information, before tracing the recent history of UK police performance management. It then explores specific themes relating to dominant forms of police performance information, before considering less widely-used alternatives. This therefore provides an assessment of the current state of the art.

After exploring these key themes, it considers their implications through the theoretical lens of *data display*. It then examines the characteristics of dominant police performance information formats, along with patterns of dysfunction associated with their use, and asks whether such adversity may be at least partially mitigated by the use of alternative formats. Finally, it presents the overarching research question.

Chapter Three sets out the philosophical and methodological foundations that underpin the research, exploring the underlying ontological and epistemological assumptions that inform research methods. An overview of the philosophical stance, *Critical Realism* (see Bhaskar, 1975; 1979; 1986) is presented, with discussion on its relevance to this study. The chapter then reviews chosen methodology and the mixed methods employed to generate findings capable of answering the research question.

Chapter Four presents the qualitative component of the research, exploring themes within the 13,135 free text comments provided by survey respondents; the data structures produced by thematic analysis are then examined, followed by the theoretical model and propositions about a mechanism that may be responsible for producing observed effects. Chapter Five then reports on statistical analysis of quantitative data produced by the survey instrument, along with results of the experimental portion of the study; this analysis assesses specific aspects of the model and tests the strength of relationships between its components and the phenomenon of interest.

Chapter Six discusses the implications of the study's findings, as well as how they connect with literature, theory and wider research. It also explores the theoretical model in greater depth, focusing on themes identified as *antecedents to data display*, as well as *core concepts* of the study. Finally, Chapter Seven provides a summary of key findings, followed by explication of the study's main contributions, overall conclusions and potential avenues for future research.

This thesis structure, and the approaches employed within it, aim to produce a theoretically sound and practically effective framework for addressing the research question and producing a novel and robust original contribution.

## Chapter Two

### Literature Review

#### 2.1 Part One: Performance Management and the Policing Context

##### 2.1.1 Introduction

This section sets the context for this research by first reviewing relevant performance management literature, along with primary documentation obtained directly from UK police forces and additional secondary material, such as open source content. It reviews the topics of performance management, performance measurement and the use of performance information, before tracing the recent history of UK police performance management. This is followed by an examination of the characteristics of dominant police performance information formats, along with discussion on their strengths and weaknesses, as well as the behavioural implications of their use.

The second part of the chapter examines the literature on *data display* and *reference dependence*, discussing prior research in these fields and identifying how application of associated theories may generate fresh perspectives about the phenomenon of interest. Next, it presents a typology of the main types of behavioural dysfunction observed in police performance management systems. Finally, it outlines the imperative for integrated research into data display and the use of performance information, before presenting the research question.

##### 2.1.2 Performance Measurement, Management and Information: An Overview

‘Performance management’ and ‘performance measurement’ are broad subjects (Bourne *et al*, 2018) and have distinct meanings, yet the links between them can be difficult to explain (Pavlov and Bourne, 2011). The terms themselves are often used interchangeably (Radnor and McGuire, 2004) and are sometimes integrated, as in ‘*performance measurement and management (PMM) systems*’ (Bourne *et al*, 2018). Although there are no universally-accepted definitions of these terms, the literature identifies common characteristics, which assists in conceptualising them.

For instance, performance *management* may be viewed as part of an overall control system (Otley, 2003), whereby officials or managers determine organisational goals,

then hold subordinates accountable for performance against these goals (Moynihan, 2005; 2008; Nielsen, 2013). Objectives can be manifold (Behn, 2003), however the ultimate goal of performance management is generally considered to be to improve organisational performance (Whitaker *et al*, 1982; Behn, 2002; Micheli and Neely, 2010; Shane, 2010; Bititci *et al*, 2018).

Meanwhile, performance *measurement* provides “...the information system at the heart of the performance management process...” (Bititci *et al*, 1997, p. 533). This system utilises performance *information* (Moynihan and Ingraham, 2004) to quantify “...the efficiency and effectiveness of past actions through the acquisition, collation, sorting, analysis, interpretation, and dissemination of appropriate data” (Neely, 1998, p.5). Therefore, these terms may be conceived taxonomically; performance management incorporates performance measurement, which in turn utilises performance information as its currency.

Performance management and measurement systems are an important means for influencing behaviour (Bourne and Bourne, 2011), as well as monitoring and controlling resources (Franco-Santos *et al*, 2012). They have broad implications for providing political bargaining power (Moynihan, 2008; Micheli and Neely, 2010) and informing public discourse (Jackson, 2011), however, Bird *et al* (2005) suggest the overriding purpose of public service performance measurement is threefold:

“...to establish ‘what works’ in promoting stated objectives of the public services; to identify the functional competence of individual practitioners or organizations; and public accountability by Ministers for their stewardship of the public services” (2005, p.2).

Performance measurement can be a highly technical task requiring expertise (Woodcock *et al*, 2019); effective measurement is also contingent on the use of appropriate types of performance information (Moynihan, 2008). Consequently, measurement systems should ideally comprise a mix of qualitative and quantitative indicators (Kaplan and Norton, 1996), be directly related to organisational goals and strategy (Ittner and Larkner, 2003) and be presented in formats that provide an accurate, contextualised picture of performance (Wheeler, 2000).

Even where it is difficult to accurately quantify performance, “It is better to imperfectly measure relevant dimensions than to perfectly measure irrelevant ones” (Bommer *et al*, 1995, p.602). Overall, measures should produce data that can be used to aid decision-making, promote effective management and drive improvement (Neely *et al*, 2002).

The effective use of performance information can therefore generate myriad benefits, such as enhancing accountability, efficiency and service improvement (De Bruijn, 2002; Collier, 2006; Ammons and Rivenbark, 2008; Heinrich, 2008; Heinrich and Marschke, 2010), informing scholarship (Moynihan and Pandey, 2010), supporting human resource management practices (Pavlov *et al*, 2017), as well as assisting “...strategic planning, resource allocation, program management, monitoring, evaluation, and reporting to internal management, elected officials, and citizens or the media” (De Lancer Julnes and Holzer, 2001, p.695).

Performance information can also aid resource allocation, act as a basis for making peer comparisons, and inform planning and budgeting decisions (Jackson, 2011). It can assist diagnostics (Nielsen, 2013), enrich programme evaluation (Kroll and Moynihan, 2017), promote pro-organisational behaviour (Fama and Jensen, 1983; Eisenhardt, 1989), support goal alignment and value creation (Foss and Stea, 2014), foster learning and improvement (Moynihan *et al*, 2009) and assist steering and controlling (Van Dooren *et al*, 2010).

Factors known to promote effective performance information use include measurement system maturity (Taylor, 2009; Kroll, 2014), stakeholder involvement (Bourdeaux and Chikoto, 2008; Moynihan and Pandey, 2010), leadership support (Boyne *et al*, 2004; Melkers and Willoughby, 2005; Dull, 2009; Kroll and Vogel, 2013), innovation (Moynihan, 2005; Johansson and Siverbo, 2009; Moynihan *et al*, 2012), support capacity (i.e. resources, expertise and information technology) (Berman and Wang, 2000; Behn, 2006; 2008a; Moynihan and Hawes, 2012) and goal clarity (Kaplan and Norton, 1996; Moynihan and Landuyt, 2009).

Centrally however, it is argued the primary objective of performance information use is to foster effective decision-making (Hatry, 1999a; De Lancer Julnes and Holzer, 2001; Ammons and Rivenbark, 2008; LeRoux and Wright, 2010; Lavertu and

Moynihan, 2012; Hvidman and Andersen, 2013; Pollanen *et al*, 2017). Moynihan (2008) posits:

“The communication of performance information is intended to act primarily as a stimulus to the decision-making process – provoking, informing, and improving the quality of decisions” (2008, p.7).

In order to support decision-making, it is imperative that performance information is relevant (Richardson, 2013) and presented in formats seen to be “...accurate, complete and trustworthy...” (Micheli and Pavlov, 2017, p.4; see also Micheli and Mari, 2014). Moore and Braga (2003) suggest using a range of indicators that provide multiple perspectives on operational activity; similarly, citing the notion of a *Balanced Scorecard* (Kaplan and Norton, 1996; Bourne, 2008; Hoque, 2014), they argue measurement systems should incorporate a combination of outcome measures, output measures, and financial measures. Likewise, drawing on Brown (1996), Moss *et al* (2007) propose that a suite of balanced indicators could be categorised under ‘input’, ‘process’, ‘output’ or ‘outcome’.

This type of categorisation is fairly representative of thinking on multidimensional performance measurement. Overall, as long as the ‘right measures are used in the right way’ (College of Policing, 2014), performance measurement systems are essential components of efficient and effective organisations. If measures “...capture the relevant characteristics of underlying operational processes” (Lohman *et al*, 2004, p.271), trends and changes in data can be interpreted and efforts made to identify their causes, learn and improve (Moynihan and Landuyt, 2009).

Moynihan (2008; 2009) and Moynihan and Lavertu (2011) cite four types of performance information use: *purposeful*, *passive*, *perverse* and *political*. Kroll and Vogel (2013) define purposeful performance information use as “...making better-informed management decisions based on performance data” (2013, p.976). Moynihan and Lavertu (2011) and Kroll (2013; 2014) assert such use promotes efficiency and effectiveness, whilst Moynihan (2009) observes:

“The central hope of performance management doctrine is that public employees use data to improve program performance” (2009, p.2).

Passive performance information use occurs when individuals do the minimum required to comply with reporting requirements, rather than using data to understand and improve performance (Radin, 2006). Perverse performance information use refers to when gaming and other dysfunctional behaviour arise in an attempt to increase performance outputs, often at the expense of underlying organisational goals (Moynihan, 2009). Finally, political performance information use occurs when data are used to claim success or advocate for resources (Moynihan, 2008).

This study focuses chiefly on the ‘purposeful’ and ‘perverse’ categories, aiming to promote the former, whilst exploring potential root causes for the latter; the following section will explore how these considerations have played out in UK policing over recent years.

### **2.1.3 UK Police Performance Management: A Recent History**

Despite the extensive literature on performance management in the public sector, there is relatively scant research examining policing, and less still in respect of the UK policing experience. Therefore, the following section will explore the recent history of UK police performance measurement and management, tracing its development during the last 35 years, and considering the implications of established models and practices.

The roots of modern day police performance management in the UK can be traced to the early 1980s, and specifically to the influence of Home Office Circular 114/83 (Home Office, 1983), which exhorted the service to demonstrate improvements in the ‘Three Es’ of efficiency, economy and effectiveness. This heralded the beginning of a paradigmatic shift towards managerialism across the UK police service, which came into being as part of wider public service reforms of the era, known as New Public Management, or NPM (Hood, 1991; 1995; 1996; 1998; Pollitt and Bouckaert, 2000; Pollitt, 2002).

The NPM reforms in the UK were part of a global paradigm, which saw a shift in the disposition of governance in public services towards a private sector ethos. In particular, governments in Western Europe, New Zealand, Canada, Australia and the US adopted NPM principles, requiring the public sector to demonstrate fiscal accountability and measurable evidence of performance. In 1993, US President

Clinton signed the Government Performance and Results Act (GPRA), which obligated agencies to establish performance targets and strategic plans, and be measured against results (Radin, 2000; Heinrich, 2008). This legislation subsequently served as a template for NPM reforms worldwide.

NPM ideology was characterised by an emphasis on marketisation, value for money, contractual relationships, the notion of citizens as ‘customers’ and a focus on performance outputs and standards (Pollitt, 1993; Dunleavy and Hood, 1994; Hood, 1995; McLaughlin *et al*, 2001; Boyne *et al*, 2004). The reforms were largely “...predicated on the belief that the public sector can learn from the private sector...” (Ghobadian *et al*, 2009, p.1520) and were promoted with the dual purpose of “...legitimizing the police service to the electorate and ordinary citizens, while encouraging efficiencies of resource use” (Hoque *et al*, 2004, p.59). Advocates of NPM asserted these aims could be achieved through:

“...new forms of budgeting, quasi-market techniques, performance-related payments, outsourcing and the expansion of internal and external contractualisation, including the closer involvement of the private sector” (De Maillard and Savage, 2012, p.364).

Proponents of NPM argued reform was necessary due to a perception that public services were inefficient, unaccountable, and laden with disproportionate bureaucratic practices (Hughes, 2003). In particular, performance management was deemed opaque and unsophisticated (if present at all); “...measures which did exist were *ad hoc* and far from systematic” (Hughes 2003, p.157). In contrast, NPM offered “...explicit standards and measures of performance” (Hood, 1996, p.257), along with accountability through the use of performance targets (Jansen, 2008).

In addition to a general shift in public attitude towards institutions during the post war decades (Kramer, 2009), the police service encountered strong public desire for greater accountability and transparency (Davis, 2012a; HMIC, 2014). Hood (1995) suggests declining levels of trust contributed towards the appetite for closer scrutiny of performance, whilst Prenzler (2007) proposes this was further fuelled by allegations of police corruption and misbehaviour during the 1960s and 1970s, which included alleged bribery (Brown, 2013), use of excessive force, (Lewis, 2010), miscarriages of justice and corruption (Loftus, 2010).

Consequently, these changing social attitudes, as well as apparent politicisation of the service (for example, during the Miner's Strike – see Coulter, 1984), led to the loss of what Reiner (2010) called the 'Golden Age' of policing. NPM reformers believed that the managerialist model represented the antidote to these ills; it was also a neat fit with the dominant political discourse of the time. However, the police service enjoyed a relatively 'light touch' in terms of the intensity of NPM reforms until the early 1990s, whereupon a programme of far-reaching police reforms was launched (Savage, 2007; De Maillard and Savage, 2012).

At the vanguard of these reforms were the 1993 White Paper on Police Reform (Home Office, 1993a) and the Sheehy Report (Home Office, 1993b). The White Paper claimed the police service lacked clarity of purpose and proposed it was necessary "...to set clear objectives and to introduce changes to the framework within which the police operate" (1993a, p.4). The paper reflected the government's intent to institutionalise performance management within policing, bringing with it increased monitoring and for the first time, comparative league tables, such as those already seen in the health and education sectors (Home Office, 1993a).

The Sheehy Report proposed 272 recommendations, including an overhaul of performance management, which, Sheehy argued, should be linked to pay. In line with the focus on performance evident in the White Paper, Sheehy proposed greater personal accountability for chief officers, plus "...performance measures for units and the officers managing them, as well as individuals operating at the front line" (Home Office, 1993b, p.53). Although many of the Sheehy reforms were ultimately never implemented, several of the recommendations pertaining to performance management were introduced into forces (Leishman *et al*, 1995).

NPM-oriented performance management philosophy was further consolidated within policing in 1994, when the Police and Magistrates Court Act (Home Office, 1994) received Royal Assent. The Act conferred authority upon the Home Secretary "...to direct Police Authorities to establish levels of performance ('performance targets')" (Home Office, 1994, p.16). It also centralised control over police performance management and radically altered the relationship between police forces and the Home Office; furthermore, it established a "...springboard for intervention by successor administrations" (Loveday, 2006, p.284).

### **2.1.3.1 Post 1997: New Labour, New Momentum**

The foundations laid by the Conservative Governments of the 1980s and 1990s were seized upon by the New Labour administration that came to power in 1997. Thereafter, performance management throughout policing became characterised by the sheer vigour with which it was applied (Royal Statistical Society, 2005; Pollitt, 2006). Hood (2007) posits, “British government departments from 1998 arguably took the target approach at the top level of government to a point hardly seen since the demise of the USSR” (2007, p.96).

Over subsequent years there followed an exponential tightening of governmental grip on police performance management, characterised by top-down control, explicit performance indicators, numerical targets, ‘league tables’, audit and inspection, as well as the holding to account of individuals for perceived ‘poor’ performance (Loveday, 1998; 2005).

This momentum intensified following the 1998 Comprehensive Spending Review (CSR), when 366 Public Service Agreements (PSAs) were introduced across the public sector, which included over 600 performance targets (Chief Secretary to the Treasury, 1998a; 1998b; James, 2004). Micheli and Neely (2010) describe how three overarching PSAs generated multiple individual targets, initiatives, actions and indicators, as they descended to the operational level. In one example:

“...these three PSAs and 10 indicators were converted into 49 separate indicators. At the local level, the 49 indicators identified in the Home Office’s interim framework were converted into 78 separate indicators” (2010, p.597).

In 1999, the Local Government Act (Home Office, 1999) introduced the ‘Best Value’ agenda, intended to ensure value for money in public services. The Act designated ‘Best Value Authorities’ (these included Police Authorities), requiring them to measure performance against a further suite of indicators and targets. In policing, many of these ‘Best Value Performance Indicators’ (BVPIs) related to crime and efficiency outcomes, or public satisfaction rates (DETR, 1999; Home Office, 2000a). The expectation was that PSAs would translate into BVPIs, which would enable performance to be tracked at the local level (Micheli and Neely, 2010).

The plethora of targets introduced under the PSA framework and Best Value initiative was expanded in 1999 by the publication of the Home Office's 'Strategic Plan for the Criminal Justice System' (Home Office, 1999) and a further supplementary publication in 2000 (Home Office, 2000b), each of which designated further numerical targets, such as:

- 30% reduction in the level of vehicle crime within five years;
- Halving the time from arrest to sentence for persistent young offenders, from 142 days to 71 days, and;
- Reduce burglary by 20%.

(Loveday, 2000; Collier, 2006).

In order to ensure that activity was focused toward meeting targets, the Prime Minister's Delivery Unit (PMDU) was inaugurated in 2001. The Unit's purpose was to monitor and drive performance in around twenty key public sector targets (Hood and Dixon, 2010). The PMDU reported directly to the Prime Minister, being renowned for its interventionist disposition, preferences for the use of comparative data, and practice of holding managers personally to account. This approach was labeled 'Deliverology' by Michael Barber, the PMDU's head (Crace, 2007).

During the same year, a further report, 'Policing a New Century: A Blueprint for Reform' (Home Office, 2001), depicted a police service in crisis, citing high levels of crime, poor detection rates, and a decline in public confidence. Drawing upon assessments conducted by the Audit Commission (Audit Commission, 2000), its authors argued there was too great a disparity between levels of performance amongst forces. The paper therefore prompted further reform and proposed a new oversight function, namely the Police Standards Unit (PSU) – a body charged with scrutinising and improving police performance. The proposed methods for achieving these improvements focused on more target-setting and centralised control.

The 2001 report paved the way for the Police Reform Act of 2002 (Home Office, 2002), which heralded a further significant movement towards the continuing centralisation of UK policing (Jones, 2003). Drawing upon the much-vaunted deficiencies exposed in the previous year's report and using them as a springboard for faster and deeper reform, the Act:

“...provided the Home Secretary with powers to set national police priorities, remove chief officers and directly intervene in the management of local police services” (Loveday, 2006, p.284).

These powers were promptly invoked by the Home Secretary, David Blunkett, who initiated a series of inspections of all Basic Command Units (BCUs), conducted by Her Majesty’s Inspectorate of Constabulary (HMIC) between 2002 and 2004. The Act inaugurated the Police Standards Unit as a statutory body and established the first National Policing Plan, which enabled BVPIs to be used as a means of ranking forces’ comparative performance (Collier, 2006; De Maillard and Savage, 2012).

In addition to its well-established inspection function, HMIC was also charged with designing and applying the new ‘Policing Performance Assessment Framework’ (PPAF) (Home Office, 2005); this was a comparative performance framework designed to assess forces in areas such as:

“...tackling crime and serious crime, protecting vulnerable people, satisfaction and fairness and resource and efficiency. Each force was judged along a scale of ‘poor/fair/good/excellent’ and on whether their performance in terms of direction was ‘deteriorating, stable or improving’” (De Maillard and Savage, 2012, p.369).

PPAF, which also influenced the character of some forces’ local performance frameworks (Hunton *et al*, 2009), later evolved into the ‘Assessments of Policing and Community Safety’ (APACS) framework, which aimed to assess the wider performance of statutory Crime and Disorder Partnerships (CDRPs), rather than solely police activity (Home Office, 2007). Despite the slight shift in emphasis, the heavy focus on crime rates as a performance measure remained, along with numerical targets and league tables; at one point, almost 300 targets and indicators applied to APACS (Golding and Savage, 2008).

### **2.1.3.2 A Change of Emphasis**

The NPM-influenced orientation of police performance management continued to ingrain itself throughout the early 2000s, despite growing opposition from within the service (Golding and Savage, 2008). Much of this disquiet was associated with

concerns that target-based performance management was responsible for generating dysfunctional behaviour. For example, Neyroud and Disley (2007) argue:

“...pressure to meet targets encourages managers to focus on volume crime investigations which are less resource intensive, at the expense of proper investigations of more serious crimes” (2007, p.563).

Apparently heeding these concerns, in 2008, the Home Secretary, Jacqui Smith, announced a move away from centralised performance management (Home Office, 2008a). This shift in emphasis resulted in the introduction of the ‘single confidence measure’; in Smith’s words, “...in future there will only be a single top down target for police forces – on improving public confidence” (Home Office, 2008a, p.6).

This was accompanied by the new ‘Policing Pledge’ – a set of standards and service level commitments that police forces were expected to adhere to (Home Office, 2008a; HMIC, 2009). Curiously, the Green Paper that proposed the single confidence measure and the Policing Pledge also initiated a more aggressive and interventionist stance for HMIC, asserting:

“HMIC will have a more hard-hitting role in exposing underperformance of police forces and authorities and ensuring it is tackled...” (Home Office, 2008a, p.82).

Nevertheless, the statement of intent to move away from centrally-driven target-based performance management represented a significant change in the direction of national policy. The notion of central government, nominally at least, relinquishing the reins of police performance management was continued by the newly-elected Conservative and Liberal Democrat Coalition Government in 2010. The new Home Secretary, Theresa May, announced that Labour’s ‘single confidence measure’ was to be rescinded forthwith and replaced with another single top down target – ‘to reduce crime’ (Greenwood, 2010).

This announcement was followed with the abolition of APACS later that same year (Home Office, 2010a). The Government stated this would “...signal a fresh start with the police service with regards to policing performance...” and pave the way for “...new arrangements to be developed that best meet the aims of strengthening local accountability” (Home Office, 2010a, p.21). These new arrangements included

the abolition of Police Authorities and the introduction of directly-elected Police and Crime Commissioners (PCCs), who would be responsible for setting priorities and holding Chief Constables to account for performance (Home Office, 2010a; 2011).

The first PCC elections took place in November 2012, placing the newly-elected Commissioners in a position of far-reaching control over their respective police forces. Each PCC was charged with producing a Police and Crime Plan which set out their priorities and details of how performance would be measured (Home Office, 2011). The Home Office's position was therefore very much that it had removed "...unnecessary interference by central Government..." (Home Office, 2010, p.4) and 'passed the baton' to PCCs, who would now determine the shape of police performance management locally.

Following the advent of PCCs, police forces experienced a great degree of variance in how performance management frameworks were devised and applied (Rix, 2013). Some PCCs opted to avoid performance targets completely; for example in Surrey, the PCC, Kevin Hurley, stated explicitly in his Police and Crime Plan:

"I do not believe in micromanagement or setting a raft of targets which could skew police activity towards chasing numbers rather than doing the right thing for the public" (Surrey OPCC, 2013, p.13).

Conversely, other PCCs introduced precisely the 'raft of targets' that Hurley and others decried. For instance, in Thames Valley, PCC Anthony Stansfield's Police and Crime Plan contained numerous specific numerical targets, such as:

"Disrupt 20 problem and organised crime groups ..." (Thames Valley OPCC, 2013, p.41).

Therefore, whilst the Government pursued the localism agenda in affording PCCs the latitude to determine the style and intensity of performance frameworks for individual police forces, the result has been a disparate landscape which still reflects many of the features of performance management that it has ostensibly distanced itself from during recent years.

### **2.1.3.3 UK Police Performance Management: Key Traits and Future Challenges**

Recent history has therefore seen a paradigmatic shift in UK police performance management, characterised by the transition from the *monitoring* of police performance to active performance *management* (De Maillard and Savage, 2012). Furthermore, Collier (2006), drawing on Sanderson (2001), observes:

“Research has shown that the development of performance indicators has been primarily top-down with a dominant concern for enhancing control and upwards accountability rather than promoting learning and improvement” (2006, p.165).

This appetite for tighter centralised control has been increasingly evident throughout this period; nationally-determined priorities, targets and reporting mechanisms have combined to create a distinctive performance management approach that has become deeply embedded in the police psyche (McDermott, 2013). Indeed, De Maillard and Savage (2012) contend:

“Britain has developed the most elaborate framework of police performance management in Europe and one of the most elaborate frameworks of managerialism in the world” (2012, p.372).

Consequently, despite the Government’s stated intentions to move away from centralised performance management, the results are incomplete and inconsistent. The arrival of PCCs, has in some cases, reverted police forces to circumstances reminiscent of the height of the New Labour era. Furthermore, as the President of the Police Superintendents Association of England of Wales, Irene Curtis, observes, targets remain so entrenched in police performance management culture, that despite Government exhortations, it is not easy to simply ‘switch them off’ (Curtis, 2013).

Additionally, HMIC ensures there remains an emphasis on comparative performance through the use of ‘league tables’, perpetuating a dominant police performance management climate that retains much of the character of the NPM-oriented practices favoured during previous years. Furthermore, due to HMIC’s insistence that police effectiveness equates to crime reduction (HMIC, 2014, p4), there remains a heavy focus on crime data as a prominent measure of performance.

This is despite disagreement about the degree to which policing activity affects crime rates. Loveday (2000), for example, drawing on Bayley (1994), argues the police service has limited influence over crime due to exogenous factors, such as economic cycles, social deprivation, unemployment and substance abuse (see also Home Office, 1990; 1997). Similarly, Coleman and Moynihan (1996) assert it is “...misleading to imply that the crime rate is so clearly within the power of the Home Secretary and criminal justice system to control” (1996, p.134), whilst Smith (2006) argues recorded crime is unlikely to reflect actual crime anyway.

Nevertheless, it seems that the use of crime data as a *source of information* could be advantageous to police managers seeking to understand crime patterns, as this can provide an insight into crime trends and inform decisions about priorities or resource deployment. The distinction here is between the use of crime statistics *to inform decision-making*, versus its use as a definitive *measure of performance*. However, there remains uncertainty about how to define and measure police performance at all (Shane, 2010). The Home Office suggests ‘good’ police performance is simply:

“A combination of doing the right things (priorities), doing them well (quality) and doing the right amount (quantity)” (Home Office, 2008b, p.4).

Whilst this definition provides general guidance, it does not specify what to measure, or how to measure it. This has fuelled an ongoing search for appropriate indicators that could provide insights into relevant policing domains, such as reliability, responsiveness, competence and fairness (Mastrofski, 1999; Davis, 2012a). However, despite exhortations for a more insightful range of police performance measures, there remains an emphasis on easy-to-measure quantitative outputs (Ammons and Rivenbark, 2008). Writing in 1956, Ridgway observed:

“Today, there is a strong tendency to state numerically as many as possible of the variables with which management must deal” (Ridgway, 1956, p.241).

This tendency persists in policing to the present day. In addition to heavy reliance on easy-to-obtain crime data, it is suggested that a key reason for its preponderance is because it can be problematic to accurately define and measure public sector performance (Rainey *et al*, 1976; Rainey, 1993; Smith 1995; Pollitt, 1999; Caers *et al*, 2006; Kelman and Friedman, 2009; Moynihan *et al*, 2009); this is because

performance information is by its very nature inexact, subjective and incomplete (Moynihan, 2008) and "...public sector goals are ambiguous, multiple, complex and frequently in conflict with one another" (Jackson, 2011, p.15).

Furthermore, it is not easy to measure quality (Likierman, 1993) or true *outcomes*, such as harm reduction (Donahue, 1989; Eisenhardt, 1989; Behn and Kant 1999; Mackenzie and Hamilton-Smith, 2010). It can also be difficult to identify causal links between a particular intervention and an eventual outcome (Pollitt, 1999). Additionally, even when an outcome can be clearly defined and causal links established, results will often vary depending on the timing of evaluations (Heinrich, 2008). Therefore, complexities associated with measuring outcomes means proxy measures are often favoured as surrogate indicators of performance (Lowe, 2013).

Although such proxy indicators (or 'tracer conditions' - see Smith *et al*, 1997) are imperfect, they are often chosen because of their simplicity and heuristic value (Schalock 2001; Moynihan and Ingraham, 2004; Liu *et al*, 2010; Lowe, 2013). However, whilst this pseudo currency of inputs and outputs may be considered sufficient by some, Collier (2006) contends that the complex nature of policing does not lend itself to simple numeric measures. Skolnick (1966) too, observes, "The goals of police and the standards by which the policeman's work is to be evaluated are ambiguous..." (1966, p.164).

These longstanding considerations have led to the evolution of police performance measurement systems founded on simple metrics to assess performance in complex operating environments. The result is a landscape where "...inexact and incomplete measures are still widely used in performance management systems..." (Heinrich and Marschke, 2010, p.184). Consequently, despite a strong mandate within the literature for multifaceted, contextualised performance information, history has shown that UK police performance measurement has gravitated towards crude indicators and proxy measures; indeed, the *status quo* has been characterised as "...only having switches when it needs dials" (Hales, 2018).

For these reasons, in 2013, the College of Policing, in conjunction with the Association of Chief Police Officers (ACPO), the Home Office, the Association of Police and Crime Commissioners (APCC) and other key stakeholders, instigated a national Commission into Police Performance Management (College of Policing,

2013). The Commission's remit was to investigate the current state of police performance management and the behaviours it generates, aiming to learn the lessons of the past and build effective frameworks for the future.

In particular, the Commission sought to promote more insightful use of performance measures, along with a move "... away from the deeply embedded target culture in policing" (College of Policing, 2013, p.5). This mandate was bolstered in 2015, when the Home Secretary appointed Chief Superintendent Curtis to chair an Independent Review of Policing Targets (May, 2015). The Review was published later that year, providing an insight into current practices; in addition to confirming the use of numerical targets was widespread, it also found police performance information as a whole tended to be presented using simplistic numerical formats, the most prevalent being:

1. Binary comparisons.
2. League tables.
3. Numerical targets.

(Home Office, 2015).

However, although the Review produced clear recommendations and provided practical guidance on effective performance measurement, the College of Policing made no commitment to implement any of its recommendations, merely releasing a brief statement of acknowledgement following its publication (College of Policing, 2015). Five years later, although the College continues to publish and promote national Authorised Professional Practice (APP) affecting almost every conceivable aspect of policing, there is still no APP on performance management, performance measurement, or the use of performance information (College of Policing, 2020).

Therefore, whilst there may now be general acknowledgement that some practices have produced unintended consequences, the challenge in promoting a more sophisticated approach to police performance measurement lacks coherence at the national level and remains unresolved. This is despite the clear message conveyed by the Review that improving the *status quo* will involve more than merely criticising targets and requires much deeper appreciation of how performance information is presented and used.

Furthermore, in January 2020, the Prime Minister, Boris Johnson, announced he intends to introduce a target to reduce violent crime by 20% (Forsyth, 2020); one month later, the Home Secretary, Priti Patel, also signalled a return to national policing targets (Shaw, 2020). In response, Bill Skelly, Chief Constable of Lincolnshire Police warned such a move was likely to create ‘unintended or perverse consequences’ (Harper, 2020); it therefore remains to be seen whether the police service is on the cusp of repeating the experience of the last 35 years.

In summary therefore, binary comparisons, league tables and numerical targets remain the most prevalent performance information formats within UK policing and are likely to remain so for some time. Therefore, this research focuses on these formats, to try and establish why they seem to exhibit such a close relationship with behavioural dysfunction; the following section examines their design characteristics and provides examples of the types of behaviour they have become associated with.

#### **2.1.4 Binary Comparisons**

A prominent feature of UK police performance measurement involves the use of *binary comparisons*. This practice may be defined as the act of comparing two isolated numeric values, then interpreting any difference between them as though it were a trend or trajectory. Comparisons are routinely made with the previous week, month, quarter, average, cumulative year-to-date total, and so on. The difference between the two values (or a percentage change) is subsequently assumed to reflect a trajectory, which is then used to identify areas of concern, inform decision-making, judge performance, and as the basis for action.

The following examples are typical:

“Crime has been reduced by 10% across the force area. This is 3,963 fewer victims of crime compared to the previous 12 months” (North Yorkshire Police, 2013).

“In April 2014 the number of priority crimes committed fell to 26,297 offences, down 10% from last month and down 11% from the same month last year. Tramlink saw the largest fall (down 23%). The rate of victims who felt they were treated fairly by the police rose to 91% (up 3%) whilst victim satisfaction rose to 80% (up 4%)” (Greater London Authority, 2014).

“The number of house burglaries reported to police fell by 36 per cent during Operation Brightshadow, which ran from Monday, October 3 to Sunday, October 9, compared to the week before” (Essex Police, 2013).

Advocates of binary comparisons often describe them as a ‘starting point for asking questions’, or ‘management information’ (see Scottish Police Authority, 2014). This conveys the impression binary comparisons are a relatively benign form of data display, where limitations are known and routinely considered. However, despite their popularity and perceived heuristic benefits, they attract criticism for being simplistic and potentially misleading. An example is provided at Figure 2.1.

*Figure 2.1: Durham Police – Antisocial behaviour (ASB) table*

	12 months to end March 2017	12 months to end March 2018	% change
<b>ASB incidents</b>	22,134	19,591	-11%
<b>- ASB incidents (Durham CSP*)</b>	17,515	15,368	-12%
<b>- ASB incidents (Darlington CSP*)</b>	4619	4223	-9%

\*Community Safety Partnership

(Durham Police and Crime Commissioner, 2018, p.3)

This example relates to antisocial behaviour metrics; it contains raw data, as well as a positive or negative percentage change, compared to the previous year. Similarly, the example at Figure 2.2 uses binary comparisons to depict year-to-date metrics, using the same period last year and the five year average as reference points, and displaying a positive or negative percentage variance to indicate the direction of perceived trajectories:

Figure 2.2: Strathclyde Police - Violent crime table

**GROUP 1: CRIMES OF VIOLENCE 1<sup>st</sup> April 2012 to 31<sup>st</sup> July 2012**

		YTD 2011/12	YTD 2012/13	Change
<b>Group 1 :Crimes of Violence</b>		<b>1,920</b>	<b>1,401</b>	<b>-27.0%</b>
Murder		14	16	14.3%
Attempted Murder		64	49	-23.4%
Culpable Homicide (common law)		0	1	-
Culpable Homicide (other)		3	3	0.0%
Serious Assault		980	660	-32.7%
Robbery and assault with intent to rob		440	284	-35.5%
<b>Violence, Disorder and Antisocial Behaviour 01/04/2012 – 31/07/2012</b>				
<b>Violence</b>		<b>YTD 2011/12</b>	<b>YTD 2012/13</b>	<b>% Change</b>
<b>1</b>	<b>Level of violent crime (Group 1 only)</b>	<b>1,920</b>	<b>1,401</b>	<b>-27.0%</b>
<ul style="list-style-type: none"> <li>The number of crimes of violence fell by 27.0% compared to the previous year and 43.2% against the 5 year average.</li> </ul>				
<b>2</b>	<b>Number of murders/attempt murders</b>	<b>78</b>	<b>65</b>	<b>-16.7%</b>
<ul style="list-style-type: none"> <li>The Force has seen an increase of 14.3% in the number of murders (a numerical value of 2) in comparison with the same period last year. Conversely, there has been a 23.1% reduction on the 5 year average.</li> <li>The number of attempt murders has decreased by 23.4% (a numerical value of 15) in comparison with the same period last year and 55.0% on the 5 year average.</li> </ul>				

(Strathclyde Police, 2012, p.1)

It can be observed that murders appear to be ‘increasing’ (by 14.3%) compared to the same period last year, yet ‘decreasing’ (by 23.1%) compared to the five year average. This prompts questions about the stability of such descriptors, as well as which (if either) comparison most accurately reflects the murder rate. Such conflict is a typical symptom of the unstable nature of binary comparisons: depending on which reference point is chosen, perceived trajectories can be diametrically opposed.

The use of binary comparisons as a method of presenting numeric data is widespread, both within the police service and amongst bodies charged with inspecting or reporting upon police performance (see HMIC, 2012; ONS, 2013; Staffordshire OPCC, 2013). In the experience of the author, substance tends to be ascribed to any apparent differences and the consequent expectation is that they must be accounted for and action taken to redress the balance. This occurs even when it is not within the gift of those under scrutiny to significantly influence a particular measure (as in the case of crime rates).

Conversely, assumptions can be drawn that suggest improvements have occurred, which invites complacency and even triumphalism. Take the following quote from one Police and Crime Commissioner's annual report:

“Recorded crime fell by 8% between April 2012 to March 2013. This represents 3,216 fewer victims of crime compared to the same period in 2011/12. This latest fall means that North Yorkshire has overtaken Norfolk to become the safest policing area in England” (North Yorkshire OPCC, 2013, p.2).

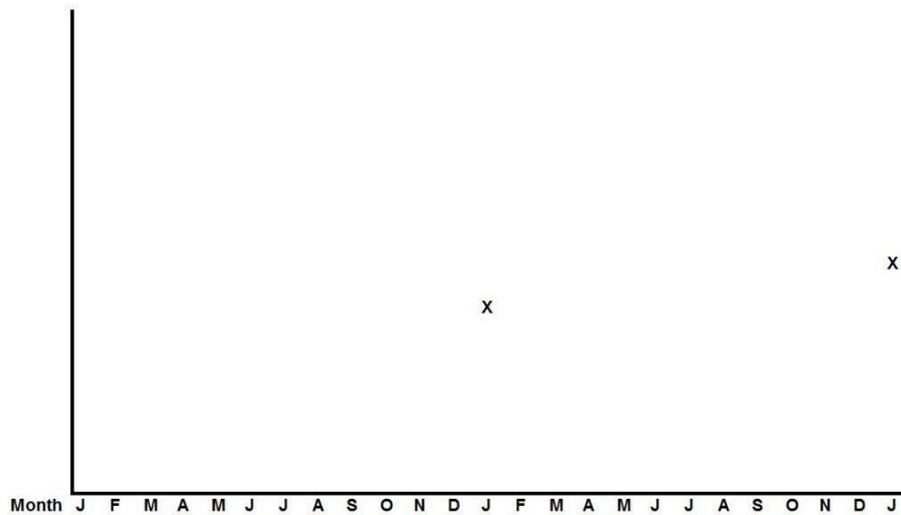
Binary comparisons tend to be used because of their simplicity. They appear to convey headlines about ‘direction of travel’, and make for easy-to-consume performance information. However, Bird *et al* (2005) warn their use tends to encourage over-interpretation of very minor changes amidst the data, leading to false conclusions:

“Very particularly the practice of concentrating on a comparison with the most recent value, this year's results compared with last year's, may be very misleading for reasons including regression to the mean” (2005, p.14).

Similarly, Rothstein (2008), Jackson (2011) and Kahneman (2011) caution against assuming performance has either improved or deteriorated, as an apparent change may simply reflect random statistical variation. In order to ensure that such statistical considerations are taken into account, Bird *et al* (2005) promote the use of time series data as an alternative; this affords a more contextualised picture than is possible when using binary comparisons.

The superficial nature of binary comparisons becomes clear when one is displayed visually, as in Figure 2.3:

Figure 2.3: Binary comparison chart<sup>2</sup>



Presenting a binary comparison in this manner exposes the gaps in the data set and highlights the lack of context. It is also apparent that if the earlier data point happens to be abnormally high or low, the risk of envisioning a steeper trajectory between the two values is increased. The practice gives the impression of trends and trajectories that simply might not exist, and makes it impossible to draw any valid conclusions. These are critical deficiencies; as Wheeler (2000) emphasises:

“Data have no meaning apart from their context” (2000, p.12).

#### 2.1.4.1 Binary Comparisons and Behavioural Change

As a primary objective of performance information is to aid decision-making, the use of binary comparisons could well affect the choices made as a result of engagement with data presented in this format. This is because negative variance may give the impression of poor performance, which could lead to managers exhorting subordinates to address perceived deficiencies, thereby potentially inducing behavioural changes.

Furthermore, the use of red and green numbering or shading to highlight perceptions of good or poor performance, along with arrows emphasising a perceived direction of travel and descriptors that ascribe blame to ‘failing’ work units, could potentially heighten these risks. The below excerpts from police performance documents are indicative:

---

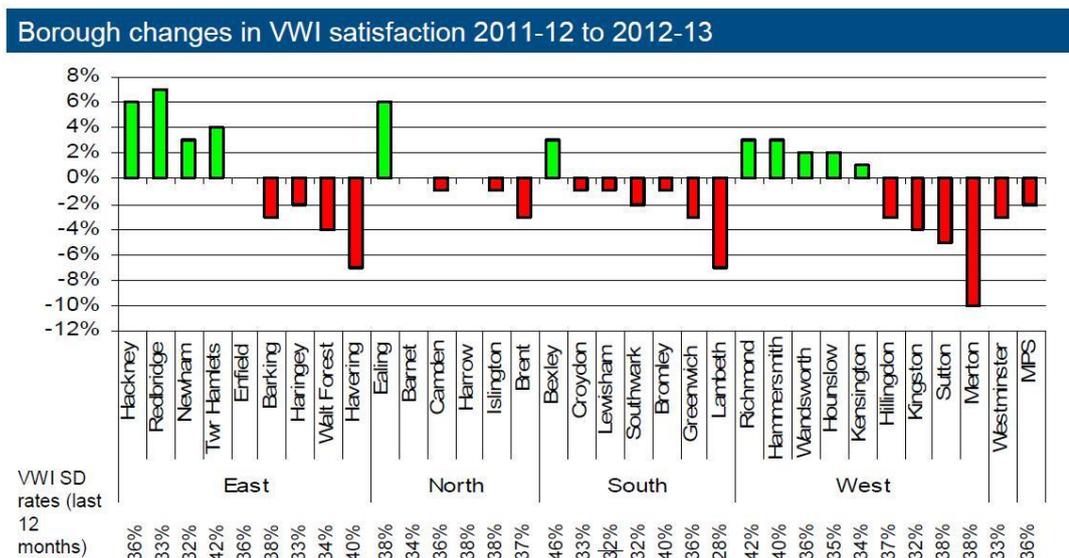
<sup>2</sup> Note: numeric values have been omitted from the vertical axis for simplicity of presentation.

Figure 2.4: Sanction detection rate (i.e. solved crimes) – Current rolling twelve months vs previous rolling twelve months

SD Rate		Primary SD Rate	
Previous R12	Current R12	Previous R12	Current R12
21.3%	21.9%	20.4%	20.7%
10.4%	11.5%	6.3%	6.0%
9.9%	12.3%	5.2%	5.5%
11.5%	9.9%	8.4%	7.1%
13.1%	14.8%	13.0%	14.8%
15.8%	15.2%	15.2%	14.2%
14.8%	14.5%	14.3%	13.4%
27.8%	25.5%	26.6%	24.6%
5.1%	6.2%	3.0%	3.3%
4.3%	5.8%	1.8%	2.3%
7.5%	7.2%	6.5%	6.6%
2.7%	2.7%	2.4%	2.2%
31.4%	35.8%	31.4%	35.8%
25.2%	28.8%	25.2%	28.8%
46.4%	50.7%	46.4%	50.7%
43.5%	46.3%	43.5%	46.3%

(Metropolitan Police, 2013a, p.6)

Figure 2.5: Violence with Injury (VWI) satisfaction rates



11 Boroughs improved their satisfaction scores for VWI last year.  
 2 unchanged  
 19 Boroughs' performance in VWI victim satisfaction have declined in the last year

(Metropolitan Police, 2013a, p.30)

Figure 2.6: Missing people - performance data

Performance Statistics - Last Month			
	Last : Month <input type="button" value="v"/>		
	This Month	Projected Comparison	Last Month
Missing Reports	▲ 47	▲ +14.42	▲ 107
Individuals	▲ 43	▲ +24.08	▲ 87
U18 Missing Reports	▼ 26	▼ -5.83	▼ 73
U18 Missing Individuals	▲ 22	▲ +3.83	▲ 53
Avg. Reports per day	▲ 3.92		▲ 3.57
Avg. Time Missing (hrs.)	▲ 9.02		▲ 11.68
Repeat Mispers	▼ 2	▼ -3.83	▼ 9
# Reports from Repeats	▼ 6	▼ -13.5	▼ 29
In-Care Individuals	▲ 11	▲ +6.42	▲ 22
In-Care Reports	▼ 15	▼ -3.25	▼ 42
Foster Care Individuals	▲ 3	▲ +4.75	▲ -
Foster Care Reports	▲ 3	▲ +3.75	▲ -
Hospital Reports	▲ 1	▲ +1.58	▲ 1
Repeat Rate	▲ 12.77%		▲ 27%
Person : Report Ratio	▲ 1 : 1.09		▲ 1 : 1.23
# Found Harmed	▲ 6	▲ +10.5	▲ 5
# Found Dead	▼ 0	▼ -1	▼ 1

(Staffordshire Police, 2019)

All of the above examples incorporate the use of colour to reinforce apparent trajectories. One also incorporates narrative stating boroughs' performance has either 'improved' or 'declined'. Another uses red or green arrows to ascribe apparent improvement or deterioration across a range of indicators relating to missing people (a complex area affected by multiple exogenous factors).

Such use of colour coding to draw attention to particular figures is commonplace in policing and appears to be a deliberate technique for initiating remedial action. Officers consistently describe being held to account on this basis; often such performance information is displayed publicly. In the words of one officer:

“We have a performance board in our station. It shows ‘up’ and ‘down’ arrows, in red (bad) and green (good) and % figures” (Guilfoyle, 2013, p.153).

When binary comparisons are augmented by such symbolism, it reinforces perceived trajectories. Unless performance information users are aware of limitations, the way the data are displayed could influence their perceptions about performance; this in turn seems capable of potentially affecting decision-making and action, in a similar fashion to that associated with the use of league tables and numerical targets.

#### **2.1.4.2 Binary Comparisons: A Summary**

Binary comparisons are a ubiquitous feature of UK police performance measurement systems, yet there is scant literature and a dearth of research on the topic. However, despite their popularity, there are well-founded concerns regarding their unstable nature. Primarily, their dependence on an isolated temporal reference point makes them susceptible to depicting trends or trajectories which may not be reflective of the longer term picture. Consequently, users may be lured into making erroneous assumptions, which could ultimately lead to unnecessary behavioural responses.

Therefore, even where managers seek to use binary comparisons in a careful manner, it seems that they are simply not a valid ‘starting point for asking questions’ at all – if their simplistic visual appearance substantially impairs data interpretation, binary comparisons could be considered so misleading as to render them counterproductive in the extreme. This study will therefore assess whether presenting data as binary comparisons does indeed affect assumptions about performance and influence subsequent choices; centrally, it also seeks to establish whether the visual appearance of this format may be an antecedent to behavioural dysfunction.

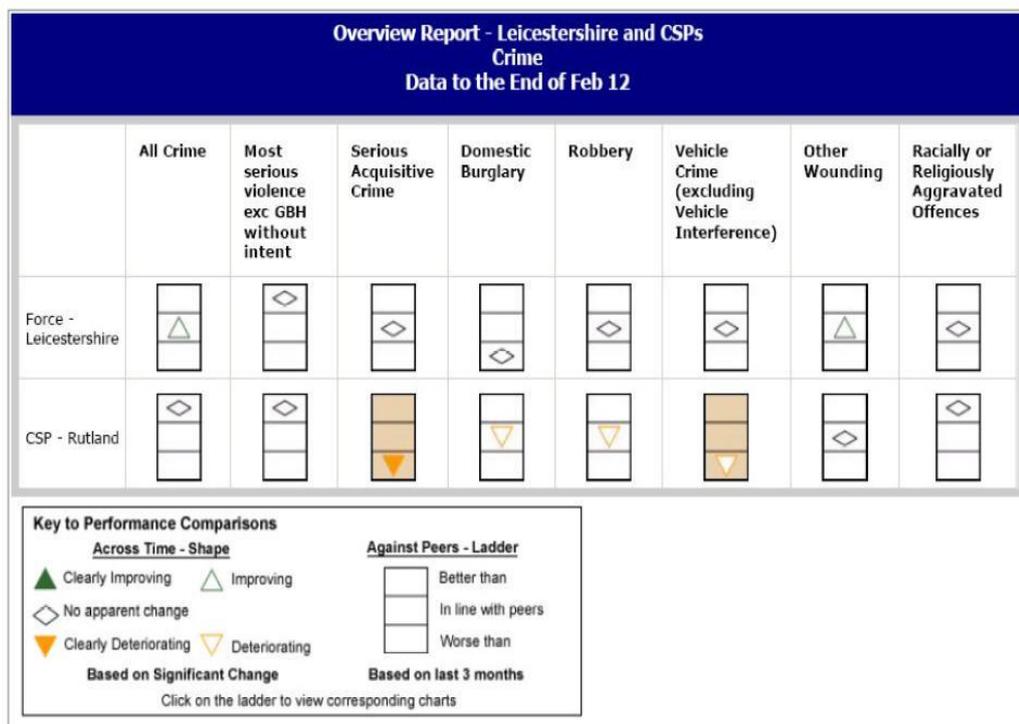
#### **2.1.5 League Tables**

Whilst potential behavioural changes associated with binary comparisons may be largely unintentional, in contrast, the use of ‘league tables’ (also known as tabular peer comparisons) is explicitly intended to influence behaviour. The practice can be traced as far back as the 18<sup>th</sup> century, when social reformer Jeremy Bentham called for what he termed the ‘tabular comparison principle’ in government accounting (Hume, 1981; Hood, 2007). League tables, ranking systems and the general practice of comparing peers for the purpose of assessing performance have subsequently taken hold throughout public services, becoming particularly well-established during the NPM reforms.

In UK policing, league tables are used nationally and locally; forces, departments, teams, and even individuals are ranked and compared against each other. Proponents argue league tables provide clarity and accountability, are useful for benchmarking performance and signal quality of service (Hatry, 1999b; Courty *et al*, 2007; Jackson, 2011). The use of league tables in the education and healthcare sectors is particularly well-established (Goldstein and Spiegelhalter, 1996; Bevan and Hood, 2006; Rothstein, 2008), but the practice is also deeply embedded in policing (Home Office, 1993c, pp.18-19).

At the national level, HMIC ranks forces according to their performance in areas such as crime rates, public satisfaction and data quality. These assessments are then disseminated through a performance measurement tool called iQuanta. The iQuanta tool presents graphical and tabular performance information, using symbols to illustrate relative performance of individual forces (or areas within forces) and perceived trajectories; comparisons are made against peers, as well as previous years' crime data (Home Office, 2010b). An example is provided at Figure 2.7.

Figure 2.7: *iQuanta Snapshot – Leicestershire Police*



(Rutland CSP, 2012, p.1)

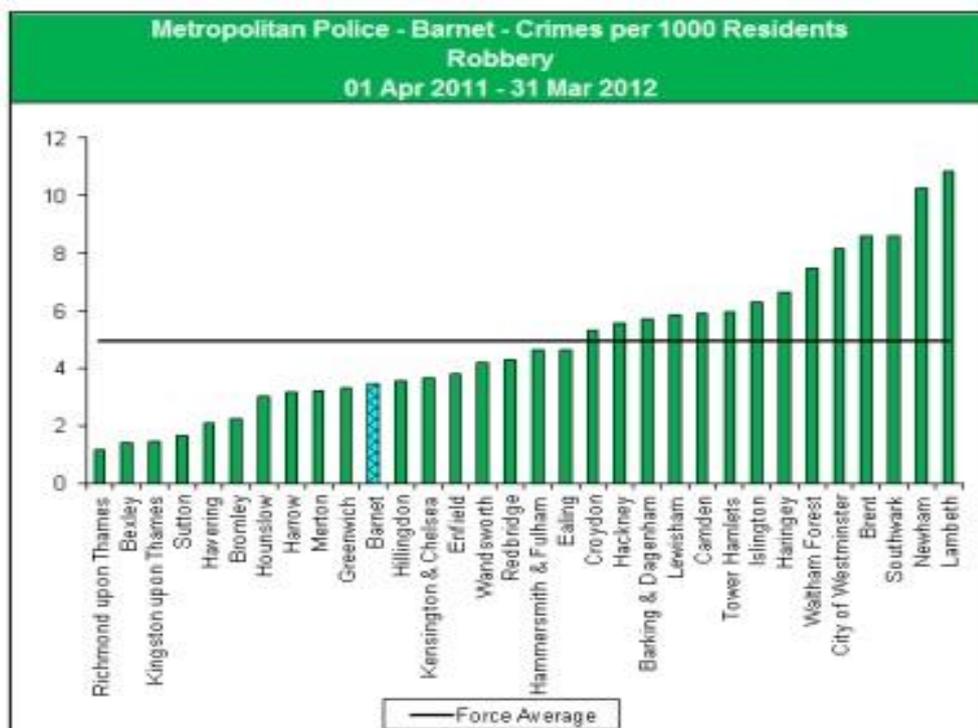
Apart from individual force ‘scorecards’ like the example above, data are often used to construct numeric tables or comparative bar charts that directly rank forces against each other. A range of metrics can determine the rankings, such as detection rates (i.e. solved crimes), speed of response to emergency calls, number of complaints, sickness levels and so on. Examples are provided as follows.

Figure 2.8: Metropolitan Police crime data dashboard

Borough Name	Crime Count
Westminster	118,809
Camden	74,362
Lambeth	73,171
Newham	71,944
Southwark	71,250
Tower Hamlets	67,382
Hackney	64,299
Croydon	63,640
Haringey	63,025
Islington	62,635
Brent	61,763
Ealing	58,468
Barnet	56,181
Wandsworth	53,269

(Metropolitan Police, 2018)

Figure 2.9: iQuanta - Robbery Comparative Bar Chart



(Home Office, 2012b, p.2)

Figure 2.10: Police Forces Comparison Table

Force	% of recorded crimes detected	Detection ranking	No of complaints per 1000 officers	Complaints rank	% 999 calls answered within target time	999 calls ranking	Average no of sickness days per officer	Sickness rank	% of immediate response incidents where target time met	Immediate response ranking
Dyfed Powys	63	1	151	3	90.5	14	11.5	14	82.7	31
Gwent	57	2	385	42	93.5	7	15.4	42	96.7	1
Suffolk	35	3	200	15	93.7	6	10.8	6	92.9	8
Devon and Cornwall	34	4	248	29	84.5	33	11.3	10	82.6	32
Cumbria	34	4	268	32	92	9	13.1	29	91.8	12
Durham	34	4	118	2	98.7	1	13	25	92.1	10
Northamptonshire	33	7	198	14	89.2	21	9.8	3	87.6	21
South Wales	32	8	255	30	85.8	30	16.1	43	87.6	21
North Wales	31	9	212	18	85.8	11	11.8	17	91.6	16
Gloucestershire	31	9	232	25	94.6	4	10.5	5	91.6	13
Northumbria	31	9	174	8	79.5	40	8.8	2	92.2	10
Wiltshire	30	12	238	27	88.4	24	11.4	12	88.3	20
North Yorkshire	30	12	344	39	84.8	31	13	26	92.2	9
Cheshire	30	12	246	28	88	29	13.7	38	77.8	41
Hampshire	29	15	153	4	73.6	42	12.2	21	94.6	3
Surrey	28	16	258	31	86.2	28	11.8	17	78.3	40
Kent	28	16	194	12	86.2	33	11.6	15	94	4
West Midlands	28	16	262	33	97.5	2	13	25	87	23
Leicestershire	28	16	167	6	90.1	16	10.9	7	82.3	34
Merseyside	28	16	233	26	88.6	23	13.3	31	88.6	19
London, City of	27	21	82	1	83.6	37	14.9	41	95.3	2
Bedfordshire	27	21	189	11	84.7	32	12.9	24	84.6	26
West Mercia	27	21	217	20	92.3	8	13.6	36	87	24
Lancashire	27	21	215	19	86.3	27	12.5	23	93.7	5
Norfolk	26	25	264	34	91.4	12	11.8	17	83.4	29
Essex	26	25	196	13	91.8	10	13.6	36	81.1	36
Derbyshire	26	25	168	7	94.5	5	11.6	15	84.1	27
Dorset	25	28	204	16	79.6	39	10	4	79.8	37
Lincolnshire	25	28	281	37	n/a	n/a	11.4	12	81.8	35
South Yorkshire	25	28	184	9	90	17	12.2	21	79.7	38
Hertfordshire	24	31	278	36	84.5	33	11.8	17	83.2	30
Cambridgeshire	24	31	227	21	88.8	22	13.3	31	82.3	33
Sussex	23	33	643	43	90	17	11.3	10	n/a	n/a
Staffordshire	23	33	370	41	82.9	38	13.3	31	89.6	17
West Yorkshire	23	33	165	5	86.4	26	13	25	89	18
Avon and Somerset	22	36	288	38	90	17	13.4	34	89.8	15
Thames Valley	22	36	205	17	88.2	25	14.3	39	79.8	38
Warwickshire	22	36	228	23	91.2	13	13.5	35	93.4	6
Greater Manchester	22	36	266	35	89.6	20	14.8	40	83.6	28
Humberside	21	40	188	10	96.2	3	8.6	1	86.2	25
Cleveland	21	40	227	21	79.2	41	13.2	30	93.1	7
Nottinghamshire	20	42	349	40	90.3	15	11.2	9	91.5	14
Metropolitan Police	15	43	229	24	83.6	36	10.9	7	76.4	42

(Observer, 2011)

Although all-encompassing national rankings are made, HMIC tend to place individual forces into ‘most similar force’ groups (known as MSFs, or MSGs), based on demographic and socio-economic factors, plus crime data (Hale *et al*, 2005, p.5; HMIC, 2013a; 2013b; 2013c). This is done with the intention of promoting “...fairer and more meaningful comparisons between areas” (HMIC, 2013b, p.1). Under this approach, conclusions about performance effectiveness are made based on each force’s position in the rankings. The rationale behind this method is:

“...not to set one individual force against another nor to use the performance indicators as a proxy for increasing inter-force competition, but rather to broaden the scope of management tools open to senior officers in the context of improving individual force performance” (Barton and Beynon, 2011, p.363).

Some commentators contend that comparing police agencies in this way is a useful method for assessing performance; indeed, ranking and grading in league tables could aid transparency and accountability (Shane, 2010). Others, however, contend there are inherent deficiencies associated with league tables that undermine their efficacy, which can lead to unwarranted conclusions being drawn and even invite unwanted behaviour. Hood (2007), for example, argues:

“Like target systems, they are likely to produce output distortions as producers learn to find ways that move their organizations up the league tables in ways that do not reflect the intentions of those who framed the rankings, or ignore non-measured activities” (2007, p.101).

Jacobs and Goddard (2007) observe that whilst composite measures (i.e. those aggregated to a single indicator) are often used to construct league tables because of their simplicity, they are unable to provide a clear picture of relative performance. McLean *et al* (2007) also conclude that aggregate measures are an insufficiently robust means of assessing performance, whilst Bird *et al* (2005) warn the approach can result in misinterpretation and even ‘severe distortion’ (2005, p.14).

Furthermore, league tables do not communicate history or context; trends, statistical anomalies or patterns therefore remain obfuscated, effectively constraining meaningful interpretation. In particular, Bird *et al* (2005) are emphatic regarding the

dangers of relying upon a ‘snapshot’ of time that is, by definition, characteristic of how data are displayed in league tables. Similarly, Shilston (2008) discourages attempts at reducing “...the complex interactions involved in service delivery to a number” (2008, p.362), whilst Wheeler (2000) warns single figure metrics lack context, making it difficult to interpret the data or assimilate information.

Jacobs and Goddard (2007) and Hood *et al* (2009) highlight further inherent instabilities and limitations, such as high degrees of uncertainty when attempting to determine the order of rankings. Likewise, Jackson (2011) warns:

“League table rankings are very unstable and are sensitive to small changes in the variables that compose them” (2011, p.20).

In one case, a ‘small change’ in a league table for local authorities led to a drastic impact on the order of ranking: “The largest jump in position for an individual authority was 54 places, more than half the league table” (Jacobs and Goddard, 2007, p.107). Similarly, Goldstein and Spiegelhalter (1996) cite a study by Green and Wintfield (1995) which examined league tables for cardiac surgeons, where “...46% of the surgeons had moved from one half of the ranked list to the other” in just one year (1996, p.403). Consequently, Jackson (2011) cautions against drawing conclusions about performance by interpreting league table rankings, arguing:

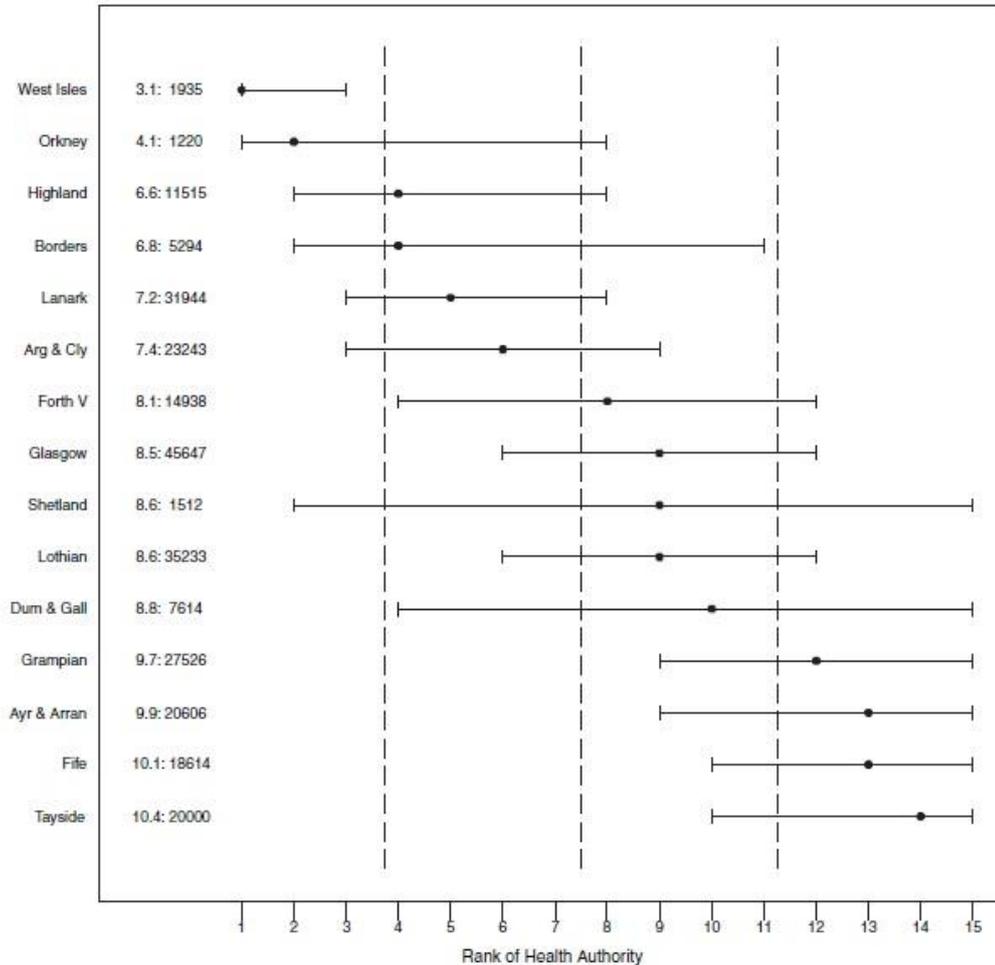
“To hold managers accountable for the position of their organization in a league table, or to reward them on this basis, is perverse” (2011, p.21).

Furthermore, Goldstein and Spiegelhalter (1996), Bird *et al* (2005) and Jackson (2011) draw attention to the critical importance of transparency regarding confidence intervals, warning of the dangers of failing to place caveats alongside rankings. Such caveats should provide clarity regarding the methodology used in constructing the table, as well as statements relating to limitations and overall context. Jacobs and Goddard (2007) insist such indications of uncertainty should always be published to “...communicate the sensitivity of the reported measure” (2007, p.109).

Despite this consensus, it is uncommon for such caveats to be provided alongside league tables and where there are high degrees of uncertainty or broad confidence intervals, this makes it impractical to separate those being ranked; furthermore, ranking is particularly difficult when differences involve small numbers. Goldstein

and Spiegelhalter (1996) contend “...even after adjustment, finely graded comparisons between institutions are impossible” (1996, p.397). In Figure 2.11 it can be seen that broad intervals result in overlaps between institutions, making neat separation problematic.

Figure 2.11: League Table – Teenage Conception Rates



(Bird *et al*, 2005, p.17; adapted from Goldstein and Spiegelhalter, 1996, p.401)

Figure 2.11 displays the median and 95% intervals for ranks of Scottish health authorities with regard to teenage conception rates between 1990 and 1992 inclusive: rates and relevant populations are shown for each health authority. By way of commentary, Goldstein and Spiegelhalter (1996) offer the following observations:

“...the medians do not always match the observed ranks. The width of the intervals is notable: in fact the firmest conclusions that can be drawn are that the Western Isles is in the lower quarter, Highland and Lanark are in the lower half and four health boards are in the top half” (1996, p.400).

Therefore, where explanatory narrative or statistical caveats are absent, it often means no useful conclusions can be drawn about performance, position, or apparent ‘direction of travel’ within a league table. In such circumstances, league tables are neither helpful in establishing how well an organisation is currently performing (Wilson and Piebalga, 2008), nor a good predictor of future performance (Leckie and Goldstein, 2009). Moreover, an organisation’s position within a league table does not necessarily correlate with its actual performance (Bevan and Hood, 2006; Jackson, 2011). Fundamentally, Goldstein and Spiegelhalter (1996) assert:

“...current official support for output league tables, even adjusted, is misplaced...” (1996, p.405).

### **2.1.5.1 League Tables and Behavioural Change**

Despite these well-documented limitations, league tables are widely used to influence performance. Unlike the relatively incidental factors associated with binary comparisons that may precipitate behavioural change, league tables are explicitly *intended* to act as a catalyst for altering behaviour. Their use often involves an implicit emphasis on ascribing failure and embarrassment in an attempt to motivate apparent underperformers. Le Grand (2010) describes the common tactic of using league tables to ‘name and shame’ institutions as:

“...the publicising of poor performance to peers, and / or to the general public, with the intention of humiliating the staff of the organisation concerned and hence encouraging them - in their own self-interest - to do better” (2010, p.61).

This underlying negative disposition characterises UK police league tables; perceived directions of travel or positions within them are often colour coded, with ‘red’ indicating failure. In the ‘most similar force’ (MSF) comparisons used in the UK, descriptors such as ‘clearly deteriorating’ accompany arrows indicating the apparent performance trajectory (see, for example, Bedfordshire Police, 2008, p.3). This can have the effect of stigmatising the forces towards the bottom of the table, or those whose performance appears to be getting worse.

In addition to the stigma of being associated with an underperforming institution, poor rankings have often been accompanied by severe personal sanctions. For

instance, following the introduction of the NHS ‘star ratings’ system in 2001 (Department of Health, 2001), which ranked institutions based on performance against targets, it was made known that those chief executives responsible for hospitals deemed to be failing would be at risk of losing their jobs (Hood, 2006; Propper *et al*, 2010). That same year, this is precisely what happened to six chief executives (Beverley and Haynes, 2005).

A similar situation has occurred in the education sector, where ‘failing’ institutions have been subject to direct Government intervention, often with the presumption of ‘leadership change’ following low rankings (Department for Business Innovation and Skills, 2011, p.28). However, Holloway *et al* (1999) argue such punitive approaches are unlikely to improve standards:

“...rather than providing a stimulus for improvement, highly critical OFSTED reports only serve to reinforce a cycle of decline” (1999, p. 199).

Bevan and Hamblin (2009) conclude the public dissemination of ranked performance data carries reputational damage to ‘poor performers’, thereby spurring them on to take action to avoid being toward the lower end of rankings in future. Indeed, they contend that the league table-based ‘star rating’ system in healthcare was explicitly “...designed to inflict reputational damage on hospitals performing poorly” (Bevan and Hamblin, 2009, p.167).

The same applies within policing, where league tables containing individual officers’ performance outputs are sometimes circulated to expose ‘poor performers’ (Metropolitan Police Federation, 2014). De Maillard and Savage (2012) observe:

“...the core rationale for police performance tables is more a reflection of a strategy of ‘naming and shaming’ - an attempt to drive up performance by identifying the best and worst performing police organisations” (2012, p.367).

This has had the effect of encouraging sensationalist headlines such as, “UK’s worst police forces named” (Daily Mail, 2006), thereby generating pressure to avoid low rankings in future. Whilst the intention of policy makers would be that organisations are encouraged to address deficiencies, experience suggests the approach doesn’t necessarily result in performance improvements (Olsen, 2012); furthermore, it risks

inducing gaming and other behavioural dysfunction (Bevan and Hood, 2006; Hood, 2007; Rothstein 2008).

Such adverse reactions are well-documented across the public sector. For example, in education, examples of perversity include ‘teaching to the test’ (Jacob, 2005; Hood, 2006; Heinrich and Marschke, 2010), preventing low-achieving students from entering exams (Rothstein, 2008) and teachers altering pupils’ test scores (Jacob and Levitt, 2003). In healthcare, Dawson *et al* (2005) discovered some hospitals focused on measured activities at the expense of other important functions. Marshall *et al* (2000), report that the publication of UK hospital mortality rates made some surgeons reluctant to operate on higher risk cases. Rothstein (2008), drawing on Altman (1990), comments on similar circumstances in the US:

“In 1989, St. Vincent's Hospital in New York City was put on probation by the state after it placed 24th in the ranking of state hospitals for cardiac surgery. The following year, it ranked first in the state. St. Vincent's accomplished this feat by refusing to operate on tougher cases” (2008, pp.41-42).

Overall, within organisations the practice of ranking teams against each other can lead to unhealthy competition (Brickman and Bulman, 1977; Tesser, 1988) and a condition known as *sub-optimisation* (Hardin, 1968; Neave, 1990, pp.232-240). This is where individual work units optimise their own performance outputs at the expense of others, leading to departmental or organisational cohesion becoming undermined and overall performance impaired (Ridgway, 1956); in some cases it can even lead to workers actively sabotaging the efforts of those perceived to be performing well (Hoffman *et al*, 1954).

Sub-optimisation is characterised by fractious relationships between departments units and behaviour designed to make internal performance look good, even if this disadvantages others (Seddon, 2003). Ultimately, this harms the overall system, distorting organisational objectives and impairing service delivery (Deming, 1986; 1994). These adverse outcomes are likely to occur in situations where peers are ranked, partly because, “...pitting agents against each other can reduce the incentive for agents to cooperate” (Heinrich and Marschke, 2010, p.194).

Whilst there is merit in understanding the reasons for differential performance between comparable work units, the drive should be for continuous improvement of the *whole*, rather than focusing on individual components (Deming, 1986; 1994). League tables frustrate these aspirations as they segregate work units and set them against each other. In policing, this has contributed towards an unhealthy performance management climate, where officers feel pressured into engaging in questionable behaviour (Cockcroft and Beattie, 2009), such as in the case below:

“We were recently told that our BCU [*Basic Command Unit – a geographical policing area*] PACE 1s [*stop and searches*] had fallen to an unacceptable level on our internal force league table and that we needed to stop-search more people” (Copperfield, 2012, p.192).

Clearly, there is an ethical dimension to be considered here, as searches may only occur when legally justifiable and if certain conditions apply (Home Office, 1984). Encouraging officers to conduct high volumes of searches in order to generate a favourable position in a league table does not meet the grounds required for the lawful and ethical use of stop and search powers, and could lead directly to improper conduct. Indeed, research into the use of stop and search powers in Scotland uncovered such questionable behaviour, such as instances of four and five year old children being stopped and searched by police (Murray, 2014).

This highlights the risk that pressures associated with the use of league tables could influence behaviour to the extent the *de facto* purpose for activity becomes to avoid low rankings. The danger is that conversations within police forces then gravitate towards, ‘How can we improve our position in the league table?’ rather than, ‘How do we provide the best possible service to the public?’

#### **2.1.5.2 League Tables: A Summary**

There are multiple examples of league tables being associated with instances of behavioural dysfunction and other unintended consequences. Furthermore, there is little dispute that the conduct of some managers has been responsible for causing dysfunction. The literature indicates that a degree of behavioural change is likely the result of pressure being exerted to avoid lower positions in league tables, suggesting that cases of adversity do indeed arise as a consequence of improper application.

However, there are also sufficient concerns about the design limitations of league tables, giving rise to the prospect their visual appearance could distort perceptions about performance. As with binary comparisons, it is conceivable that managers' assumptions about relative performance could influence their decision-making and possibly even contribute towards behavioural change. Therefore, even where managers seek to use league tables with care, it could be that there is a latent risk of invalid assumptions, which in turn may catalyse inappropriate responses.

This study will investigate this prospect further, aiming to establish whether adverse behavioural tendencies<sup>3</sup> are likely to be triggered when league tables are used, even in the absence of improper application or other aggravating factors. Specifically, it seeks to ascertain whether the way in which league tables display data is a traceable antecedent to behavioural dysfunction.

### **2.1.6 Numerical Targets**

By way of providing a clear definition, the type of targets this research focuses on are *numerical* targets, of the type most commonly used in UK policing; these tend to be manifested as an exhortation to achieve an aspirational numeric value, such as to 'increase 'X' by 20%' / 'produce 'A' number of outputs within 'B' timeframe' / 'attain a standard of 'C' in 90% of cases', and so on.

The use of targets in performance management systems has been the subject of much debate and there is extensive literature to consider. Therefore, the forthcoming section is more substantial than those exploring binary comparisons and league tables. It begins with a brief insight into the historical antecedents of performance targets, before discussing Goal Setting Theory, arguments for and against the use of targets, and implications for behavioural change.

#### **2.1.6.1 Numerical Targets: A Brief Timeline**

Targets have a long history within performance management systems, but they came to prominence through the work of US industrialist Frederick W. Taylor, whose 'scientific management' techniques of the early 20<sup>th</sup> century became synonymous

---

<sup>3</sup> For the purposes of this research, 'adverse behavioural tendencies' are defined as: "Recurrent forms of behaviour characterised by disproportionate, unwarranted, or perverse responses to police performance information".

with the notion of controlling work through the imposition of explicit standards. Taylor's methods involved measuring relatively simple inputs and outputs, such as the time taken to complete a unit of work, or the number of items produced, and then setting targets for production levels (Taylor, 1911; Hood, 2007).

Subsequently, target-driven performance management became widespread, being used in armaments production during both World Wars, tax administration and job placement programmes (Jowett and Rothwell, 1988; Hood, 2007; Hood and Dixon, 2010). Meanwhile, the Soviet 'Targets and Terror' regime became renowned for harsh penalties imposed for failure to attain quotas (Berliner, 1956; Rothstein 2008). During the late 20<sup>th</sup> century, the use of targets proliferated in both the public and private sectors, with the NPM reforms ensuring they were integral to performance management in schools, healthcare, government and law enforcement (Hood, 1991; 1996; Bevan and Hood, 2006).

#### **2.1.6.2 Numerical Targets and Goal Setting Theory**

In order to understand why the use of numerical targets is so widespread, it is helpful to explore the theoretical perspectives offered by *Goal Setting Theory*, as they provide insights into reasons why targets are considered by many to be a useful performance management tool. Although Goal Setting Theory is primarily associated with Edwin Locke and Gary Latham (Locke, 1968; Locke *et al*, 1981; 1989; Locke and Latham, 1984; 1990; 1991; 2002; 2006), English psychologist Cecil Alec Mace is credited as 'the man who discovered goal setting' (Phillips-Carson *et al*, 1994).

During the 1930s, Mace conducted experiments which suggested specific, clearly-defined goals increased his subjects' performance; he also found that tightly-defined objectives tended to lead to better performance increments than broad 'do your best' type goals (Mace, 1935). Locke and Latham have expanded upon this field of research over recent decades and offer their definition of a 'goal'<sup>4</sup> as follows:

“A goal is the object or aim of an action, for example, to attain a specific standard of proficiency, usually within a specified time limit” (2002, p.705).

---

<sup>4</sup> The term 'goal' in American English may be considered synonymous with 'target'.

Locke and Latham assert goal-directed behaviour results in improved performance and they are explicit about the advantages of ‘specific, difficult goals’ (Locke and Latham, 2002). Behn (2003) also argues targets provide a clear objective to be achieved; according to Moynihan and Pandey (2005) such clearly-articulated objectives counteract goal ambiguity. Separate research highlights generally positive effects, such as improved exam results (Ashforth and Anand, 2003; Boyne and Chen, 2006; Barsky, 2008; CIPD, 2016a).

Others claim challenging, specific goals act as a catalyst for stimulating innovation and action (Sitkin *et al*, 2011), guiding effort and persistence (Cyert and March, 1963), promoting flexible thinking (March, 1991), building collective enthusiasm (Bass and Riggio, 2005) and motivating high performance (Rousseau, 1997). In particular, some suggest ‘stretch goals’ (or ‘stretch targets’) (see Tully, 1994; Locke, 1991; See *et al*, 2003; Sitkin *et al*, 2011) provide a unifying purpose (Colbert *et al*, 2008) and promote organisational learning (Levitt and March, 1988).

Targets are also claimed to enhance managerial responsibility, motivate workers, and foster accountability (Capon *et al*, 1987; Jennings and Haist, 2004). Additionally, they can promote transparency and legitimacy (Hoque *et al* 2004) and have led to improvements in reported levels of performance in hospitals and schools (Hood, 2006; Hood *et al*, 2009; Le Grand, 2010; Propper *et al*, 2010). Overall, Locke and Latham (2006) propose, “Goals direct attention, effort, and action toward goal-relevant actions at the expense of non-relevant actions” (2006, p.265).

Many commentators envision targets as a key feature of performance management systems (Simons, 1995; Bourne *et al*, 2009; Micheli and Manzoni, 2010; Micheli, 2012b). Neely, for example, argues for multidimensional, stakeholder-driven approaches that incorporate targets as legitimate components of ‘learning systems’ (Neely, 2007; Neely *et al*, 1997; 2001; Neely and Powell, 2004). Similarly, Micheli, suggests they should be used to “...direct attention, mobilise effort and inform strategy development” (2012a, p.90).

Experimental research has shown that the presence of goals in a variety of settings has led to increased performance outputs. Whilst tending to focus on simple physical tasks, this research does suggest participants perform better when assigned a target to aspire towards. For example, subjects:

- Squeezed a grip more tightly (Botterill, 1977);
- Lifted heavier weights (Ness and Patton, 1979);
- Exercised at a higher rate (Bandura & Cervone, 1983);
- Pedalled faster (Roberts and Hall, 1987), and;
- Endured pain for longer (Stevenson *et al*, 1984).

Nevertheless, there is general acknowledgment of an association between targets and unintended consequences, although the consensus seems to be the target itself is not responsible (Latham, 2004). For instance, Locke and Latham (2006) suggest adverse behaviours arise primarily when goals are framed in a manner that associates them with threat or failure (i.e. the framing is responsible, rather than the target).

Others, however, contest the effectiveness of target-setting. For example, Deming (1986, 1994) asserts that merely setting a numerical target does not provide a method or capacity for achieving it. Locke and Latham (1991) appear to accept this; whilst insisting that goal setting promotes effort, persistence and direction, they observe “...there are times when these three mechanisms, including the use of previously learned skills, are insufficient to attain a goal” (1991, p.228). They also find that self-set goals are no more effective at increasing performance than assigned goals, or participatively-set objectives (Locke and Latham, 1990).

Separate research concludes that targets do not improve performance (Shapira, 1989; Ghobadian *et al*, 2009); specifically, Huber and Neale (1987) and Northcraft *et al* (1994) found that targets can cause individual performance outputs to be boosted at the expense of collective organisational objectives, echoing the dangers of sub-optimisation, as discussed in respect of league tables. Indeed, Locke and Latham (1991) acknowledge, “Trying for specific, challenging goals may actually hurt performance in certain circumstances” (1991, p.229).

Furthermore, the perception stretch goals improve performance is claimed to be overstated (Sitkin *et al*, 2011). Citing effusive endorsements of the practice (see Hamel and Prahalad, 1993; Collins and Porras, 1994; Tully, 1994; Hamel, 1998), Sitkin *et al* (2011) argue their usefulness is exaggerated and often based on anecdotes, outlier cases and non-representative samples (see also Denrell, 2003). Additionally, whilst Locke (2001) claims stretch goals can improve the performance

of high ability managers, he acknowledges they may also provoke a fall in overall morale due to other employees becoming demotivated (Locke, 2001). See *et al* (2003) warn this impact on morale may even lead to a drop in performance.

Research also suggests targets fail to improve performance in complex settings (Earley *et al*, 1989; CIPD, 2016a) and can even have a negative effect on performance (CIPD, 2016b, p.17). For instance, in studies relating to air traffic control, Kanfer and Ackerman (1989) found targets actually impeded performance. Likewise, Horton and Smith (1988), Loveday (2006) and Mackenzie and Hamilton-Smith (2010) claim complex systems (such as public services) do not respond well to target-driven performance management.

This is perhaps, in part, because multiple conditions affect outputs in complex settings (Collier, 2006). This point is generally accepted by goal setting theorists; Locke *et al* (1989) and Locke and Latham (2002; 2006) acknowledge targets are ineffective in environments where subjects do not have full control over their performance capability. Additionally, Seijts and Latham (2001) and Locke and Latham (2009) accept specific goals do not always lead to improved performance. Instead, Locke and Latham (2009) suggest broader ‘learning goals’ (such as to acquire task knowledge) can be beneficial as an alternative.

Targets are posited to be a particularly ill fit in the public sector (Behn and Kant, 1999), as it is different, complex, and fosters values such as political independence, professionalism and neutrality, rather than the attainment of profit (Hughes, 2003). Moreover, Armstrong (1997) contends market theory “...is not robust enough to embrace the full range of public sector activities such as governance and guarding public interest” (1997, p.3), making it difficult to transplant private sector methodologies into non-indigenous settings (Hvidman and Andersen, 2013).

Therefore, it may well be that goal setting can help increase outputs in the short term, and for non-complex tasks, whilst being less effective in more complex environments. This is perhaps unsurprising, as numerical targets have their origin and greatest application in settings where outputs are relatively easy to quantify, such as in manufacturing plants.

### 2.1.6.3 Setting Numerical Targets

Although the use of numerical targets is widespread, it is rare to see guidance on how to actually define them. For example, Fortuin (1988) advises targets should be revised annually and set at a ‘more challenging’ threshold once achieved, but does not elaborate on how this adjustment should be calculated. Similarly, whilst Bourne and Franco-Santos (2010) emphasise the importance of thorough data analysis when setting ‘high but achievable’ targets, they do not explain how to calculate where the precise target should be placed.

Similarly, the Department for Education (2012) exhorts schools to establish ‘ambitious and realistic’ targets by first examining the previous two years’ data, but provides no supporting methodology on how these targets should be determined. Courty *et al* (2007) cite various approaches; they note that experimentation and statistical methods are sometimes employed, but warn even the more sophisticated techniques are not particularly robust (see also Bird *et al*, 2005, pp.7-8).

In policing, Shane (2007; 2008) recommends analysing historic data to help forecast future trajectories, but then simply suggests “...you now want to set an end outcome: to reduce the average number of monthly crimes by 10%” (2007, p.188). No rationale or method for fixing the 10% target is offered. Indeed, it seems the most common method of setting targets is to simply make an arbitrary adjustment against previous years’ figures (Wheeler, 2000; Mackenzie and Hamilton-Smith, 2010).

Despite obvious limitations, in the experience of the author, these types of approaches remain the dominant method for target-setting in the UK police service. Although comparisons to previous years’ figures, averages, or peer data usually provides the *baseline* from which numerical targets are derived, the consistent theme at the *actual point of setting the target* is that the exercise tends to be an arbitrary adjustment against this figure.

One striking example in UK policing was the ‘20 / 20 / 20 Challenge’ issued by the London Mayor’s Office for Policing and Crime (MOPAC), which required the Metropolitan Police to “Cut crime by 20%, boost public confidence by 20% and cut costs by 20%” (MOPAC, 2012, p.2).

The process for arriving at the 20% crime reduction target was described by Stephen Greenhalgh, the Deputy Mayor for Policing and Crime, as follows:

“The Mayor and Commissioner both wanted crime to fall. The argument was by how much. This was settled – over a curry – at 20% over four years” (Greenhalgh, 2014, p.3).

This approach of establishing numerical targets by benchmarking and then adjusting against previous years’ figures (or selecting an arbitrary value), results in extremely specific (and often curious) police performance targets, such as those listed below:

- “Reduce Burglary on business premises (where the total value of the property stolen was over £1000) by 5% compared to 2009/10” (Leicestershire Police Authority, 2010, p.24);
- “90% of terrorist scenes managed to a good or very good standard” (Metropolitan Police, 2005, p.27);
- “43% reduction in people seriously injured in road traffic accidents” (Transport Scotland, 2018, p.2), and;
- “Reduce bureaucracy by decreasing the volume of manually-produced performance reports by at least 20%” (British Transport Police, 2012, p.5).

Even the FBI prescribed a numerical target to disrupt 125 terrorist acts during 2015 (US Department of Justice, 2016, p.8), although it is unclear why the threshold was set at 125. Similarly, in 2013, Strathclyde Police set a target of 459,438 stop and searches (Scottish Centre for Crime and Justice Research, 2014, p.7) and 243,206.3 during 2009-2010 (Strathclyde Police, 2010, p.6). (The decimal point is not a typographical error!) Below are further examples:

Figure 2.12: Strathclyde Police – Performance report

Violence, Disorder and Antisocial Behaviour OMIS (Force Targets) 30/03/2009 - 21/03/2010	Target	2009/10	% above/below target <sup>1</sup>
Increase the detection rate for serious assault from 48% to 50%	50%	53.5%	3.5%
Decrease the number of serious assaults by 5%	3521.9	3277	-7.0%
Increase detection rate for Robbery from 39% to 41%	41%	44.8%	3.8%
Decrease the number of robberies by 6%	1674.2	1 519	9.3%
Increase the detection rate for petty (common) assault from 65% to 67%	67%	66.6%	-0.4%
Decrease the number of petty (common assault) by 3%	32320.3	34074	5.4%
Increase the detection rate for total crimes and offences in domestic abuse incidents from 70% to 72%	72%	71.1%	-0.9%
Decrease the proportion of repeat offender incidents in domestic abuse incidents by 2%	60.1%	61.5%	1.4%
Decrease the proportion of repeat victim incidents in domestic abuse incidents by 2%	61.1%	62.1%	1.0%
Increase the number of stop searches by 6%	243206.3	312844	28.6%
Increase the proportion of positive stop searches by 1%	7.4%	7.2%	-0.2%
Increase the number of detections for consuming alcohol in a public place by 7% <sup>7</sup>	21977.0	28407	29.3%
Increase the number of detections for carrying knives/bladed instruments by 7%	2651.0	2114	-20.3%
Increase the number of detections for urinating in a public place by 5% <sup>7</sup>	5901.2	8100	37.3%
Increase the detection rate for racially motivated crimes and offences in racist incidents from 53% to 56%	56%	60.2%	4.2%
Increase the detection rate for homophobic crimes and offences in homophobic incidents from 64% to 67%	67%	70.9%	3.9%

(Strathclyde Police, 2010, p.6).

Figure 2.13: West Yorkshire Police - Key Performance Indicators and Targets

Priority	Key Performance Indicators	Actual Performance 2011/12	Target 2012/13
Local Policing	Reduce the level of acquisitive crime	42,051	41,604
	Reduce the level of burglary dwelling	17,806	17,359
	Continue to tackle ASB to impact on the proportion of residents who believe that ASB has increased in their local area	14.3%	14.3%
Protecting the Public from Serious Harm	Continue to tackle the level of serious violent crime	787	787
	Improve the repeat victimisation rate for domestic violence	38.2%	Less than 38.2%
	Maintain the sanction detection rate for serious sexual offences	29.5%	29.5%

(West Yorkshire Police, 2012, p.4).

Figure 2.14: British Transport Police – Sickness data table

	A Div'n	East	South	TfL	B Div'n	Midlands	Pennine	Wales	Western	C Div'n
<b>Overall sickness to be less than 7.3 days per employee</b>										
YTD Performance	7.55	5.25	5.21	6.32	6.45	6.92	8.22	6.64	7.81	7.58
YTD Linear Target	7.30	7.30	7.30	7.30	7.30	7.30	7.30	7.30	7.30	7.30
Last month performance	7.13	4.54	4.57	5.33	5.85	6.38	7.30	5.61	7.56	6.85
Last month target	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69
Better / worse than LM	↑	↓	↓	↓	↓	↓	↓	↓	↓	↓
Same period last year	7.32	n.a.	n.a.	n.a.	6.63	n.a.	n.a.	n.a.	n.a.	7.84
Note: Last Year sickness data is for the Interim Divisions (averages for the appropriate old Areas).										
<b>Police officer sickness to be less than 7.3 days per officer</b>										
YTD Performance	6.52	4.44	4.94	7.12	6.32	8.43	10.35	8.16	9.52	9.17
YTD Linear Target	7.30	7.30	7.30	7.30	7.30	7.30	7.30	7.30	7.30	7.30
Last month performance	6.14	3.67	4.28	5.91	5.75	7.79	9.20	7.12	9.19	8.30
Last month target	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69
Better / worse than LM	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
Same period last year	5.05	n.a.	n.a.	n.a.	5.39	n.a.	n.a.	n.a.	n.a.	8.73

(British Transport Police Authority, 2015, p.40)

Figure 2.14 uses binary comparisons to compare the current value to last month and the same period last year, as well as year-to-date and monthly targets. Although not explicitly ranked, geographic areas are also compared. Furthermore, the symbolism of red and green shading and ‘up’ and ‘down’ arrows reinforces perceptions of performance being ‘good’ or ‘bad’; even the descriptors employ the language of ‘better’ or ‘worse’, compared to last month.

Figure 2.15 illustrates how the Metropolitan Police presents stop and search data:

Figure 2.15: Metropolitan Police – Stop and search performance data

BOCU Location of Search	Volume (PACE + S60)				Arrest Rate - PACE & S60 TARGET 20%					
	Month	R3 Period			Month	R3 Period				BOCU Rank
	Aug 13	Jun 13- Aug 13	Mar 13- May 13	R3 Period Move	Aug 13	Change vs. Jul 13	Jun 13- Aug 13	Mar 13-May 13	R3 Period Move	
Barking and Dagenham	424	1,500	1,990	-490	17.2%	-0.7%	18.5%	18.5%	-0.0%	13
Barnet	500	1,610	2,543	-933	10.8%	-9.5%	15.9%	13.8%	+2.1%	21
Bexley	310	1,020	1,195	-175	18.1%	-2.7%	20.5%	18.7%	+1.8%	6
Brent	1,065	2,981	3,860	-879	19.0%	-3.1%	20.9%	18.5%	+2.3%	5
Bromley	754	2,814	3,367	-553	14.7%	-3.9%	16.5%	14.9%	+1.6%	20
Camden	730	2,791	3,535	-744	15.9%	+2.7%	13.8%	11.9%	+1.9%	29
Croydon	688	2,461	2,959	-498	11.0%	-5.7%	14.4%	14.2%	+0.2%	27
Ealing	456	1,854	3,070	-1,216	14.7%	-3.1%	15.8%	13.0%	+2.8%	22
Enfield	636	1,895	2,779	-884	19.0%	+5.1%	17.0%	12.1%	+4.9%	19
Greenwich	541	1,990	2,373	-383	25.0%	+6.1%	19.6%	15.1%	+4.5%	8
Hackney	783	2,246	2,135	111	20.8%	-2.2%	22.4%	23.8%	-1.5%	3
Hammersmith & Fulham	891	2,843	4,218	-1,375	19.4%	-0.5%	18.3%	16.8%	+1.5%	14
Haringey	335	1,267	1,783	-516	23.6%	+3.1%	21.2%	16.5%	+4.7%	4
Harrow	561	1,864	1,927	-63	16.9%	+3.2%	14.6%	11.6%	+3.1%	25
Havering	339	1,071	1,327	-256	17.1%	+5.9%	15.4%	16.2%	-0.8%	23
Heathrow	61	328	310	18	24.6%	+4.3%	22.9%	25.2%	-2.3%	2
Hillingdon	242	1,049	1,357	-308	14.0%	-5.6%	17.8%	19.2%	-1.4%	17
Hounslow	556	1,933	2,586	-653	11.7%	+0.0%	12.0%	11.5%	+0.5%	33
Islington	385	1,384	1,867	-483	13.8%	-5.0%	18.1%	18.3%	-0.1%	15
Kensington And Chelsea	755	1,530	2,087	-557	12.5%	-3.7%	14.5%	11.9%	+2.6%	26
Kingston-Upon-Thames	326	950	1,552	-602	12.6%	-13.0%	19.7%	17.7%	+2.0%	7
Lambeth	1,002	2,654	2,189	465	15.2%	-4.6%	17.2%	15.4%	+1.7%	18
Lewisham	528	2,049	2,213	-164	19.1%	-1.2%	19.4%	15.7%	+3.7%	10
Merton	216	722	989	-267	20.8%	-2.8%	24.0%	24.5%	-0.5%	1
Newham	779	2,591	3,263	-672	19.9%	+0.8%	19.3%	13.9%	+5.4%	11
Redbridge	715	2,639	3,748	-1,109	12.7%	-3.4%	13.4%	12.9%	+0.6%	30
Richmond-Upon-Thames	378	1,093	1,539	-446	16.4%	+1.0%	14.1%	7.7%	+6.4%	28
Southwark	1,582	5,048	5,319	-271	12.9%	+0.7%	12.7%	12.6%	+0.1%	31
Sutton	234	769	953	-184	20.9%	-2.6%	19.5%	11.3%	+8.2%	9
Tower Hamlets	884	3,367	4,192	-825	14.9%	-1.2%	14.7%	13.0%	+1.8%	24
Waltham Forest	593	2,244	2,895	-651	18.2%	-3.2%	18.6%	15.2%	+3.4%	12
Wandsworth	496	1,637	2,111	-474	20.4%	+4.1%	17.8%	14.8%	+3.0%	16
Westminster	2,229	7,023	6,722	301	12.3%	+0.1%	12.0%	12.3%	-0.3%	32
<b>MPS</b>	<b>20,974</b>	<b>69,217</b>	<b>84,953</b>	<b>-15,736</b>	<b>16.0%</b>	<b>-1.0%</b>	<b>16.4%</b>	<b>14.6%</b>	<b>+1.7%</b>	<b>--</b>

Metropolitan Police (2013b, p.4)

This table also uses raw data, colour-coding and comparisons to various numeric reference points, typifying much police performance information. Features include:

- Binary comparisons against the previous month;
- Ranked comparisons between Boroughs, and;
- Numerical targets, namely (i.e. 20% arrest rate to result from stop and search).

Similarly, narrative about performance often combines numerical targets with binary comparisons and comparisons between peers:

“The target for 2012/13 is to achieve a detection rate of 21%. Between April and December 2012 the force achieved a detection rate of 18.6% which is off target and is 3.5% lower than the same time last year...The level of performance achieved places the Force in 1st position within its MSG but with a deteriorating trend” (Staffordshire OPCC, 2013, p.4).

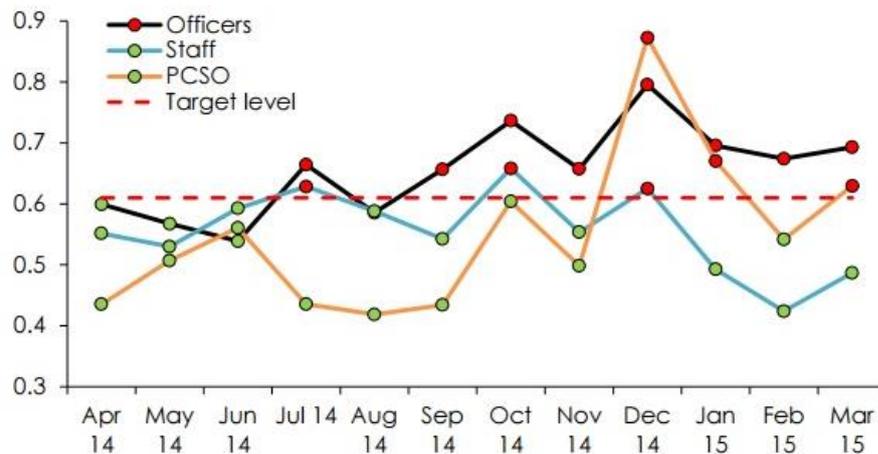
This demonstrates how these three dominant formats are routinely integrated; firstly, binary comparisons are used in an attempt to understand data, then numerical targets are devised with the intention of securing a favourable place in a league table. Occasionally however, policing targets are depicted graphically rather than as percentages. Whilst this may at first appear more sophisticated, the notion of fixing a target in this way disregards the statistical concept of *variation* (Wheeler, 2000).

Crucially, if a target is placed anywhere within the expected range of variation, then despite constant effort and ability, sometimes it will be hit and other times missed (Bird *et al* 2005). Conversely, if it is placed outside the expected range, then it cannot predictably be achieved under current system conditions. Consequently, citing Castellano *et al* (2004), Heinrich and Marschke (2010) highlight the:

“...failure to understand variation in process or the inherent inability of a stable process to regularly achieve point-specific targets as a fatal flaw of performance measurement systems” (2010, p.200).

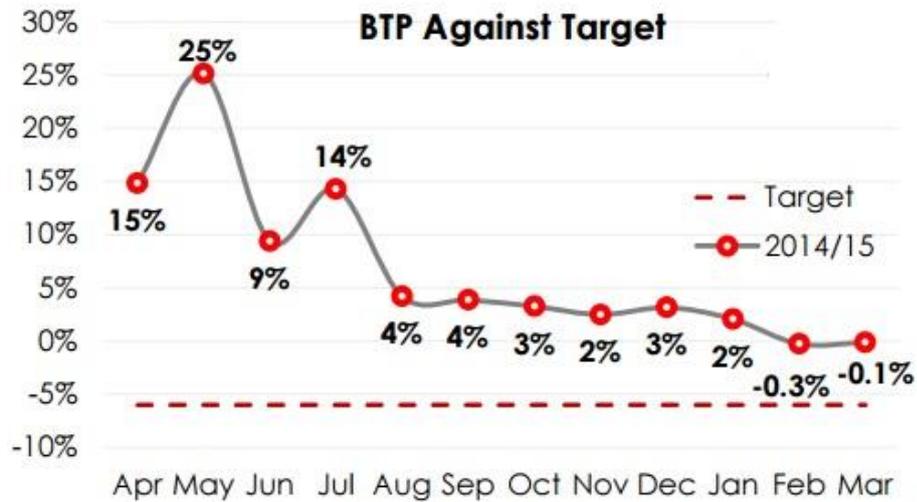
The following examples illustrate these points; in Figure 2.16, the target for sickness levels appears to have been placed in the midst of normal variation, whilst in Figure 2.17 it has been set below all data points, suggesting it is likely to be outside the normal range (a ‘stretch target’). In both cases, control limits have been omitted, meaning it is impossible to identify what is statistically ‘normal’.

Figure 2.16: British Transport Police – Sickness target chart



(British Transport Police Authority, 2015, p.11)

Figure 2.17: British Transport Police – Railway disruption target chart

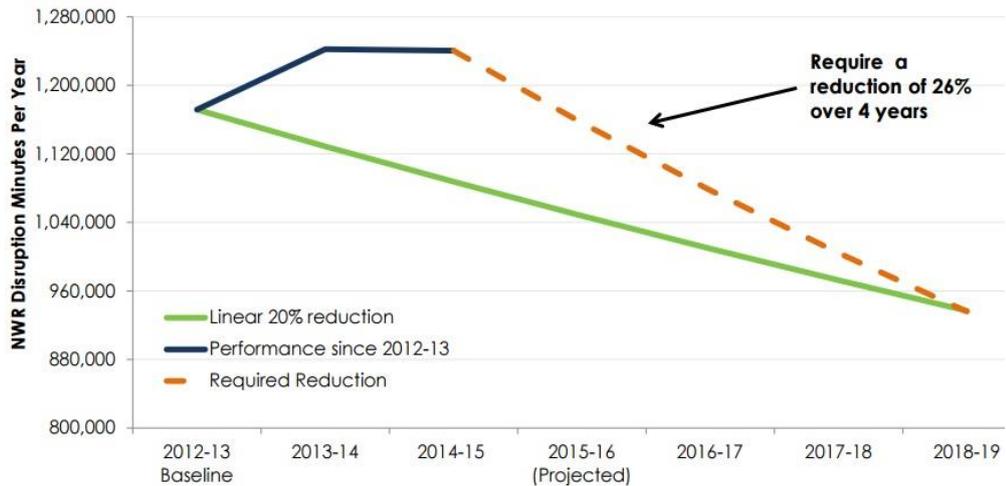


(British Transport Police Authority, 2015, p.9)

Furthermore, whilst it may be feasible to predict future performance *within a range* (subject to constancy of system conditions and confidence intervals), it is not possible to foresee the precise numeric value of some future state in order to designate it as a target. For example, if the crime rate was steadily decreasing it may be possible to predict there will be between 90,000 and 100,000 offences committed in a years' time, but it would not be possible to state precisely how many (e.g. 94,500 offences).

Figure 2.18 below illustrates this point, as it displays linear trajectories towards a precise aspirational end point; the chart fixates on a specific numeric value to be achieved by the 2018-2019 financial year, yet there is nothing to indicate whether this target falls within the expected range, nor any acknowledgement of variation.

Figure 2.18: British Transport Police – Railway disruption progress chart



(British Transport Police Authority, 2015, p.5)

Overall, the use of numerical targets in UK policing is firmly embedded and interwoven with binary comparisons and league tables, within performance measurement systems characterised by simplistic data display. The following section will next examine behaviour commonly associated with target-driven performance management, as well as ethical implications and unintended consequences.

#### 2.1.6.4 Targets and Behavioural Change

Proponents and critics of targets tend to agree that they act as a trigger for behavioural change; indeed, that their *purpose* is to change behaviour. Bevan and Hood (2006) assert:

“Governance by targets rests on the assumption that targets change the behaviour of individuals and organizations...” (2006, p.8).

However, perhaps even more so than with league tables, the literature suggests many of the behavioural changes associated with target-driven performance management are damaging (Smith, 1990; 1995; De Bruijn, 2002; Pidd, 2005; Bevan and Hood, 2006; Hood, 2006; 2007; Loveday, 1999; 2006; 2008; Courty *et al*, 2007; Rothstein, 2008). In particular, concerns regarding behavioural dysfunction and unintended consequences have existed for several decades (see Wagner, 1954; Blau, 1955; Berliner, 1956; Ridgway, 1956). More recently, Locke and Latham (2009) highlight various unwanted side effects:

“...excessive risk taking, increases in stress, feelings of failure, using goals as a ceiling for performance, ignoring non-goal areas, short-range thinking, and dishonesty / cheating” (2009, p.20).

There is extensive commentary regarding potential reasons for this apparent association. Goodhart’s eponymous Law warns, “Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes” (Goodhart, 1975; Chrystal and Mizen, 2001, p.4); the inference being that pressure to meet targets is likely to trigger counterproductive reactions. Wheeler (2000), posits targets initiate one (or more) of the following responses amongst workers:

1. They work to improve the system.
2. They distort the system.
3. They distort the data.

(Wheeler, 2000, p.20).

Proponents of target-driven performance management anticipate the positive response described at (1); however, the possibility of the adverse reactions listed at (2) and (3) cannot be ignored. As with league tables, the intention of target setters is to improve performance by focusing attention on organisational priorities. However, narrow focus on areas subject to targets can lead to organisational purpose becoming supplanted by a *de facto* purpose of attaining the target (Hammer, 2007; Barsky, 2008); a condition which the Home Office (2008c) likens to “...hitting the target but missing the point” (2008c, p.12).

Although Locke and Latham (2006) insist, “Goals direct attention, effort, and action toward goal-relevant actions at the expense of nonrelevant actions” (2006, p.265), others agree about the directive effect of targets, but highlight adverse side effects (Jacob, 2005; Heinrich, 2008). Typically, Spitzer (2007) suggests workers will try and attain targets “...by whatever means are available” (2007, p.43), whilst Wu *et al* (2008) warn this focus can lead to other performance dimensions being neglected. Shorrocks and Licu (2013) argue:

“Experience shows that numerical targets do indeed set direction; they set people in the direction of meeting the numerical target, not necessarily achieving a desired system state” (2013, p.11).

Furthermore, Bandura (1999) argues performance targets can cause moral disengagement, leading to unethical conduct. Similarly, Schweitzer *et al* (2004) and Barsky (2008) observe that specific goals tend to drive perverse behaviour, whilst others highlight an association between targets and gaming (Bevan and Hood, 2006).

Types of gaming include:

- ‘Ratchet effects’: where next year’s targets are based on previous years’ performance, resulting in a perverse incentive for managers to under-report current performance in order to secure less demanding future targets;
- ‘Threshold effects’: where performance across different functions is reported as a whole, obscuring poor individual or departmental performance and even encouraging high performers to allow their performance to deteriorate, and;
- ‘Output distortions’: where targets are prioritised at the expense of unmeasured performance dimensions. (Also termed ‘effort substitution’ - see Kelman and Friedman, 2009, p.917; Benaine and Kroll, 2019).

(Adapted from Bevan and Hood, 2006, p.9).

Smith (1990) and Pidd (2005) identify other themes:

- ‘Tunnel vision’: where managers select some targets (usually the easiest to achieve or measure) and ignore others;
- ‘Sub-optimisation’: where departments act in a way that protects their own interests but damages the performance of the overall system, and;
- ‘Myopia’: where managers focus on achievable short-term objectives at the expense of longer-term aims.

Such behaviours have been observed throughout the UK public sector, chiefly in the fields of healthcare and education, although cases have also arisen in Children’s Services, (Reason, 2000; Munro 2005; Fish *et al*, 2008; Broadhurst *et al*, 2010), job training and placement schemes (Heckman *et al*, 2002), assessment of benefits claims (Wintour and Domokos, 2013) and immigration processing (Mulley, 2012).

Examples from healthcare include falsification of hospital records (British Medical Association, 2007; Healthcare Commission, 2009; Francis, 2013a; 2013b, Longman,

2013), patients being left in ambulances outside A&E departments to avoid breaching handover targets (Loveday, 2006; Gretton, 2013), people being admitted or retained on wards unnecessarily (Smith, 1993; Loveday 2005), and wheels being removed from gurneys to enable them to be reclassified as ‘beds’ (Commission for Health Improvement, 2002).

In other cases, ambulance response times were falsified to ensure compliance with attendance targets (Commission for Health Improvement, 2003; Audit Commission, 2007; Bevan and Hamblin, 2009; Murfitt, 2016), along with various types of output distortions (Holmstrom and Milgrom, 1991; Gibbons, 1998; Carvel, 2003) such as the cancellation of ophthalmic appointments to ensure other waiting time targets were met (Public Administration Select Committee, 2003) and ‘risk adjustment’; i.e. where targets generate, “...a strong incentive to select healthy and compliant patients and to avoid severely ill and noncompliant patients” (Eijkenaar, 2011, p.119).

In education, examples include exam grade inflation (Koretz, 2002), multiple registrations to different exam boards (Adams, 2013; Paton, 2013a; 2013b), exclusion of low-achieving students (Kantrowitz and Springen, 1997; Loveday *et al*, 2004; Figlio, 2005), ‘cream-skimming’ during placement allocation (Carnoy *et al*, 2005) and manipulation of attendance records to hit truancy targets (Rothstein, 2008; Shorrocks and Licu, 2013).

Finally, there are also concerns about the psychological effects of target-driven performance management; it is claimed to degrade trust and autonomy (CIPD, 2003; Taylor-Gooby, 2009), demoralise workers (O’Neill, 2002; Jackson, 2005), disempower middle managers (Fitzgerald *et al*, 2002), constrain professionalism (De Bruijn, 2007), destroy goodwill (Ghobadian *et al*, 2009), discourage cooperation (Barsky, 2008), suppress innovation (Shane, 2010) and dehumanize (Western, 2007).

Overall, the literature suggests that where behavioural dysfunction or unintended consequences are observed in target-driven performance management systems, they tend to be remarkably consistent in character.

#### **2.1.6.5 Targets in Policing**

The literature relating to targets in policing is less well-developed than in healthcare or education, but it appears there are strong similarities with the experiences of other

sectors. Perhaps one of the most well-known models of target-driven performance management in policing is that of the Compstat regime, introduced into the New York Police Department (NYPD) by Mayor Bill Bratton in 1994 (Bratton, 1998; Kelling and Bratton, 1998; Bratton and Malinowski, 2008; Pasha *et al*, 2018). Described as “...a managerial system, aimed at holding commanding officers accountable” (Eterno and Silverman, 2012, p.24), Compstat involves twice-weekly public meetings where crime statistics are examined and local police commanders are questioned by senior officers about crime rates (Willis *et al*, 2007).

Proponents claim these meetings promote accountability, organisational learning, and that (alongside other reasons) the Compstat model has been responsible for dramatic reductions in recorded crime (Behn, 2008b; Bratton and Malinowski, 2008; Behn, 2014). One of the benefits of Compstat is that data can be used to inform managers about crime patterns, although precinct commanders are also directly compared against each other, and management dissatisfaction with performance can result in individuals being removed from their post. This gives the Compstat system immense behavioural power (Moore and Braga, 2003).

The approach has also become popular in other US cities (Weisburd *et al*, 2003; Behn, 2006), such as Baltimore, where ‘Citistat’ has taken root as a progeny of Compstat (Filichio, 2005; Behn, 2008a), and Los Angeles, with ‘Compstat Plus’ (Bratton and Malinowski, 2008). Whilst Compstat’s objective is to identify crime hotspots and emerging trends through data analysis, it is pertinent to note that this ‘analysis’ is routinely limited to comparing current crime data to that of “...the same period last year at three levels: weekly, 28 days, and year to date” (Eterno and Silverman, 2012, p.24); in essence, binary comparisons.

Furthermore, the drive for accountability sometimes manifests itself in aggressive and bullying behaviours on the part of senior officers charged with holding local commanders to account. Eterno and Silverman (2012) recount, “At times, commanders are berated and embarrassed in front of their peers...” (2012, p.25) and “...peppered with questions” (Moore and Braga, 2003, p.447). Critics argue that instead of fostering accountability, Compstat creates a high pressure environment (Kroll, 2017) and reinforces authoritarianism (Willis *et al*, 2004; 2007); Behn even suggests that ‘accountability’ often simply translates as ‘punishment’ (2001, p.3).

The behavioural impact of target-based systems such as Compstat is profound; Eterno and Silverman (2012) cite examples of the lengths officers have gone to in order to meet quotas. For instance, the number of prosecutions and forcible searches (known as ‘stop and frisk’) increased exponentially; citizens were also prosecuted for eating doughnuts in a playground (2012, p.215) and playing chess in a public park (2012, p.216).

Similar experiences have been observed elsewhere. For example, in Chicago, some murders were allegedly downgraded to ‘non-crime status’ to avoid breaching a target of 500 murders per year (Bernstein and Isackson, 2014), whilst in the UK, targets actively reduced some forces’ willingness to investigate offences (Frakes, 2018). In Alabama, officers were required to issue fixed numbers of traffic citations per month (Reason, 2013), whilst Dutch officers issued large volumes of fines in order to meet targets, in cases where they otherwise would not have done so (Hoogenboezem and Hoogenboezem, 2005; Terpstra and Trommel, 2009).

In the UK, the practice of aiming for ‘low hanging fruit’ to meet targets is renowned; it occurs when effort is focused towards generating large numbers of outputs, such as for arrests, stop and searches, tickets or detections (Cumming, 2014; Metropolitan Police Federation, 2014). This practice, which also been documented by HMIC (1999b; 2008), Neyroud and Disley (2007), Chatterton (2008), Loveday (2008) and the Home Office (2008c), has led to people being unnecessarily criminalised (Flanagan, 2008); examples include children being arrested for throwing slices of cucumber, or drawing on pavements with chalk (Barratt, 2007; Leapman, 2007; Tendler, 2007; Wright, 2007; Monro, 2008).

In other areas, investigators would prioritise easy cases toward the end of monthly or quarterly reporting periods to meet targets (HMIC, 1999). This phenomenon, known as ‘storming’ (Argyris, 1952; Granick, 1954; Berliner, 1956) has been observed as a common response to quotas. Activity intensifies as the threshold approaches and outputs tend to peak, followed by a slump at the beginning of the subsequent reporting period. ‘Storming’ has been widely identified in the economic literature (Schick, 1998; Pollitt, 1999), but is also evident in policing.

As with league tables, the implications of targets for ‘stop and search’ are particularly topical. For instance, officers have been accused of under-recording such

interactions in order to artificially minimise the proportion of unsuccessful searches. Such fears were raised at the MOPAC Stop and Search Working Group:

“What some of the young people are saying is that the police officers are not filling in slips. When they get a negative stop they do not fill in their slips and when they get a positive stop they do” (MOPAC, 2013, p.3).

Targets can also act as a perverse incentive not to record crime (Young, 1991), or to downgrade offences inappropriately (Seddon, 2008; Eterno and Silverman, 2012; Crockett, 2013; Shaw, 2013). Examples include classifying burglaries or robberies as simple ‘theft’, or recording attempted burglaries as ‘damage’ (Patrick, 2013; see also UK Statistics Authority, 2010). Furthermore, targets can even discourage proactive police activity; this paradox was identified in a report on street robberies:

“It is a moot point whether it made sense for the government to set a target to reduce police recorded robbery in the first place, given that increases might well reflect enhanced police action in this area. Ironically, the government’s target on street crime has risked creating a perverse incentive for police forces to avoid identifying and recording robbery offences” (Centre for Crime and Justice Studies, 2007, p.33).

Internal reviews by UK police forces have also found multiple examples of targets adversely affecting behaviour (Kent Police, 2013a; 2013b; Wiltshire Police, 2014a; 2014b). Such cases include ‘fiddling’ response time data (Whitehead, 2010), with some officers even making bogus 999 calls in order to improve call answering times (Loveday, 2006); police call centre staff have also been known to answer, then disconnect, incoming calls to meet call-handling targets (Wiltshire Police, 2014b).

Instances of extremely unethical behaviour have also been reported. For example, officers from the Metropolitan Police allegedly encouraged some rape victims to retract their statements and reclassified serious sexual offences so they were no longer registered as crimes (IPCC, 2013; Johnson, 2013). Separately, other officers apparently achieved detections by encouraging suspects to admit offences they had not committed; some even created records for non-existent crimes, thereby generating an artificially high detection rate (Young, 1991). Patrick (2009) proposes a typology of police gaming practices which classifies such unethical behaviour:

- ‘Cuffing’ – under-recording or downgrading crime;
- ‘Nodding’ – inducing suspects to admit offences for which they were not responsible;
- ‘Skewing’ – focusing activity towards some performance dimensions at the expense of others (i.e. output distortions), and;
- ‘Stitching’ – fabrication of evidence and / or detection records.

(Adapted from Patrick, 2009).

Citing Chatterton and Bingham (2006) and Chatterton (2008), Patrick (2009) asserts that such conduct has occurred throughout UK policing, as a direct consequence of target-driven performance management. Overall, the police target culture has become synonymous with ‘corner-cutting’, systemic failure to accurately record crime, distortion of activity, erosion of integrity, as well as damage to trust, morale and officers’ wellbeing (HMIC, 1999; Rowson *et al*, 2012; Kent Police, 2013a; 2013b; Metropolitan Police Federation, 2014; Wiltshire Police, 2014).

In 2013, as a result of such concerns, the House of Commons Public Administration Select Committee (PASC) commenced an inquiry into whether crime statistics were accurately recorded by police forces, and if not, which factors influenced recording practices (PASC, 2013a). The Committee heard evidence of widespread malpractice, consistently traceable to crime reduction targets (PASC, 2013b).

The Committee’s final report was published in April 2014. It remarked upon an “...entrenched target culture...which persists to this day” (PASC, 2014, p.27). The report concluded targets “...tend to affect attitudes, erode data quality and to distort individual and institutional behaviour and priorities” (PASC, 2014, p.31). Consequently, the Committee issued the following recommendation:

“The Home Office...should make clear in its guidance to PCCs that they should not set performance targets based on Police Recorded Crime data as this tends to distort recording practices and to create perverse incentives to misrecord crime. The evidence for this is incontrovertible. In the meantime, we deprecate such target setting in the strongest possible terms” (2014, p.52).

### **2.1.6.6 Numerical Targets: A Summary**

The consensus throughout the literature is that targets change behaviour. Advocates anticipate positive behavioural changes; indeed, goal setting theorists have found targets can drive performance improvements. Others, however, highlight instances of behavioural dysfunction, arguing this outweighs perceived benefits. Concerns also exist about the lack of statistically reliable methods available for setting targets.

However, uncertainty remains about whether targets are relatively benign *per se*, with dysfunction only arising as a result of misuse. The literature certainly recounts aberrations that derive from improper application, such as bullying associated with Compstat. Nevertheless, it is unclear whether a portion of observed dysfunction may arise due to the visual appearance of numerical targets; does reliance on an (often arbitrary) aspirational reference point heighten the likelihood of flawed assumptions about performance, which then leads to a particular pattern of responses?

In other words, what *causes* managers to behave in certain ways when using targets to interpret performance data and can dysfunction arise regardless of the manner in which they are applied? This study will examine these questions, to try and establish whether the fashion in which numerical targets inherently display data is a direct antecedent to dysfunction, even in the absence of other aggravating factors.

## **2.2 Part Two: Data Display**

### **2.2.1 Introduction**

At the heart of this thesis is the prospect that the visual appearance of performance information may ultimately be a contributory factor towards dysfunction. Therefore, this section examines the subject of *data display* (or *data visualisation*), exploring underlying principles, insights from *reference dependence* theories and an overview of relevant research. It concludes by offering observations about antecedents to behavioural dysfunction, making the case for further study into the effects of data display on performance information use, and presenting the research question.

### **2.2.2 Data Display: An Overview**

The data display literature is extensive and varied (see Tufte, 1990; 1997; 2001; 2013; Card *et al*, 1999; Chen, 2006; Few, 2009; Lindquist, 2011; Ward, 2015; Isett and Hicks, 2018) and provides guidance relevant to the presentation of performance information. Good practice advises formats should offer “...clarity, precision and efficiency” and must “...avoid distorting what the data have to say” (Tufte, 2001, p.13). This is especially important where data are used to inform decision-making, as oversimplified, incomplete or irrelevant data can lead to inappropriate inferences being made (Tufte, 2013). In other words, *how* information is presented is just as important as *what* is presented (Lazard and Atkinson, 2015).

Effective data display requires context, depth, clarity and accuracy, as this is necessary to promote truthful, credible and precise findings (Schmid, 1983; Cleveland, 1985; Steinbart, 1989; Few, 2012; Yau, 2013). Formats should also be appropriate to the setting (Tufte, 2013), informative (Holmström, 1979; 1999; Linder and Foss, 2013), and useful for posing and answering questions (Kosslyn, 1994). Meaningful data formats are also necessary for making legitimate comparisons (Behn, 2006). Data display can therefore aid or impair interpretation (Wainer, 1984; Ammons and Rivenbark, 2008), affect decision-making (Elting, 1999) and influence task performance (Lusk and Kersnick, 1979).

The literature on data display covers a much wider field than the presentation of performance information; for example, it extends to the study of information systems (Claramunt *et al*, 2006; Shipley and Chakraborty, 2018) and computer science (De

Fanti *et al*, 1987; Chen, 1999; Ltifi *et al*, 2009). Similarly, data display research addresses pictorial representations (Ward *et al*, 2015), dynamic visualisations (Di Biase *et al*, 1992), 3D displays (Schroeder *et al*, 2006), virtual worlds (Proctor and Winter, 1998), animation (Tobler, 1970), use of colour (Ware, 2013), real time modelling (Dayley and Mayers, 1999), graphical simulation (Henderson and Mason, 2005) and techniques that integrate two or more such approaches (Pursula, 1999).

However, as the focus of this research is firmly on the implications of data display for the *design and use of numeric performance information*, these areas remain chiefly outside of its scope. Nevertheless, there are some general lessons from data display research which are directly relevant; for example, studies advise graphical formats tend to convey greater depth than tabular formats, thereby enabling better comprehension and aiding decision-making (Vessey, 1991; Speier, 2006; Smerecnik *et al*, 2010; Hirsch *et al*, 2015; Huber, 2016). Furthermore, separate research suggests the way in which data are framed can have a substantial impact of users' perceptions about performance (Olsen, 2013a; 2015; James and Van Ryzin, 2015).

Different types of data display lend themselves to different contexts and the provision of supporting narrative can assist users in understanding the content (Vernon, 1950). Olsen (2017), for example, found a preference for 'episodic' performance information (i.e. accounts, narratives, case studies) over statistical data. Furthermore, users' prior knowledge, personality, cognitive abilities and attitudinal attributes can influence behaviour towards information management systems (De Sanctis, 1984). Additionally, some managers feel a strong 'conditioning bond' for data formats they have used in the past (Lusk and Kersnick, 1979). For these reasons it is important to recognise that multiple variables affect the choice of format.

Nevertheless, it would appear that the selection and use of meaningful data display formats is a necessary first step towards fostering effective decision-making (Cleveland and McGill, 1984), especially as increased usability promotes purposeful performance information use (Kroll, 2013). Hill and Milner (2003) warn of the dangers of substandard graphical displays, insisting data must be displayed responsibly so as not to mislead. Similarly, Grainger *et al* (2016) assert that arbitrary or otherwise irrelevant data should be excluded. As poor data display can adversely

affect interpretation and impair decision-making, this has important implications for performance information use. As stated by Tufte (2013):

“There are right ways and wrong ways to show data; there are displays that reveal the truth and displays that do not” (Tufte, 2013, p.23).

There have been various studies exploring data display in the context of performance information use; these are generally positioned within the field of healthcare research and largely focus on data interpretation and decision-making. For example, Anhøj and Blok Hellesøe (2016) and Mountford and Wakefield (2016) conducted research into the use of ‘RAG’ reports (‘red / amber / green’ or ‘traffic light’ designs) as healthcare performance information, concluding that whilst they possess some heuristic value, such presentations are potentially misleading.

Separate research by Hawley *et al* (2008) and Faber *et al* (2009) observed that easy-to-read visual representations tend to be popular with performance information users. Similarly, other studies have found strong preferences for relatively simplistic formats, on account of their perceived accuracy and ease of use (Elbel *et al*, 2014); for example, RAG reports were found to be a popular choice, due to their uncomplicated visual appearance (Hildon *et al*, 2012).

The choice of data display format has also been found to affect the accuracy of users’ interpretation of performance information (Kurtzman and Greene, 2016). For example, during an experimental study, Gerteis *et al* (2007) tested the effects of a variety of formats, establishing that the appearance of various presentations significantly affected participants’ comprehension of data. Similar findings arose from separate research conducted by Damman *et al* (2011) and Brewer *et al* (2012).

Other research has examined the impact of presenting healthcare performance information in Statistical Process Control (SPC) charts<sup>5</sup> (see Shewhart, 1939; Wheeler, 2000), finding that this format provides greater context (Adab *et al*, 2002; Woodall, 2006; Sherlaw-Johnson and Bardsley, 2016). This tends to lead to improved accuracy of data interpretation (Marshall *et al*, 2004; Schmidtke *et al*, 2017a), however, despite these benefits, it was found that visually complex formats

---

<sup>5</sup> There is a brief overview of the main characteristics of SPC charts and their application toward the end of this chapter.

such as SPC are comparatively under-used in healthcare (Hibbard *et al*, 2001; Neuberger *et al*, 2017; Schmidtke *et al*, 2017b).

Further studies have shown that data display can affect decision-making (see Uhrig *et al*, 2006; Peters *et al*, 2007; NHS, 2018). Additionally, despite the general preference for simple, easy-to-digest formats, studies have demonstrated that decision-making tends to be adversely affected by their use (Elting, 1999; Hibbard *et al*, 2002). Overall, the consensus is richer formats are more strongly associated with enhanced comprehension of data and better choices, whilst simplistic methods of presenting data tend to mislead users and impair decision-making.

Considering the well-established guidance on effective data display, these conclusions are perhaps unsurprising; however, although the literature is clear regarding the influence of data display on decision-making, there is nothing that directly explores any potential relationship between such decisions and dysfunction. Similarly, other than highlighting the frailties of some formats, there is limited research into the reasons *why* they tend to mislead users. Below, Table 2.1 displays a synopsis of the healthcare studies referred to within this section.

*Table 2.1: Studies relating to data display and performance information use*

<b>Title</b>	<b>Focus of study</b>	<b>Key Findings</b>
Anhøj and Blok Hellesøe (2016)	Healthcare performance information (‘RAG’ reports)	Presenting performance information in nominal categories of ‘red’, ‘amber’, ‘green’ (i.e. RAG statuses) is overly simplistic and can be misleading.
Mountford and Wakefield (2016)	Healthcare performance information (‘RAG’ reports)	‘Stoplight reports’ (i.e. RAG) have some heuristic value, however a general lack of awareness of their limitations causes an overextension of their application.
Woodall (2006)	Healthcare performance information (SPC charts)	SPC charts that display control limits are useful for distinguishing systematic changes amidst the data from normal statistical variation (i.e. they are superior to time series charts that do not include control limits).
Sherlaw-Johnson and Bardsley (2016)	Healthcare performance information (SPC charts)	SPC charts are an effective data display format for monitoring clinical performance. Different types of SPC charts have particular applications within the healthcare field and can assist in assessing patient safety and improving quality of care.

Schmidtke <i>et al</i> (2017a)	Healthcare performance information  (SPC charts)	<p>The format of many performance graphs does not facilitate the statistical reasoning necessary to make necessary distinctions between 'signals' and 'noise'.</p> <p>In tests, participants more accurately identified outlying data using a control chart than using a non-control chart, but their ability to then apply that information to more complicated questions was limited.</p> <p>Compared to time-series graphs which do not feature control limits, SPC charts allow users to better understand variation in performance measures and in doing so, this impacts positively on decision-making and quality improvement efforts.</p>
Schmidtke <i>et al</i> (2017b)	Healthcare performance information  (SPC charts)	<p>SPC charts can help NHS board members interpret data effectively (by enabling them to discriminate between 'signals' and statistical 'noise'); however they are currently underused. Wider use would improve board members' decision-making and positively impact on patient care.</p>
Neuburger <i>et al</i> (2017)	Healthcare performance information  (SPC charts)	<p>Although time series charts are increasingly used by clinical teams to monitor performance, SPC charts are not widely used, partly due to uncertainty about which chart to use. Different types of SPC charts can be useful for different applications within the field of healthcare when presenting performance data.</p>
Adab <i>et al</i> (2002)	Healthcare performance information  (League tables and SPC charts)	<p>League tables are a well-established means of assessing comparative performance, but are easy to misinterpret. Conversely, SPC charts are a more robust method for displaying performance information; depending on whether data are presented in one or other of these formats can produce different interpretations.</p>
Marshall <i>et al</i> (2004)	Healthcare performance information  (League tables and SPC charts)	<p>Health service decision-makers perceive fewer outliers when presented with performance data in SPC charts than when displayed in league table format. This reduces tendencies to over-investigate what may appear to be unusual performance.</p>
Faber <i>et al</i> (2009)	Healthcare performance information  (Preferences for different formats)	<p>Easy-to-read presentation formats and explanatory messages are popular as they are believed to improve knowledge about performance information and promote a positive attitude towards the use of quality information.</p>

Hawley <i>et al</i> (2008)	Healthcare performance information  (Preferences for different formats)	Simple visual representations (e.g. pictographs) are preferred formats for communicating probabilistic information to patients in shared decision-making environments, particularly among lower numeracy individuals.
Hildon <i>et al</i> (2012)	Healthcare performance information  (Preferences for different formats)	Various performance information formats (tables, bar charts, caterpillar and funnel plots) and content (uncertainty displays, volume of outcomes, colour, icons, and ordering) were considered by patients' focus groups; simple displays (e.g. star ratings and RAG colour coding) were found to be most visually appealing to users, whilst unfamiliar formats (e.g. error bars) tended to be misinterpreted.
Elbel <i>et al</i> (2014)	Healthcare product information  (Preferences for different formats / effect on decision-making)	Experimental study testing participants' choices regarding hypothetical health care options, where different data display formats containing the same information were presented to participants. It was found that consumer choice was directly influenced by the method of presentation.  It was also found data displays that incorporate interpretive features (e.g. symbols or explanatory narrative) are most influential in users' decision outcomes.
Elting (1999)	Healthcare performance information  (Effect on decision-making)	Experimental study where participants were shown various types of data from a hypothetical clinical trial (tables, pie charts, bar graphs and icon displays) and asked to decide whether to continue with the trial. Accuracy of decisions was affected by the type of data display, as well as whether the data were positively or negatively framed.
Hibbard <i>et al</i> (2002)	Healthcare performance information  (Effect on decision-making)	Some data display formats make it easier for users to process and integrate quality data into their choices. However, other presentation formats influence consumers' decisions in ways that undermine their self-interest. Overall, the findings demonstrate the way information is presented affects how it is weighted in decisions.
Uhrig <i>et al</i> (2006)	Healthcare product information  (Effect on decision-making)	Experimental study testing the impact of different information formats in healthcare plans. It was found that presenting the same information using different formats affected consumers' choices.

Peters <i>et al</i> (2007)	Healthcare performance information  (Effect on decision-making)	Results of three studies support the idea that 'less is more' when presenting consumers with comparative performance information on which to make hospital choices.  The study found that the chosen data display format will have a significant influence on which information is attended to and used. It was found that better choices were generally made when the data display format was designed to ease the cognitive burden and emphasise the meaning of important information.
NHS (2018)	Healthcare performance information  (Effect on decision-making)	Different data presentation formats affect decision-making. SPC charts are recommended as they assist effective decision-making by displaying contextualised data, whilst simplistic formats (such as binary comparisons) are criticised for being misleading.
Hibbard <i>et al</i> (2001)	Healthcare performance information  (Effect on interpretation / comprehension)	Data display formats which appear complex tend to be under-used. However, some presentation strategies that involve simplifying performance information (e.g. star ratings) assist users with lower skills to comprehend and use information more effectively. The study also found that evaluative labels appear to aid those in the midrange of comprehension ability.
Gerteis <i>et al</i> (2007)	Healthcare performance information  (Effect on interpretation / comprehension)	Experimental study testing the impact of different data display formats on users' comprehension. Seven reporting templates in different formats were shown to participants and it was found that different presentation formats significantly affect users' comprehension of the data.
Damman <i>et al</i> (2011)	Healthcare performance information  (Effect on interpretation / comprehension)	Experimental study investigating the effects of using various types of performance information and different designs of visual components (bar charts, star ratings, ordering of data, type of stars, number of stars and inclusion of a global rating). The findings indicate that the type of performance information and presentation features affected interpretation and use.
Brewer <i>et al</i> (2012)	Healthcare performance information  (Effect on interpretation / comprehension)	Compared to horizontal bar graphs, tables required more time and experience to interpret, suggesting that tables can be a more burdensome format to use than graphical formats.

Kurtzman and Greene (2016)	Healthcare performance information  (Effect on interpretation / comprehension)	<p>Numeric formats that include considerable text are generally not as effective as simple formats employing straightforward graphs or familiar icons.</p> <p>Graphical displays are generally more beneficial than numeric formats and consumers tend to better understand and make more informed choices when the information display is less complex.</p> <p>The choice of data display can also have a significant impact on how well consumers understand comparative information and affect health care choices.</p>
----------------------------	--	--

Although these studies are sited within the healthcare domain, it is proposed this thesis speaks directly to this school of research. Principally however, whilst the studies explore the impact of data display on interpretation and decision-making, this thesis aims to extend the reaches of prior research by assessing whether there is also a significant relationship between data display and behavioural dysfunction.

Furthermore, this study focuses specifically on the design characteristics of numeric police performance information, by considering how the use of reference points can frame assumptions about performance. For instance, when performance information users perceive deficiencies due to variance from reference points such as a numerical targets, how does this shape their perceptions? To answer this, the following section explores *reference dependence* theories, to help explain why interaction with certain formats can activate cognitive processes that decode performance information in terms of variance from temporal, social, or aspirational reference points.

### 2.2.3 Reference Dependence

When it comes to understanding the mechanics involved in the interpretation of numeric performance information, the concept of *reference dependence* (Tversky and Kahneman, 1986; 1991; Quattrone and Tversky, 1988; Kahneman, 1992; Greve, 1998; Heath *et al*, 1999; Garcia *et al*, 2006; Olsen, 2013b) may help shed light on the thought processes that arise when performance information users engage with simplistic formats; it could also provide insights into why such formats have gained currency in the field of performance information use.

Reference dependence is a well-established facet of Prospect Theory (Kahneman and Tversky, 1979), which proposes performance losses or gains are expressed in terms of variance from a reference point. Kahneman (1992) asserts, “Reference points are important because other outcomes are compared to them, and are coded and evaluated in terms of this comparison” (1992, p.296), whilst Heath *et al* (1999) posit “...whenever a specific point of comparison is psychologically salient, it will serve as a reference point” (1999, pp.105-106). This accounts for the operation of binary comparisons, as well as other forms of performance information that define ‘success’ or ‘failure’ by comparison to a relative position.

Reference points are used because of their perceived heuristic value (James, 2011); they act as a judgemental shortcut for when individuals do not have the time, inclination or resources to consider performance information more deeply (Olsen, 2013b). Lack of absolute information, bounded rationality (Simon, 1955; 1978) and the cognitive limitations of the human brain (Quattrone and Tversky, 1988; Birnberg *et al*, 2009) mean individuals often prefer to judge performance by comparing a current value to a reference point. This may be temporal, social or aspirational, as reference points can be influenced by “...aspirations, expectations, norms and social comparisons” (Tversky and Kahneman, 1991, p.1047).

Therefore, a key implication of reference dependence is that the same piece of information may be interpreted differently depending on its relative distance and direction from the reference point. For example, depending on the position of a current value in relation to a datum pertaining to ‘the same time last year’, this can influence perceptions about whether performance is improving or deteriorating. In such circumstances, the performance information user perceives a linear trajectory from the historic reference point towards the current value and formulates a perception based on any difference between the two.

The general notion of reference dependence is a useful expository vehicle for helping to understand the operation of formats that define performance trajectories by depicting variance from reference points; however, binary comparisons, league tables and numerical targets can each be further examined through the lens of one of a subset of theoretical perspectives within this field.

In respect of binary comparisons, Albert's (1977) *Temporal Comparison Theory* is most relevant, as it proposes individuals often compare their present identity or attributes to historical reference points. Albert argues that if individuals compare current self-descriptions with dissimilar previous ones "...they will perceive themselves as undergoing change" (Albert, 1977, p.491). This process is undertaken with the intent of understanding self-progression and realising predictive utility:

"Temporal comparison between past and present is likely to occur whenever the individual desires to predict the future" (Albert, 1977, p.500).

Applying these principles to the performance context, Olsen (2013b) observes, "Historical reference points denote a temporal comparison between current and previous performance" (2013, p.3). Similarly, James and Moseley (2014) and Hansen *et al* (2015) note citizens routinely draw on such reference points to make comparisons, although Charbonneau and Van Ryzin (2015) found that comparisons against peers or an average were considered more salient than temporal reference points when making judgments about performance.

Nevertheless, it is suggested those who use binary comparisons to interpret performance information do so believing they are a useful method for identifying change and anticipating future trajectories; this applies whether the comparator datum is an historic value or an average (Moore and Klein, 2008). The principles of reference dependence and temporal comparison theory therefore help explain how assumptions are formulated when performance is depicted as binary comparisons.

With regard to the implications of reference dependence for league tables, *Social Comparison Theory* (Festinger, 1954; Goethals and Darley, 1977; Miller, 1983; Goethals, 1986; Garcia *et al*, 2006; Corcoran *et al*, 2011; Olsen, 2013b) provides relevant insights. Festinger (1954) argues there is a fundamental urge within the human condition that drives a search for information about status amongst peers, whilst Miller (1983) talks of an innate drive to make comparisons with those who are 'salient and available'. Corcoran *et al* (2011) propose:

"Social comparisons are a fundamental psychological mechanism influencing people's judgments, experiences, and behavior" (2011, p.119).

Therefore, when it comes to performance, individuals naturally compare themselves with others, particularly those possessing similar characteristics (Goethals and Darley, 1977; Garcia *et al*, 2006). Ammons (1997) argues such comparative benchmarking is more useful than reflexive benchmarking against historic performance data, whilst Hansen *et al* (2015) found that performance information users ascribe substantial weight to differences between peers when social reference points are utilised to present data.

In ranking systems, relative positions of peers act as reference points, with competition intensifying in proximity to the top or bottom of a league table, or close to thresholds for rewards or sanctions (Garcia *et al*, 2006). Therefore, it could be argued that ranking systems could act as a catalyst for performance improvements; indeed, Moynihan and Landuyt (2009) suggest it is beneficial to make comparisons between organisations to obtain continuous feedback on performance and identify opportunities to improve.

However, it may be the case that league tables invite flawed assumptions about relative performance as they are constructed using fixed hierarchical reference points. Due to the equal distances between ranks, marginal differences in outputs between work units (or naturally-occurring variation) can be overlooked. Even where numeric outputs are published alongside ranked positions, neat separation of work units creates a kind of forced distribution where half of those ranked are ‘below average’, and someone is always bottom. Overall, this can create the impression of ‘good’ or ‘bad’ performance, defined by a work unit’s position alone.

Furthermore, there can be risks to individuals’ psychological wellbeing if they are labelled as poor performers on the basis of peer comparisons. Pettigrew (1967) proposed that when humans compare themselves to peers, this results in positive, negative or neutral self-ratings. Individuals therefore feel worse after comparing their performance to an apparently superior peer than when they compare it to an inferior peer (Gilbert *et al*, 1995). This affects self-esteem (Morse and Gergen, 1970), meaning that negative comparisons can be a determinant of feelings of resentment and deprivation (Goethals, 1986).

A further foray into the reference dependence literature finds limited but pertinent content that helps explain how numerical targets act as reference points. In a similar

fashion to temporal comparisons that hinge upon an isolated historic value, numerical targets create a precise aspirational value against which performance is judged. As stated by Heath *et al* (1999):

“When people set a goal, their goal acts as a reference point and systematically alters the value of outcomes” (1999, p.95).

Consequently, the reference point (i.e. target) becomes a dividing line between ‘good’ and ‘bad’, being intended to “...simplify evaluation by transforming a continuous measure of performance into a discrete measure of success or failure” (Greve, 1988, p.59; see also March, 1988). As a result, those who are below target will perceive current performance as a loss relative to the reference point, which can affect the cognitive processes that guide individuals’ choices (Heath *et al*, 1999).

In such circumstances, loss is felt more keenly due to *negativity bias* (see Lau, 1985; Baumeister *et al*, 2001; Rozin and Royzman, 2001; James and John, 2007; Olsen, 2013c) and *loss aversion* (see Tversky and Kahneman, 1986; 1991; Quattrone and Tversky, 1988; Kahneman, 1992); it therefore follows that failure to hit targets, outperform peers, or exceed prior levels of performance, could induce undue concern and feelings of failure.

This occurs as negativity bias causes greater weight to be ascribed to comparisons where the current value falls below the reference point, as “...negativity bias implies that individuals pay more attention to negative information than positive information of the same magnitude” (Olsen, 2013c, p.2). Furthermore, loss aversion dictates “...losses (outcomes below the reference state) loom larger than corresponding gains...” (Tversky and Kahneman, 1991, p.1047), thereby potentially causing disproportionate focus on perceived poor performance.

Although loss aversion is relevant to temporal and social comparisons, the aspirational nature of numerical targets puts success or failure into even sharper focus. Therefore, it may well exert greater influence than in temporal or social contexts, even where performance targets are not linked to extrinsic rewards (Heath *et al*, 1999); indeed Greve (1998) warns explicitly “...aspiration levels have behavioral consequences” (1998, p.80). It could therefore be that the strict ‘pass’ /

‘fail’ configuration of numerical targets is responsible for feelings of failure and perhaps even contributes towards some of these behavioural consequences.

Overall, reference dependence theories help to explain the functioning of cognitive processes that affect data interpretation. Similarly, the concepts of negativity bias and loss aversion act as useful explanatory antecedents to the operation of these reference dependence theories. Therefore, the theoretical insights they provide will be used to inform this study’s findings and gain a deeper understanding of how performance information users interact with particular data display formats.

#### **2.2.4 Statistical Process Control**

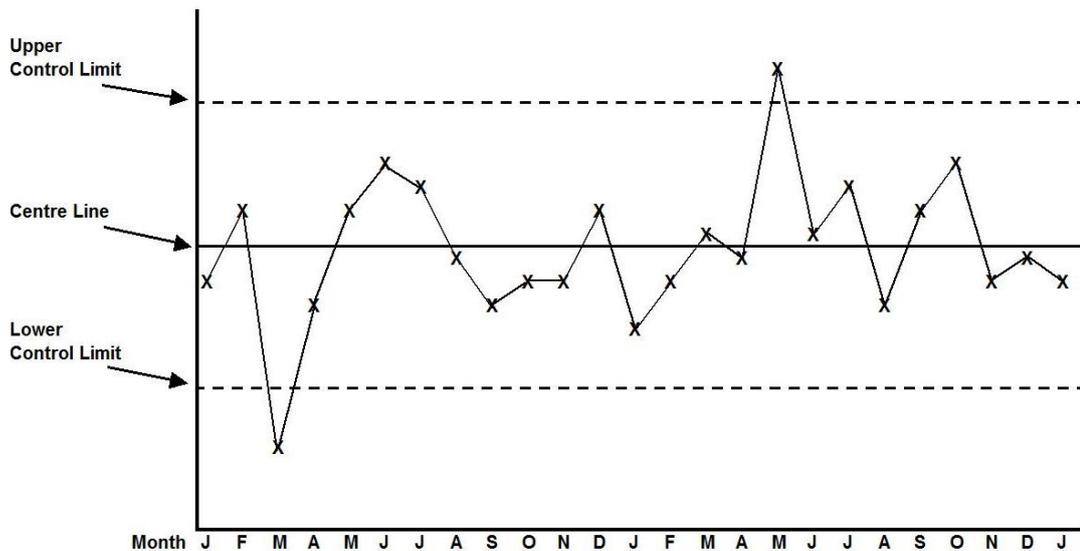
In contrast to reliance on isolated reference points, other methods of presenting numeric data promote greater accuracy, by incorporating multiple reference points and conforming to established principles of statistical rigour. One such method is Statistical Process Control, or SPC (see Shewhart, 1939; Nielsen, 2015). Despite its origins and most common application being in engineering and the manufacturing industries, SPC methodology can also be applied within the field of public sector performance information use (NHS, 2018).

A major advantage of SPC charts is that they allow users to differentiate between normal statistical variation and ‘signals’. Terminology varies, but the former refers to non-significant values in a data set that fall within an expected range, whilst the latter denotes outliers or unusual patterns. Variation is inevitable, being caused by a multitude of factors, often outside of human control (Deming, 1994). Therefore, there is no merit in attempting to ascribe meaning to differences between data points that fall randomly within the expected range. As stated by Wheeler (1998):

“While all data contains noise, some data contain signals. Before you can detect a signal within any data set you must first filter out the noise” (1998, p.31).

An example of a SPC chart is displayed at Figure 2.19.

Figure 2.19: Example of a SPC chart <sup>6</sup>



The control limits are usually positioned at either two or three *standard deviations* (see Field, 2013, pp.27-28) from the mean average (centre line) and demarcate the anticipated range of variation with 95% or >99% confidence, respectively. <sup>7</sup> Crucially, the control limits are determined by calculations involving the specific data set; if artificial boundaries are inserted, or control limits are set within two standard deviations, this risks non-significant values being interpreted as signals.

In this case, all points except those outside of the control limits indicate normal variation; therefore the only areas where it is appropriate to further explore data are the two outliers. Whilst breaches of the control limits are the most obvious type of signal, others can occur between them. These include:

- Six or more consecutive points in the same direction;
- Eight or more consecutive points on the same side of the centre line;
- 14 or more consecutive alternating points, and;
- Three or more consecutive points closer to a control limit than the centre line.

(Joiner, 1994; Wheeler, 2000).

<sup>6</sup> Note: numeric values have been omitted from the vertical axis for simplicity of presentation.

<sup>7</sup> For methodology on constructing SPC charts, see Joiner, 1994, pp.148-149.

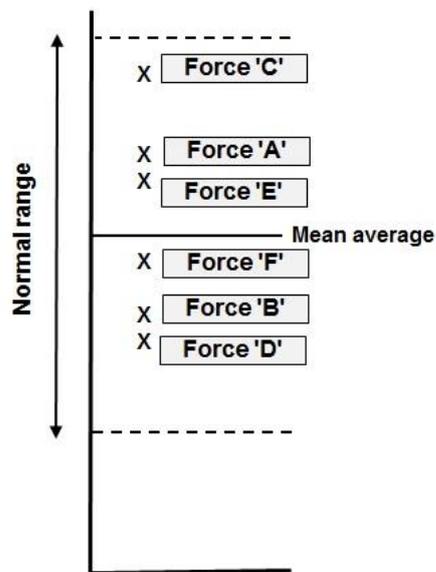
Where such patterns are identified, there is a mandate to inquire further into reasons why the data present themselves in this manner. Clearly, this method of displaying data is useful for preventing unwarranted assumptions about non-significant differences. Rather than making a comparison against a single reference point, SPC charts direct attention towards issues of genuine concern; they also provide users with the confidence not to ‘knee-jerk’ to non-significant differences between values that are simply reflective of normal variation. For this reason, Joiner (1992) argues:

“One necessary qualification of anyone in management is to stop asking people to explain ups and downs (day to day, month to month, year to year) that come from random variation” (Joiner, 1992; cited in Deming, 1994, p.216).

A further advantage of SPC charts is that they enable prediction. For example, where control limits are set at two standard deviations, there is 95% probability the data will continue to populate within the parameters. Therefore, users are informed with high degrees of confidence whether data are stable and what is likely to happen next. The objective is then to continuously improve performance, producing a long-term trajectory in the desired direction (Deming, 1986).

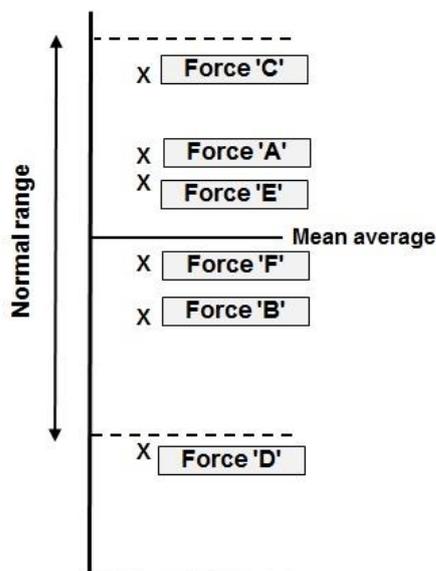
However, not only are SPC principles useful for presenting time series data, but they can also be used to construct statistically robust data display formats for making contextualised comparisons between peer groups. Examples are provided at Figures 2.20 and 22.1, below:

Figure 2.20: Contextualised peer comparison chart 1<sup>8</sup>



In Figure 2.20 it can be seen that all six forces fall within the normal range. Therefore, if the control limits were set at two standard deviations, this provides 95% confidence that no force is significantly different from its peer group. Therefore, it is inappropriate to draw inferences from relative positions, as only normal variation is present. Conversely, in Figure 2.21, Force 'D' sits outside the expected range and is therefore an outlier. This provides a mandate to scrutinise potential reasons for this significant difference from peers.

Figure 2.21: Contextualised peer comparison chart 2



<sup>8</sup> As with the previous SPC chart diagram, numeric values are omitted for simplicity.

Overall, it is suggested that data display formats utilising multiple reference points and / or founded on SPC methodology may improve the accuracy of interpretation and mitigate the likelihood of unwarranted assumptions being made. Whilst such formats are less commonplace in UK police performance management systems, it seems that their wider use may present an opportunity to insulate users against some of the common errors associated with simplistic formats.

In summary, reference dependence theories, along with negativity bias and loss aversion, offer insights into psychological and cognitive undertakings relevant to engagement with numeric data display formats. These insights could ultimately aid understanding of the way in which different types of police performance information are coded and interpreted; ultimately, they may also shed light on whether initial engagement with reference-dependent forms of performance information is a factor ultimately capable of influencing dysfunction.

### **2.3 Chapter Two: Summary and the Imperative for Further Research**

The literature review identifies an apparent association between certain types of police performance information and behavioural dysfunction, as well as consistency in the types of dysfunction observed. It also highlights factors that could potentially act as antecedents to dysfunction; these tend to be organisational conditions, along with improper use of performance information or overbearing conduct by managers.

Examples include the presence of overly-intrusive audit and inspection regimes (Loveday, 1998), or an organisational disposition towards authoritarianism (Willis *et al.*, 2004; 2007). Other possible antecedents involve managers demanding answers for perceived poor performance (Eterno and Silverman, 2012), holding subordinates personally accountable for performance levels (Behn, 2001; Collier, 2006; Hood, 2006), as well as the practice of ‘naming and shaming’ those considered to be ‘poor’ performers (Le Grand, 2010).

Similarly, dysfunction seems to arise where managers insist on explanations for apparent deficiencies, even when variance within performance data may not be significant (see Joiner, 1992; Moore and Braga, 2003), or where subordinates have little direct control over performance levels, such as in the case of crime rates (Bayley, 1994; Coleman and Moynihan, 1996). Similarly, their demands for

improvement, when based on flawed interpretation of data, also seem capable of inducing adverse reactions (Eterno and Silverman, 2012; Home Office, 2015).

Additionally, where managers' initial interpretation of data suggests deterioration or otherwise poor performance, there can be a tendency for them to initiate unnecessary operational activity (Home Office, 2015). This may well trigger perverse responses on the part of subordinates, but could also be considered dysfunctional in its own right, as such activity damages organisational cohesion and effectiveness (Ridgway, 1956; Deming, 1986).

However, whilst such conditions may suggest possible explanations for dysfunction, they do not explain *why* managers seem to react in such a consistent fashion after engaging with particular types of performance information. Therefore, it seems appropriate to investigate whether the data display format itself may influence how managers respond. Certainly, studies into the effects of data display indicate it can affect comprehension and decision-making; however, it is unknown if its influence could also extend to affecting behavioural dysfunction.

Consequently, it is proposed there is a substantial imperative for deeper research into the effects of data display upon the design and use of performance information and specifically whether it behaves as a direct antecedent to dysfunction. This study will therefore examine UK police officers' experiences of the use of various performance information formats to identify common themes and ascertain whether they are reflective of those within the literature; it will also employ experimental methods to assess cognitive and behavioural responses to different formats, aiming to establish if data display can exert a significant influence on the phenomenon of interest.

The research question is therefore:

*“Does data display influence the likelihood, nature or extent of behavioural dysfunction in police performance management systems, and if so, why?”*

## Chapter Three

### Philosophical and Methodological Foundations

#### 3.1 Introduction

This chapter first sets out the philosophical foundations that underpin the research, exploring the underlying ontological and epistemological assumptions. It then discusses the chosen methodological position, as well as the methods utilised to generate findings capable of providing an explanation for the phenomenon of interest.

Additionally, it details the construction, validity and reliability of a large-scale survey instrument used to collect quantitative and qualitative data, which aims to identify key themes relating to the use of performance information in UK police forces. The instrument also contains an experimental component, designed to assess cognitive and behavioural responses to various data display formats; this, too, is comprehensively explored.

#### 3.2 Critical Realism: Background and Key Characteristics

The underpinning philosophical position underpinning this research is *critical realism* (see Bhaskar, 1975; 1979; 1986; 1998a; Wilson, 2006; Fleetwood, 2013). Critical realism interrelates ontology and epistemology, whilst aiming to explain ‘why what happens actually happens’ (Danermark *et al*, 2002). It offers an ontology that recognises the existence of a world independent of researchers’ knowledge, alongside a fallibilist epistemology in which knowledge is socially produced (Miller and Tsang, 2010). It asserts theories are fallible and no knowledge is 100% certain (Longshore-Smith, 2006). Bergin *et al* (2008) conclude:

“Critical realism is first and foremost concerned with ontology and starts from questions about what exists... a philosophy of reality must begin with a theory ‘of being’ (ontology) as distinct from a theory ‘of knowledge’ (epistemology)” (2008, p.170).

Critical realism originated during 1960s Britain, through the work of Roy Bhaskar and his mentor Rom Harré. Bhaskar’s views were, and remain, largely contrary to

some elements of more established philosophical traditions, being critical of strong forms of positivist reductionism, alongside equally dogmatic strains of interpretivism and constructivism (Gorski, 2013). It differs from positivism by asserting that a constant conjunction of variables does not provide causal explanation, whilst also criticising ‘anything goes’ interpretivist approaches that are unclear about by what standards one explanation is judged better than another (Easton, 2010).

Instead, critical realism “...combines a modified naturalism with the necessity of interpretive understanding of meaning in social life” (Sayer, 2000, p.3). Critical realists reject judgmental relativism (i.e. the notion that multiple theories or interpretations are equally valid), whilst endorsing epistemological relativism and judgmental rationality (i.e. in principle, it is possible to make contingently reasoned and accurate judgments about reality) (Lipscomb, 2011; Bygstad and Munkvold, 2011). It also rebuts the proposition of a sharp divide between the natural and social sciences (Bhaskar, 1979); this resonates with the concept of an ‘objectivist-subjectivist continuum’ (Burrell and Morgan, 1979).

### **3.2.1 Reality and Knowledge**

Critical realism makes a fundamental distinction between two dimensions of knowledge; the intransitive and the transitive (Bhaskar, 1998a; Bergin *et al*, 2008). The intransitive dimension comprises real entities constituting the social and natural world (Outhwaite, 1987). Science attempts to comprehend the intransitive dimension through socially-produced transitive theories, models and paradigms, which are fallible (Bhaskar, 1998a; Bergin *et al*, 2008; Gorski, 2013); knowledge is therefore historically and socially located (Mingers, 2002), meaning its production occurs in the transitive dimension (Bhaskar, 1989).

As critical realists believe science is a social product (Sayer, 1992), it follows that whilst theories, models and paradigms may evolve or challenge each other, the reality they attempt to interpret does not necessarily change (Collier, 1994; Sayer, 2000). Therefore, when theories change it does not mean what they are about changes too; as Sayer (2000) observes, “...there is no reason to believe that the shift from a flat earth theory to a round earth theory was accompanied in a change in the shape of the earth itself” (2000, p.11).

However, access to an external reality is affected by social factors. Critical realists argue some phenomena are entirely socially constructed (e.g. ‘disability’), but refute constructivist claims that there exists no external reality independent of us, or that we cannot gain any reliable knowledge about it (Danermark, 2001). Consequently the critical realist position is that whilst interpretive understanding is a necessary and important feature of social science, this does not preclude scope for causal explanation (Sayer, 2000).

Furthermore, critical realism conceives reality as differentiated and stratified; a perspective that demarcates it from other ontologies. Most of reality is unobservable (Bunge, 2004), comprising three domains of the social and natural world; these are the *real*, the *actual* and the *empirical* (Bhaskar, 1998a). The ‘real’ domain is all that exists, whether or not it is known or experienced (Sayer, 2000). The ‘actual’ domain “...refers to what happens in reality when the powers or mechanisms of the real are activated and events or experiences are produced” (Bergin *et al*, 2008, p.171; see also Danermark *et al*, 2002). The ‘empirical’ domain is comprised solely of what is experienced (Collier, 1994).

### **3.2.2 Entities, Structures and Mechanisms**

Bhaskar (1998a) posits the real world is populated by *entities*, which have *powers* (to act and be acted upon) and *liabilities* (i.e. limitations / constraints). These entities (or ‘objects’) are the basic building blocks for critical realist explanation and may be physical (e.g. atoms / organisms), social (e.g. market forces), or conceptual (e.g. categories / ideas) (Bhaskar, 1997). Other examples could be organisations, people, attitudes, resources, wars, identities, or economic activities (Sayer, 2000, p.19). Easton (2010) asserts, “Entities can be tangible or intangible, social or physical, dormant or active” (2010, p.125); therefore, tangible entities could be performance documents, whilst intangible entities could include ‘performance culture’.

Entities are usually structured and can possess internal structures, which in turn have their own powers. However, such powers can exist unexercised (e.g. the power of workers to ‘work’, but who choose not to) (Sayer, 2000). Structures exist *in* systems and structures *are* systems (Gorski, 2013), being “...a set of internally-related objects or practices” (Sayer, 1992, p.92). For example, organisations may comprise various entities (e.g. departments, processes, or resources), whilst people also exhibit

structures, such as personal characteristics, gender or psychological structures (Easton, 2010). Structures “...enable what *can* happen through the workings of their mechanisms in geo-historical contexts” (Longshore-Smith, 2006, p.202).

Although pre-existing structures exert causal influences on actors “...this does not mean that the behaviour of actors is determined by social structure” (Lewis, 2002, p.20). Structure and agency are each a condition for and consequence of the other; agency is the capacity of actors to act, whilst structure reflects the conditions within which agents operate, such as social norms, rules, or language. Therefore, social activity occurs within contexts provided by pre-existing social structures, where individuals are enabled or constrained by structures, but still maintain a degree of voluntarism (Lewis, 2002; Bygstad and Munkvold, 2011). However, the presence of particular social structures militates in favour of certain actions (Lewis, 2002).

Critical realists believe events involve *necessary* and *contingent* relationships (Easton, 2010, p.121). The former occurs where entities are mutually dependent; the latter is when relationships exist but are not necessary. However, whilst independent elements may combine to produce a new feature or experience “...they cannot be reduced to their components only, even though their components are necessary for their being” (Bergin *et al*, 2008, p.171). Sayer (2000) cites the example of water to illustrate this - although it possesses causal powers to extinguish fire in its ‘whole’ form, its constituent components (hydrogen and oxygen) in isolation do not account for its properties; indeed they actually accelerate combustion.

The continual operation and interaction of entities within particular structures therefore gives rise to certain causal powers, tendencies, or ways of acting, called *mechanisms* (or ‘generative mechanisms’) (Bhaskar, 1979, p. 170). Machamer *et al* (2000), define mechanisms as follows:

“Mechanisms are composed of both entities (with their properties) and activities. Activities are the producers of change. Entities are the things that engage in activities” (2000, p.3).

Therefore, the interaction of mechanisms causes the presence or absence of events (Mingers, 2002). Bhaskar (1975) writes:

“The world consists of mechanisms not events. Such mechanisms combine to generate the flux of phenomena that constitute the actual states and happenings of the world” (1975, p.106).

Mechanisms are simply “...ways in which structured entities by means of their powers and liabilities act and cause particular events” (Easton, 2010, p.122). Mechanisms operate in the real domain and are “...nothing but the way of acting of things” (Bhaskar, 1975, p.14). Highly complex systems (e.g. living cells, schools) can exhibit several concurrent mechanisms at different levels (Bunge, 2004, p.193) and different effects may be caused by the same mechanisms (Danermark, 2001).

Mechanisms are system-specific (Bunge, 2004) and their outcomes are context-dependent (Bygstad and Munkvold, 2011); examples of mechanisms could include electrical and chemical signals, division of labour, publicity, or military expeditions (Bunge, 2004, p.191). Mechanisms can reinforce each other (Danermark *et al*, 2002), operating upward and downward across levels, or laterally within the same level (Anderson *et al*, 2006; Miller and Tsang, 2010).

Furthermore, there are micro and macro mechanisms; for instance, some mechanisms operate at the molecular level, whilst others operate at biological, psychological or social levels (Danermark, 2001). Events are simply conjunctures of all the mechanisms operating in a given situation (Bhaskar, 1998b); they are the external and visible behaviours of people, systems and things (Easton, 2010).

### **3.2.3 Causation**

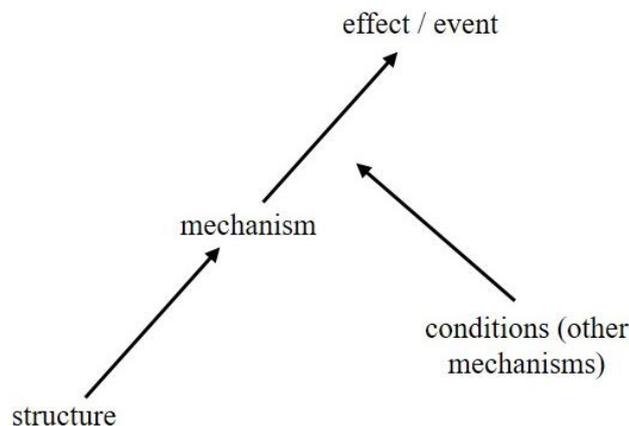
Critical realism disapproves of the Humean ‘successionist’ notion of causation (Sayer, 2000; see also Hume, 1967). Instead, it proposes an alternative conception of science founded on a non-deterministic notion of causality that seeks to answer underlying ‘why’ questions, whilst focusing on meanings, interpretation and context (Longshore-Smith, 2006). Whilst positivists concentrate on cause-and-effect relationships, critical realists “...redefine laws so that they are better understood as explanatory and non-predictively normic” (Bergin *et al*, 2008, p.173).

Secondly, whilst constructivists understand causality as a non-linear process of constructing meaning from complex individualised social realities (Appleton and King, 2002), critical realists locate causal relationships at the level of mechanisms,

whether or not events are observed or detected (Tsang, 2006). Furthermore, they conclude generalisations should be described as *tendencies* rather than predictions (Bhaskar, 1998b; Longshore-Smith, 2006). These tendencies are located at the level of the real, yet may be observed at the empirical level (Longshore-Smith, 2006).

Despite differences with other philosophical positions on causation and prediction, critical realism acknowledges mechanisms and laws are not incompatible, as “...any mechanism unsupported by some law(s) must be regarded *ad hoc*...” (Bunge, 2004, p.199). Therefore, a critical realist account of causal explanation “...is one that identifies entities and the mechanisms that connect them and combine to cause events to occur” (Easton, 2010, p.122). This may be depicted visually, as below:

*Figure 3.1: Critical realist view of causation*



(Sayer, 2000, p.15)

The critical realist position on causation also enables the production of a narrative explicating the objects, structures, causal powers and liabilities that combine to activate mechanisms responsible for producing events. For example:

*Entities* (a salesperson) having *structures* (knowledge and personality traits) and necessarily possessing *causal powers* (to persuade a buyer, who is another entity) and *liabilities* (to be rejected by prospective buyers / to become tired) will, under different specific conditions (such as whether the buyer has a need for the product), produce different types of *events*.

(Adapted from Easton, 2010, pp.121-122).

Crucially, such an account goes beyond describing *what* happens, to postulating mechanisms capable of explaining *why* an observed phenomenon occurs. As stated succinctly by Fleetwood (2013):

“We don’t explain why the bus is late today by stating that it is always, or regularly, late” (2013, p.35).

### **3.2.4 Retroduction**

Critical realist explanation occurs through a process of *retroduction* to possible causes (Bhaskar 1998c, xvii). Retroduction is a “...mode of inference in which events are explained by postulating (and identifying) mechanisms which are capable of producing them...” (Sayer, 1992, p.107). It involves ‘moving backwards’ from a phenomenon of interest to a conception of a thing (e.g. power or mechanism) that could have caused it (Easton, 2010). Retroduction is akin to ‘conjecturing’ (Bygstad and Munkvold, 2011) and may be likened to an art rather than a technique; this is because mechanisms are typically unobservable (Bunge, 2004, p.200).

Hodgkinson and Starkey (2011) define retroduction as “...a logic of exploration, starting with tentative hypotheses that are explored until they lead to new practicable ideas” (2011, p.363); these hypotheses can explain (or lead to) other hypotheses (Bylander *et al*, 1991). It is analogous to the pragmatist notion of *abduction* (Peirce, 1931), as it helps explain events by proposing “...hypothetical mechanisms that, *if they existed*, would generate or cause that which is to be explained” (Mingers, 2002, p.300). Thagard and Shelley (1997) observe:

“Many important kinds of intellectual tasks, including medical diagnosis, fault diagnosis, scientific discovery, legal reasoning, and natural language understanding have been characterized as abduction” (1997, p.1).

Like pragmatism (see Peirce, 1878; 1905; Dewey, 1931; Cherryholmes, 1992; 1994; House, 1992; Menand, 1998; Shields, 1998; Hookway, 2008; Talisse and Aikin, 2011; Bacon, 2012), critical realism is *performative*; that is, although critical realists assume there is a real world ‘out there’, they accept this assumption cannot be proven or disproven; nevertheless, they behave as if the world was real and generally this supposition works (Easton, 2010, p.119).

### 3.2.5 Critical Realism and Statistical Analysis

The critical realist quest for deeper explanation has generated much criticism of statistical methods. For example, Bhaskar (1998d) is dismissive of ‘statistical models of explanation’ (1998d, p.225), whilst others argue statistical analysis:

- Assumes one-way, linear, conjunctive, Humean causality;
- Provides an ‘impoverished and empiricist’ viewpoint with weak claims to causation;
- Lacks ontological depth;
- Disregards the impact of unobservable and / or unmeasurable variables, and;
- Assumes closure, when it can be difficult to induce this in open systems / social contexts.

(See Fildes, 1985; Lawson, 1997; Pawson and Tilley, 1997; Olsen, 1999; Ron, 1999; Mingers, 2003; 2006).

Additionally, null-hypothesis significance testing is criticised for confusing statistical significance with theoretical or practical significance, as results may be statistically significant but have tiny, or unreported, effect sizes. Furthermore, the 95% significance threshold implies an arbitrary ‘all or nothing’ cut-off point. As Mingers (2006) quips:

“4.9% is significant – hooray, publish – 5.1% is not” (2006, p.211).

However, others insist statistical analysis can be legitimately applied in critical realist research, particularly for identifying *patterns* (or potential ‘demi-regularities’ – see Pratschke, 2003, pp.24-25; Lawson, 1998, p.149) amidst data that may indicate the presence of underlying generative mechanisms (Mingers, 2006; Miller and Tsang, 2010; Mingers *et al*, 2013). Indeed, Pratschke (2003) argues, “Mathematical ‘formalisation’ can actually *enhance* theoretical clarity by spelling out the empirical consequences of a theoretical hypothesis” (Pratschke, 2003, p.22).

Indeed, statistical analysis can be extremely useful as long as it is not used to claim a causal explanation in its own right (Danermark, 2001, p.10; Mingers, 2006). Porpora (1998) asserts “...analytic statistics are not explanatory tools at all” (2003, p.4),

concluding the issue is not one of legitimacy *per se*, but rather that positivism mistakenly conflates evidence and explanation. Fundamentally, it is argued:

“...statistical analysis should not be thrown out entirely but can be utilized critically within a practical and realistic framework...” (Mingers, 2006, p.203; see also Lawson, 1997; Layder, 2003).

It may even be possible to approximate social conditions in experimental settings (i.e. to induce artificial closure on part of an open system), in order to test subjects' behaviour when confronted with different variables (Mingers, 2002). However, it must be recognised that variables are measures of things and not the things themselves - therefore they can only register quantifiable change and not its cause (Sayer, 1992; Easton, 2010). Consequently, whilst there are merits to this approach, caveats apply (Harré and Secord, 1972; Caldwell, 1984).

### **3.2.6 The Benefits of a Critical Realist Account**

The critical realist notion of a stratified reality, along with its distinction between the intransitive and transitive dimensions of knowledge, allows for an interpretation of reality that respects theories of natural science, whilst acknowledging the socially-constructed nature of much of what is known. It presents a “...point of entry into epistemology and metaphysics for practicing social scientists” (Groff, 2004, p.23) capable of reconciling the divide between social theory and empirical research (Carter, 2000). Sayer (2000) asserts:

“...critical realism provides an alternative to both hopes of a law-finding science of society modelled on natural science methodology and the anti-naturalist or interpretivist reductions of social science to the interpretation of meaning” (2000, pp.2-3).

Critical realism therefore offers a solid ontological basis for theoretically sound and highly relevant research, as it promotes “...scientific understanding of generative mechanisms for knowledge creation, [and] a pragmatic concern for effectiveness (‘does it work?’) rather than ‘truth’ (‘is it true?’) as a guiding research principle...” (Hodgkinson and Starkey, 2011, p.363). Its pragmatic vein supports “...rigorous research that describes and evaluates what is going on in practice” (Markus, 1997, p.18; see also Goles and Hirschheim, 2000), being particularly suited to management

research due to the “...multi-level and socially constructed character of organizations...” (Hodgkinson and Rousseau, 2009, p.540).

This study therefore follows critical realist research principles in order to generate theory that possesses explanatory utility *and* incorporates a practical dimension (see Wacker, 1998). This dual objective is consistent with the views of Suddaby *et al* (2011), who assert “...researchers should focus explicitly on what practitioners actually do...” (2011, p.243). Similarly, drawing on Zikmund *et al* (2008), Gay and Weaver (2011) ask:

“In short, what is the purpose of theory if it cannot bridge the research-practice gap and thus, has no practical application to the real world?” (2011, p.30).

Overall, it is argued critical realism provides a firm philosophical foundation for this doctoral research, which utilises a four-step model proposed by Easton (2010):

1. Developing a research question that identifies a phenomenon of interest, in terms of discernible events, and asking what causes them to happen.
2. Provisionally identifying key entities and structures, their powers and liabilities, as well as necessary and contingent relationships.
3. Capturing data and asking why events happened / are happening, whilst taking into account the issues associated with interpreting empirical data.
4. Identification of one or more mechanisms that could have caused the events.

(Adapted from Easton, 2010, p.128).

Easton’s framework translates as follows for this research:

1. *Phenomenon of interest:* Behavioural dysfunction within police performance management systems.
2. *Entities and structures:*
  - a. UK police service (which comprises entities and socially-constructed structures, such as hierarchies, relationships, procedures, norms, departments, people, processes, resources, culture).

- b. Users of police performance information who have their own structures (e.g. knowledge, personal characteristics, psychological structures) and possess individual agency, yet are influenced by specific conditions (i.e. UK policing structures).
  - c. Performance information of varying formats, which may possess powers and liabilities that influence users, thereby contributing towards the likelihood of particular tendencies being exhibited.
3. *Data*: Review of literature and additional documentation, along with quantitative and qualitative data generated by a survey instrument.
  4. *Mechanisms*: To be conjectured through retrodution.

This approach represents a robust and coherent philosophically-underpinned framework that is consistent with the study's overarching methodological position, as well as the methods employed. These are outlined below.

### **3.3 Methodology and Methods**

The following section examines methodological considerations pertinent to this study. It first considers how the philosophical position influences methodology and discusses approaches that were considered but discounted. It then sets out a rationale for chosen methods, as well as providing an overview of the analytical strategy. Finally, it explores the design and implementation of the large-scale survey instrument forming the backbone of the research.

#### **3.3.1 Methodological Framework**

This study employs *mixed methods* (see Creswell, 2008; 2009; 2013). Critical realists advocate blending quantitative and qualitative methods (Sayer, 2000; Mingers, 2001; 2002; 2003; Rousseau *et al*, 2008; Hodgkinson and Rousseau, 2009) and assert deductive, inductive and retroductive approaches may legitimately be employed within a mixed methods framework (Danermark *et al*, 2002; Anderson, 2009; Easton, 2010; Williams and Karahanna, 2013). Similarly, Garrison (1994) and Lipscomb (2011) argue for the coexistence of research traditions from the positivist /

empiricist (quantitative) camp, alongside a phenomenological / interpretivist (qualitative) perspective.

Mixed methods research is considered particularly suitable for problem-solving, as it enhances "...the accuracy of judgments by collecting different kinds of data bearing on the same phenomena" (Jick, 1979, p.602). Therefore, the study adopts this approach, integrating *sentiment analysis* (see Turney, 2002; Wilson *et al*, 2005) and *thematic analysis* (see Gioia *et al*, 2012), alongside statistical analysis of the output of embedded psychometric micro-experiments.

Whilst it is acknowledged critical realists do not claim causation based on apparent relationships between variables, it is suggested that where statistical methods are used, there is merit in adhering to established methodological conventions in order to demonstrate transparency and rigour. Therefore, although forthcoming statistical analysis does not claim to possess explanatory utility in its own right, it is conducted and reported upon in a fairly traditional fashion. Consequently, where terms such as 'significant' are used, this is merely to highlight notable patterns or tendencies.

At the heart of this research is the desire to identify and understand mechanisms responsible for triggering behavioural dysfunction in police performance management systems. Therefore, it embraces a problem-solving ethos and emphasis on 'pragmatic science' (Anderson *et al*, 2001; Hodgkinson and Healey, 2008), where theory informs practice, resulting in research that is both academically rigorous *and* socially useful. Pettigrew (1995; 1997), Hodgkinson and Rousseau (2009) and Hodgkinson and Starkey (2011) call for research where methodological rigour and practical relevance are high; this study adheres to those principles.

### **3.3.2 Research Design Considerations**

In designing the research, a number of potential approaches were considered, each of which exhibited varying strengths and weaknesses. For example, consideration was given to conducting semi-structured interviews (Britten, 1999; Legard *et al*, 2003) and focus groups (Kitzinger, 1994) as these methods can generate rich data (Morgan, 1998; Silverman, 2000; Bloor *et al*, 2001). However, researchers often experience difficulties in obtaining access to potential participants (Johl and Renganathan, 2010)

and conducting interviews can require significant time and financial commitments (Seidman, 2006); the same applies to the transcription stage (Bryman, 2001).

The overriding consideration, however, was the risk of *interviewer effects* (Marshall, 1998; Opdenakker, 2006) (or *moderator bias* in focus groups – see Marlowe, 2000), and what Scheurich (1997) calls researcher's 'baggage'. In face-to-face or telephone interviews researchers may unconsciously impose their own voices, preconceptions and prejudices (Groves and Kahn, 1979; Quantz, 1992; Cassell and Johnson, 2006). Even an interviewer's gender (Landis *et al*, 1973; Groves and Fultz, 1985), age (Norris and Hatcher, 1994) or race (Cotter *et al*, 1982) are factors known to affect responses. Furthermore, as researchers may inadvertently seek confirmation rather than disconfirmation (Wason, 1960) distortion could arise due to question design, interviewer behaviour, or selectivity in emphasising certain data.

Although some degree of bias is inevitable in any study (Pannucci and Wilkins, 2010), these concerns were of particular relevance to this research, as the author is a serving police officer, actively involved in the development of national police performance management policy. Therefore, even with steps being taken to mitigate bias, there was a risk that data collection involving personal interaction could become tainted by the factors outlined above; there was also a danger that interviewees or focus group participants might be influenced by prior knowledge of the author's work (see Guilfoyle, 2011; 2012; 2013; 2015; 2016).

For these reasons, it was decided that a survey instrument would be the most appropriate means of gathering data, as respondents would remain at a 'safe distance' (see Bryman, 2001), thereby mitigating the possibility of interviewer effects. For added sterility, the researcher's identity would not be disclosed and the instrument would be administered by independent third parties. Furthermore, from a practical perspective, surveys have the advantage of being more feasible to implement than interviews (Easterby-Smith *et al*, 2002) and enable wider coverage (Kelley *et al*, 2003).

An additional consideration was that the nexus of the research is very specific, making it potentially difficult to construct interview or focus group templates capable of eliciting the type of data required. The concern was that interviews or focus groups might produce general narrative about behavioural dysfunction in

police performance management systems without fully addressing the research question; consequently, whilst such content may nominally enrich the literature, it could fall short of generating a theoretical contribution.

Therefore, it was concluded that the deployment of a large-scale survey instrument was the most appropriate means of data collection for this study. Whilst it is acknowledged survey data may not generally be as descriptively rich as interview data, in this case the output is considered to be of an appropriate type and depth to facilitate the search for mechanisms.

### **3.3.3 Overarching Analytical Strategy**

The first phase of the analytical strategy adopted for this research involves *critical analysis* (Cottrell, 2005) of relevant performance management literature and primary documentation obtained from UK police forces, alongside secondary material (e.g. open source content). This approach enables rich insight into the domains being studied and is endorsed by Suddaby *et al* (2011):

“Organizational theorists can develop new insights not through inductive analysis of empirical data alone but by considering the current body of literature (including papers, books, presentations, working papers) as another source of explicit empirical data... researchers can then focus their attention on specific aspects of the literature” (2011, p.243).

Empirical data collection then follows via the deployment of a large-scale survey instrument. This instrument captures free text data documenting respondents’ experiences of performance information use, alongside embedded psychometric tests that assess their reactions to various experimental visual stimuli. Such an approach is deemed consistent with a critical realist research ethic, especially as the use of questionnaires can be an effective method for eliciting evidence about mechanisms (Miller and Tsang, 2010).

Next, qualitative data collected via the survey instrument will be analysed in two stages; firstly, *sentiment analysis* (see Turney, 2002; Wilson *et al*, 2005) will be conducted, to identify any notable polarities or recurrent patterns associated with particular data display formats. This will be followed by *thematic analysis* utilising Gioia’s method (Gioia *et al*, 2012) to identify prominent concepts, themes and

dimensions; these will be arranged into *data structures*, enabling the generation of an empirically-grounded theoretical model.

Statistical analysis of quantitative data produced by the micro-experiments will then be conducted, in order to assess respondents' cognitive processes, decision-making and behavioural responses. This sets the study apart from much research into performance information use, as experimental methods are relatively rare within this field (Kroll, 2014)<sup>9</sup> and it is even more unusual to find such research that involves practitioners as subjects (Moynihan, 2016a; Moynihan *et al*, 2016).

Finally, outcomes of the qualitative and quantitative analysis will be synthesised, to consolidate the model and provide focus for discussion. Overall, this approach seeks to evaluate any identified relationships between data display and behavioural dysfunction, as well as identify mechanisms whose operation may be responsible for producing the phenomenon of interest.

Consequently, specific questions to be addressed are:

1. Are any recurrent patterns, themes, or significant differences observed in respect of sentiment polarity and / or narrative, and if so, why might this be?
2. Does engagement with particular data display formats appear to influence the likelihood, nature or extent of behavioural dysfunction, and if so, why might this be?
3. Do examples of behavioural dysfunction provided by respondents reflect those observed within the literature, and if so, what are the similarities and differences?

In tackling these considerations, the primary objective of this analytical strategy is therefore to answer the research question by providing a robust explanatory account for the phenomenon of interest, whilst explicating the role of data display as a potential catalyst for dysfunction.

---

<sup>9</sup> Aside from experimental methods used in the healthcare studies previously discussed, perhaps the work of Olsen (2013a; 2013b; 2015; 2017; 2018) is the main exception, alongside James, 2011; James and Moseley, 2014; Moynihan, 2015; Nielsen and Baekgaard, 2015; Andersen and Moynihan, 2016a; 2016b; Nielsen and Moynihan, 2017a.

### 3.3.4 Survey Instrument: Background and Design

This section provides an overview of the survey instrument. Firstly, it outlines the instrument's objectives and construction; next, it provides commentary on the pilot phase, validity and reliability, as well as the instrument's operationalisation. Analysis, results and conclusions are presented in subsequent chapters.

The instrument was constructed using Qualtrics software (Qualtrics, 2014), incorporating experimental and non-experimental dimensions, which capture a combination of quantitative and qualitative data. It was designed to assess two main areas:

1. How respondents interpret and react to visual stimuli depicting police performance information, presented using various data display formats.
2. Respondents' experiences of UK police performance management; in particular, the extent that certain types of performance information are used in their own forces and what the effects of this might be.

In designing the survey instrument, consideration was given to the order in which these two components would be exposed to respondents; whilst the first stage of *analysis* concentrates on the qualitative data pertaining to respondents' experiences, it was imperative that they were not inadvertently primed to react differently to the experimental component by having already recounted their experiences in their own forces. For this reason, the experimental component was positioned first in the survey and it is in this order that it is discussed in this chapter.

The primary objective of the experimental section was to elicit responses to various stimuli in as sterile an environment as possible; therefore, there should be no question of respondents being coerced to respond in certain ways by third parties or organisational pressures, as may be the case in operational environments. This approach therefore aimed to assess whether their interpretations and behavioural responses arose as a consequence of design, rather than latent factors (i.e. do particular tendencies arise even in the absence of external influences?)

To achieve this, the instrument was divided into four parts:

1. Experimental testing of responses to stimuli.
2. Data capture of personal experiences.
3. Data capture of respondents' rank / force.
4. Qualifying questions.

The rationale for this structure, along with the analytical strategy, is outlined below:

### 3.3.4.1 Section One: Experimental Testing

The first section of the instrument comprises five thematic blocks of psychometric micro-experiments of identical construction. The first three blocks are designed to test respondents' reactions to the following data display formats:

1. Binary comparisons.
2. League tables.
3. Numerical targets.

The sequence of these blocks is randomised; however, the questions within them follow a structured sequence, so are not further randomised. This is as follows:

1. *Visual stimulus* (for example, respondents are presented with a table containing a binary comparison between this month's and last month's crime figures, as below).

*Figure 3.2: Crime figures table*

Crime Figures			
Last month	This month	Difference	Percentage difference
1,598	1,745	147	8.4%

Respondents are then asked to provide an interpretation, by choosing one of four options, as follows:

*“In respect of the crime rate, which of the following does the table appear to indicate?” Crime is:*

- *Increasing*
- *Decreasing*
- *Stable*
- *Don't Know*

The purpose of this question is to establish whether respondents make any assumptions about trajectories based on variance from reference points. (For example, in this case, it is anticipated that many respondents will conclude crime is increasing).

2. *Level of concern* (i.e. how concerned do respondents become as a consequence of their interpretation?) They are asked:

*“As a result of the information contained in the table, how likely is it that you would be concerned about the crime rate?”*

Responses are captured on a 5-point Likert-type scale (see Likert, 1932; Colman *et al*, 1997; Vagias, 2006; Rattray and Jones, 2007; Sauro, 2007; Losby and Wetmore, 2012). Scalar descriptors range between ‘1’ (Very unlikely), ‘2’ (Unlikely), ‘3’ (Don’t know), ‘4’ (Likely) and ‘5’ (Very likely).

The purpose of this question is to assess whether respondents’ interpretation of the data generates concern about perceived deficiencies, and if this may ultimately affect subsequent behavioural responses. For the binary comparison stimulus, it is anticipated that many respondents will gravitate towards the higher end of the scale, as a result of assumptions made at Question 1.

3. *Behavioural response* (i.e. what action will respondents take?) Responses are captured on a series of identical 5-point Likert-type scales that assess the extent of five types of reactions. Respondents are asked:

*“As a result of the information contained in the table, how likely is it that you would respond as follows?”*

Respondents are asked, would they:

- *Do nothing.*
- *Ask for an explanation about performance.*
- *Communicate an expectation there should be an improvement.*
- *Offer praise regarding current level of performance.*
- *Initiate an operational response (e.g. commission further research, change tactics, deploy resources).*

Each option requires a response. These options provide a mix of ‘positive’ and ‘negative’ responses, intended to reduce the likelihood of ‘acquiescent response bias’; namely, “...the tendency for respondents to agree with a statement, or respond in the same way to items” (Rattray and Jones, 2007, p.237).

The options include behaviours identified as potential antecedents to dysfunction; this choice of variables should enable an assessment of whether relationships exist between them and output from previous questions, as well as confirm if there are strong associations between such behaviours and particular data display formats. In this case, it is anticipated that where respondents perceive negative variance they will be more likely to take action rather, than ‘do nothing’ (i.e. hold one’s nerve).<sup>10</sup>

These three thematic blocks of tests are then followed by a further two, which contain stimuli representing more reliable (albeit relatively visually complex) data display formats, namely:

1. A Statistical Process Control (SPC) chart.
2. A contextualised peer comparison diagram.

These blocks follow an identical format to those already described and their order of appearance is also randomised. It is not known how familiar respondents will be with these alternative formats, as they are less common in UK police forces. As they represent richer forms of data display, it is speculated respondents might be more likely to accurately interpret the content.

### **3.3.4.2 Section Two: Data Capture of Personal Experiences**

The purpose of this section is to gather data about respondents’ experiences of police performance management and establish the extent and impact of the use of various data display formats within UK police forces. Respondents are presented with identical stimuli to those in Section One<sup>11</sup> and asked the following question:

---

<sup>10</sup> The types of responses listed as options reflect behaviours identified in the literature review and are also comparable with the author’s experience. For obvious reasons, it was considered impractical to extend the range of options to include more extreme types of dysfunction, such as, “*Would you falsify documents / bully subordinates / misrecord crime?*” etc. See Olsen *et al* (2019), who assert measures of dishonesty and self-reported unethical behavior in surveys will be severely biased.

<sup>11</sup> The only exception is that the contextualised peer comparison diagram is omitted, as in the author’s professional experience, such diagrams are rarely used in UK police forces.

*“In your own organisation, how often are data presented using this format?”*

Responses are captured on a 5-point Likert-type scale with ascending scalar descriptors ranging from ‘1’ (Never) to ‘5’ (All of the time), along with a separate ‘Don’t know’ option.<sup>12</sup> Unless respondents select ‘Don’t know’ or ‘Never’, they are then asked:

*“If data are presented using this format in your organisation, do you believe this has any effect?”*

Respondents are invited to record their response in a free text box.

### **3.3.4.3 Section Three: Respondents’ Rank / Force Data**

Here, respondents are asked to provide their rank and the name of their force (or specify if they are not a police officer). No information capable of identifying individuals is captured. The purpose of this section is to enable stratification by rank to establish if there is variance at different levels of the organisation, as well as to identify whether there are significantly different patterns of responses across forces.

### **3.3.4.4 Section Four: Qualifying Questions**

In this section, respondents are asked additional questions to identify those who possess prior knowledge of statistical process control principles. They are asked:

*“How familiar are you with Statistical Process Control charts?”*

Responses are captured on a 5-point Likert-type scale with ascending scalar descriptors ranging from ‘1’ (Not at all familiar) to ‘5’ (Very familiar). There is no ‘Don’t know’ option. The purpose of this question is to establish whether respondents who accurately interpreted the SPC stimulus did so because of prior knowledge, or because the chart was relatively easy to interpret, even for those unfamiliar with the format. It should also determine if those familiar with SPC principles react differently to others in respect of the first three thematic blocks.

Respondents are then asked:

---

<sup>12</sup> ‘Don’t know’ is not presented as the central option as this scale does not contain positive and negative steps in either direction away from the mid-point, unlike those in the experimental section.

*“How familiar are you with the work of Simon Guilfoyle?”*

Responses are captured on a 5-point Likert-type scale identical to the previous question. The purpose of this question is to establish whether some respondents react differently due to being familiar with the author’s prior work, thereby identifying a potential source of ‘contamination’; each of these two questions should therefore help determine whether such familiarity behaves as a moderating condition (see Baron and Kenny, 1986).

Finally, respondents are given the opportunity to explain, add or clarify any of their answers, or make any further comment considered necessary. This is done via a free text option and is optional.

### **3.3.5 Survey Instrument: Pilot Phase**

During February and March 2014, the survey instrument was piloted with 17 participants, consisting of police officers of five different ranks, ex-police officers, police staff members and non-police personnel. The purpose of this approach was to obtain as wide a perspective as possible when gathering feedback about the survey design. During this period, the survey instrument underwent continuous revision, resulting in four iterations prior to finalisation.<sup>13</sup>

A key theme that emerged during this phase related to respondents who were already familiar with SPC; many stated they responded differently to the stimuli than they otherwise would have done. This confirmed the necessity of the qualifying questions included within the instrument. A further consideration involved respondents desiring additional information prior to making a decision. This was mitigated by introducing a ‘Don’t know’ option where appropriate, although from the author’s operational experience, it is commonplace for decisions to be made based on the same type and level of information as that provided in the tests.

---

<sup>13</sup> Whilst the primary objective was to gather data from police officers, the final instrument included options for police staff (i.e. employees of the police service who are not warranted police officers) and non-police personnel to identify themselves where they had participated. This was to either enable such responses to be placed into control groups if sufficient data were forthcoming, thereby enabling comparison with police officers, or allow them to be quarantined if not.

It was also recognised that whilst efforts were made to ensure the tests accurately reflected the operational policing context, it would be impossible to model every condition likely to influence respondents' decision-making. However, based on professional insight and feedback from pilot participants, it is deemed that reasonable steps were taken to ensure the tests replicated operating conditions as closely as possible. Comments from participants and rationale for changes or retention of particular design features were recorded within a contemporaneous document; a redacted version of this may be viewed at *Appendix 'A'*.

### **3.3.6 Survey Instrument: Live Phase**

In April 2014, a web link to the finalised survey instrument was circulated by the Association of Chief Police Officers (ACPO), the Police Federation of England and Wales, and the Police Superintendents Association of England and Wales (along with their Scottish and Northern Irish counterparts), following an agreement with all parties that they would expose the survey instrument to their members. (Copies of relevant correspondence may be viewed at *Appendix 'B'*, and a MS Word-exported version of the survey instrument can be inspected at *Appendix 'C'*).

The survey was made available for a period of 30 days, from 28<sup>th</sup> April 2014 to 27<sup>th</sup> May 2014 inclusive. A unique record was generated each time a respondent clicked the survey link, even when they accessed the front page but subsequently decided not to proceed. In total, the survey front page was accessed 7,157 times, of which 1,565 participants choose not to proceed further (21.9%). According to Solomon (2000), a high percentage of drop-outs at this stage is not uncommon for web-based surveys.

As the primary objective of the instrument was to obtain data from police officers, responses from police staff (N = 208) and non-police personnel (N = 18) were removed from the data set prior to analysis. The remainder of the data set was used for analysis, utilising IBM SPSS (IBM Corp, 2013). Of the remaining 6,930 occasions the survey link was accessed, 5,374 respondents engaged beyond the opening page (77.5%), with 4,173 of these completing all sections of the survey.

Although N = 4,173 for fully-completed surveys, additional data from partially-completed responses were also retained for analysis. This is because the instrument was partitioned into thematic blocks containing forced answer fields, thereby

ensuring each question produced a complete response. As non-parametric methods were used to analyse the data, differences in sizes of comparator data sets would not destabilise analysis.

Therefore, had partially-completed surveys been discarded, useful data (both quantitative and qualitative) would have been lost. As SPSS automatically disregards missing values during statistical analysis, it seemed prudent to include all valid data; where such additional data are available this adds greater depth to the analysis, irrespective of whether they are drawn from a respondent who completed the entire survey. Furthermore, including data from partially-completed surveys is considered to be good practice for guarding against attrition bias (Jüni and Egger, 2005).

Consequently, each of the data sets analysed involve larger sample sizes than 4,173, producing findings at the indicative levels shown in Table 3.1.

*Table 3.1: Indicative statistics*

<b>Measure</b>	<b>Value</b>
Confidence level	99.99%
Probability (two-tailed)	$p = <0.001$
Statistical power	
<ul style="list-style-type: none"> <li>• to detect an effect size of 0.1</li> <li>• to detect effect sizes <math>&gt;0.15</math></li> </ul>	<p>0.989</p> <p>1.000</p>

These values confirm the analysis is capable of establishing findings at very high levels. Confidence and probability thresholds are well in excess of the standard criteria of 95% and  $p = <0.05$  generally used in statistical analysis (Field, 2013), whilst statistical power is much greater than the recommended level of 0.8 (Cohen, 1988; see also Faul *et al*, 2007; Kalpana, 2011).

### **3.3.7 Robustness of the Survey Instrument**

Various methods of assessing validity were considered. As the instrument measures hitherto-untested criteria, it was not feasible to conduct criterion validity testing against an established scale (Litwin, 1995). Neither was it considered appropriate to undertake extensive analysis of the data generated during the pilot phase, or attempt reliability testing using statistical methods. This was partly because the sample size was small ( $N = 17$ ), but moreover it was recognised that ongoing revision of the instrument could render statistical analysis problematic.

Therefore, a *content validity* (see Litwin, 1995) approach was adopted to enable initial assessment. This involved seeking advice on the instrument's construction from academics experienced in survey design, as well as feedback from the police practitioners and lay persons who participated in the pilot phase. This approach enabled various perspectives to be considered when refining the instrument prior to circulation. Data were also visually inspected to identify any obvious patterns (which suggested the instrument was performing as expected), although no firm conclusions were drawn for the reasons described above.

In addition to the above, it was also necessary to test the instrument's underlying structural robustness (i.e. *construct validity* – see Trochim, 2006a), as well as select appropriate analytical techniques. The first step in doing so was to determine the measurement levels of the data and identify whether parametric or non-parametric methods would be most appropriate.

It was established that the first question in each thematic block produced nominal (or categorical) data, as it simply classified the type and frequency of responses. The second question (level of concern) and the third (behavioural response) utilised Likert-type scales, producing data measured at the ordinal level. The remainder of the instrument captures nominal data about respondents, as well as ordinal data regarding prior knowledge of SPC and the author's previous work.

Together with these considerations, decisions regarding appropriate analytical approaches took into account assumptions about distribution. Tests for normality indicated that all ordinal data exhibited patterns of non-normal distribution; Kolmogorov-Smirnoff tests and Shapiro-Wilks tests confirmed this ( $p = <0.001$ ). Histograms and Q-Q plots for each ordinal variable also identified skewness and kurtosis, although just five of the ordinal variables exhibited skewness  $>1$  or  $<-1$ .

Therefore, the measurement levels and non-normal distribution patterns suggest non-parametric tests were more appropriate than parametric tests. Nevertheless, there is disagreement within the literature about how strictly these rules should be adhered to, particularly when dealing with large sample sizes; these discussions were also considered (see Knapp, 1990; Kuzon Jr *et al*, 1996; Fagerland and Sandvik, 2009; Fagerland, 2012).

For instance, Carifio and Perla (2007), challenge the assertion Likert data must always be treated as non-parametric. Similarly, Fagerland (2012) insists *t*-tests are robust (even to heavily-skewed distributions) if sample sizes are larger than 20. Furthermore, Carifio and Perla (2008) argue Likert data may legitimately be treated as interval in some circumstances, such as when anchors are absolutes (e.g. ‘Never’).

Others however, remain unconvinced. Jamieson (2004), for example, argues the presence of skewness violates assumptions about normal distribution, which renders parametric analysis inappropriate. Furthermore, she is adamant that Likert data are ordinal and should only be analysed using non-parametric methods, asserting “...the response categories have a rank order, but the intervals between values cannot be presumed equal” (2004, p.1217). Paraphrasing Kuzon Jr *et al* (1996), she points out “...the average of ‘fair’ and ‘good’ is not ‘fair-and-a-half” (Jamieson, 2004, p.1218).

Taking this debate into account, and considering the nature and measurement levels of the data produced by this instrument, it was deemed most appropriate to use non-parametric methods, as they make few assumptions about distribution patterns and do not require measurement levels to be interval or ratio. Whilst it is broadly accepted that non-parametric analysis is generally less statistically powerful than parametric techniques, this is of limited consequence when large samples are involved (Lehmann, 1998; Hoskin, 2014).

### **3.3.8 Reliability**

In order to ascertain the instrument’s reliability (i.e. whether it effectively measures intended dimensions, is robust to measurement error, and data are highly reproducible – see Litwin, 1995), it was necessary to conduct a series of assessments prior to analysing output from respondents. This included evaluating *convergent validity* (i.e. measures of constructs that theoretically should be related, are in fact observed to be related) and *discriminant validity* (i.e. measures of constructs that theoretically should not be related, are in fact unrelated – see Trochim, 2006b). For this instrument, however, convergent validity was deemed more relevant than discriminant validity, as it was constructed to assess related dimensions.

All ordinal variables were tested to establish the instrument’s Cronbach’s Alpha coefficient, a common measure of internal reliability (Field, 2013, pp.708-710).

Kline (1999) suggests that whilst the accepted standard for internal reliability is generally acknowledged to be 0.7 or higher (see also Brace *et al*, 2009, p.368), for psychological constructs, lower values may be expected due to the diversity of the dimensions being measured. Nevertheless, the Cronbach's Alpha coefficient for the instrument was found to be well in excess of this threshold, at 0.897, indicating high internal reliability (Nunnally, 1967).

Furthermore, the vast majority of inter-item correlations produced values in excess of 0.3, confirming high internal reliability (see Field, 2013, p.685). The consistent exception in each of the five categories was the variable '*Offer praise regarding current level of performance*', which produced low values ranging between -0.139 and 0.083. This suggests this item did not effectively capture the same underlying dimensions as the other variables; in these circumstances Field (2013) recommends such variables be omitted. By doing so, it was observed that the overall reliability of the instrument was strengthened and the Cronbach's Alpha coefficient rose to 0.921.

A complementary test of correlation between individual variables using Spearman's *rho* (see Fink, 1995, pp.38-39) confirmed significant correlations between ordinal variables at the  $p = <0.001$  level in all categories, except for the '*Offer praise regarding current level of performance*' variable, where many inter-item correlations were not significant. After this variable was omitted, a re-run of the test with the remaining variables confirmed correlations between all variables ( $p = <.001$ ).

Factor analysis using the ordinal variables was then conducted, which produced output of 0.906 for the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO); this falls within the strongest category (i.e.  $>0.9$ ), indicating compact patterns of correlation (Hutcheson and Sofroniou, 1999). Bartlett's Test of Sphericity (see Brace *et al*, 2009, p.354) produced output of Chi-Square = 56190.223,  $p = <0.001$ . These highly significant results indicate that a large proportion of variance within the data is explained by factors. Furthermore, anti-image matrices comparing negative partial covariances and negative partial correlations also exhibited patterns which support the likelihood of there being an underlying factor structure.

Extraction using Principal Component Analysis (PCA) produced communalities for individual variables ranging between 0.612 and 0.808. *Kaiser's Criterion* states that where the sample size is greater than 250 and the average extraction communality is

$\geq 0.6$ , this indicates factorability (Field, 2013, p.877). PCA also identified there were eight factors with an Eigenvalue  $> 1$ , accounting for 74.45% of total variance. It was found that one factor, accounting for 34.42% of total variance, loaded strongly onto the majority of the variables, suggesting consistency in how the instrument measures underlying dimensions.

Additionally, upon examining reproduced correlations and their residuals, it is noted that when observed correlations and predicted correlations were compared, the residuals tended to be small (in some cases, values of  $-0.001$  and  $0.001$  were recorded). This provides a further strong indication of factorability (Brace *et al*, 2009, p.359). Overall, the tests results assure the presence of convergent validity. Direct oblimin rotation was then conducted (see Kline, 1994); this produced pattern and structure matrices, showing which variables load most strongly onto which factors.

In summary, it can be stated that the instrument's design is sufficiently rigorous to be considered capable of reliably assessing underlying dimensions.

### **3.3.9 Sampling and Biases**

At the time of the survey in 2014, data on UK police officer numbers indicated there were 129,584 full-time equivalent (FTE) police officers in England and Wales (Berman and Dar, 2013, p.3), 17,496 FTE officers in Scotland (2013, p.15) and 6,885 FTE officers in Northern Ireland (2013, p.17). Therefore, the total population of UK police officers was 153,956.

Due to the prohibitive logistical and administrative requirements of conducting fully randomised sampling in 48 separate police forces, it was acknowledged that such an approach would not be feasible for this study. As the alternative method chosen was national dissemination via web link, consideration was given to potential biases. For instance, coverage bias (see Bethlehem, 2007) was deemed not to be a significant issue, as UK police forces are routinely internet-enabled, ensuring particular demographic groups would not be inadvertently excluded.

However, a possible risk was *sampling bias*, and in particular, *self-selection bias*. Knapp (2014) identifies web-based surveys as being prone to self-selection bias, which can result in a non-representative sample and difficulties in generalising findings. Nevertheless, De Vaus (2002) argues web-based surveys can still be "... a

very useful means of obtaining representative samples of specific populations” (2002, p.79). For this study, the sampling frame matches the specific population of ‘all UK police officers’, with access being equitable across age, sex, length of service, ethnicity and other personal characteristics.

Consequently, although the possibility of self-selection bias is a valid consideration, it is suggested any practical impact in these circumstances would be minimal. Therefore, to discount a web-based survey on the basis of self-selection alone was considered inappropriate, as the opportunity to gather relevant data at the national level would still be of immense value. As posited by Winship and Mare (1992):

“To rely exclusively on observational schemes that are free from selection bias is to rule out a vast portion of fruitful social research” (1992, p.328).

However, options for controlling survey bias were still considered; namely, adjustment weighting and reference surveys. In order to undertake the former, Cuddeback *et al* (2004) suggest the source and direction of bias should be understood prior to applying statistical correction. As this was not known, it was decided that a reference survey involving a smaller group, controlled through randomised sampling, would be a more appropriate option, as this “...can substantially reduce the bias of web survey estimates” (Bethlehem, 2008, p.13).

Therefore, in order to control for self-selection bias, the survey instrument was also deployed on a smaller scale within a single police force (Leicestershire) for a period of 30 days, between 31<sup>st</sup> October and 29<sup>th</sup> November 2014 inclusive. At the time of the survey, the force comprised 2,074 FTE police officers.

Prospective respondents were selected through systematic sampling (De Vaus, 2002, pp.72-74), where an independent observer drew lots to determine a random starting point from within a register of all Leicestershire officers. Every subsequent fifth officer was then added to a panel and emailed individually with a unique web link. This method of selecting respondents was deemed appropriate as there would be no anticipated periodicity issues within the sampling frame, and no requirement for stratification. The process resulted in a fully randomised sample: N = 375.

The purpose of the reference survey was to establish if the results produced by this fully randomised sample closely reflected those of the larger, national sample. If so,

this would suggest participants were indeed representative of the UK police officer population, thereby strengthening the legitimacy of the national study and confirming that the instrument possesses *external validity* (see Trochim, 2006c).

At the closure of the reference survey, 195 of the 375 respondents (52%) had completed all sections, although earlier questions enjoyed higher levels of participation. Babbie (1990) proposes that anything over a 50% response rate is adequate and certainly within a range not untypical for survey returns; Welch and Barlau (2014) noted returns typically fall in the 50-65% range, whilst Baruch (1999) found the average to be around 55%.

Ary *et al* (1996) recommend it is necessary to check for the presence of nonresponse bias wherever the response rate is less than 75%, whereas Borg and Gall (1983) suggest the appropriate threshold is 80%. The purpose of investigating potential nonresponse bias is to establish if there are significant differences between responders and nonresponders. One recognised method is to examine known characteristics of each group (e.g. gender), to see if there are differences that may need to be controlled for.

Another method is to use late responders as a surrogate for nonresponders, by comparing their responses with those of early responders. Whilst it is necessary to test for late response bias as a matter of course, this approach can also provide an insight into the extent of nonresponse bias, being considered a well-established and legitimate method for identifying its presence and probable direction (Pace, 1939; Miller and Smith, 1983; Groves, 2006).

Consequently, both techniques were employed, although personal characteristics were deemed less relevant to this particular study. This was due to the sample being drawn from a sample frame of police officers without attempting to stratify by gender, age etc. Nevertheless, the email distribution panel recorded whether respondents were male or female, meaning it was possible to track the gender split amongst respondents and non-respondents.<sup>14</sup> The breakdown is provided below:

---

<sup>14</sup> Nationally, the gender split for police officers is Male: 72.7% / Female: 27.3% (Berman and Dar, 2013, p.7). These proportions also happened to be replicated exactly in Leicestershire Police at the time of the survey (Leicestershire Police, 2014).

Table 3.2: Gender frequencies

	Male	Female
<b>Total invitations:</b> 375	278 (74.1%)	97 (25.9%)
<b>Total responses (including partial):</b> 227	169 (74.4%)	58 (25.6%)
<b>Total nonresponses:</b> 148	109 (73.6%)	39 (26.4%)

It can be seen the gender split was closely replicated across groups (variance for each gender was <1% between respondents and non-respondents). This indicates minimal difference between groups, meaning gender-based nonresponse bias is unlikely to be an issue. As stated, however, it was anticipated greater insight into potential late and nonresponse bias was likely to be gleaned from analysis of early and late responses.

Therefore, the first 50 and last 50 respondents were placed into respective groups and Mann-Whitney U tests were conducted for each variable within the survey instrument, to establish if there were any significant differences between early and late responders. The analysis found no significant differences between groups for any of the variables. Consequently, this suggests late response bias and nonresponse bias were unlikely to have an adverse impact. It also suggests the instrument itself is of sufficiently rigorous design to enable reliable ‘test-retest’ replication of the study (see Field, 2013, p.885).

In respect of the national survey, identical tests for late response / nonresponse bias were also conducted. Data from the first and last 1,000 respondents were compared, revealing the presence of potential bias in 16 of the 30 variables, although in all cases, the effect size was small or very small ( $r = 0.03$  to  $0.23$ ). Furthermore, for all but one variable, differences involved *greater skewness* in the expected direction. This suggests early respondents were more conservative than late respondents, potentially indicating a slight *underestimation* of the strength of anticipated responses amongst the unobserved population.<sup>15</sup>

Next, further Mann-Whitney U tests were conducted to compare responses generated by the reference survey with those from the national survey. Of the 30 variables individually tested, only one significant difference was found ( $p = 0.04$ ), with a very

<sup>15</sup> The exception related to the variable that captures respondents’ interpretation of the league table stimulus; early respondents were less likely to select ‘Don’t know’ (44.3%) than late respondents (59.3%).

small effect size ( $r = 0.03$ ). Of note, this negligible divergence related to there being greater skewness in the expected direction for a variable in the reference survey. As this could be considered the more reliable of the two due to the randomised sampling, this provides reassurance that the output is reflective of the population.

It is therefore suggested that the design and sampling methodology of the reference survey are sufficiently robust, and that its output so closely matches that of the national survey, to be able to state that both data sets are highly likely to be representative of the UK police officer population.

### **3.4 Summary**

In summary, it may be stated that the construction, testing and deployment of the survey instrument have been systematic and rigorous, thereby ensuring the output it produces should accurately reflect behavioural tendencies. In a positivist study, it would be claimed this instrument capably produces inferential statistics that enable findings to be generalised to the entire UK police officer population; however, from a critical realist standpoint, generalisability comes about as a result of:

“...identifying the deep processes at work under contingent conditions via particular mechanisms...The best explanation, that is the one most consistent with the data, is what is being sought” (Easton, 2010, p.126).

In other words, the data act as a raw material from where patterns and tendencies can be identified through analysis and used as a basis for developing theory that is applicable beyond that particular case. This represents the first step in a process that leads to the production of an explanatory account characterised by retroductive logic and underpinned by the highest levels of statistical rigour.

Overall, taking into account the combined strengths of this study’s philosophical foundations, attention to methodological rigour and careful choice of methods, it is argued these ingredients enable the production of a robust explanatory account for the phenomenon of interest. The following chapter builds on this foundation, discussing results of the qualitative analysis and presenting interim conclusions.

## Chapter Four

### Qualitative Analysis and Findings

#### 4.1 Introduction

This chapter reports on the analysis of qualitative data generated by the survey instrument. As discussed in the previous chapter, this instrument contains a series of questions designed to elicit free text responses; four fields record respondents' experiences of the effects of the use of binary comparisons, league tables, numerical targets and SPC charts respectively, whilst a fifth provides an opportunity to make general comments regarding police performance management or how police performance information is used.

This text was imported into Excel, producing one row of entries per respondent. Not all respondents answered all questions and some text was removed due to it being unsuitable for analysis (e.g. 'Good luck with your PhD' or 'no comment'). Nevertheless, after sanitisation, the total number of viable entries was as follows:

- Binary comparisons: 3,241 responses.
- League tables: 3,173 responses.
- Numerical targets: 3,361 responses.
- SPC charts: 2,233 responses.
- Optional free text commentary: 1,127 responses.

In total this provides 13,135 text entries for analysis.

The forthcoming sections report on the implementation and output of the analytical processes, beginning with sentiment analysis (see Turney, 2002; Wilson *et al*, 2005), conducted to identify any patterns or notable differences amidst the data. Next, Gioia's method of thematic analysis (Gioia *et al*, 2012) is utilised to identify key concepts, themes and dimensions; this process produces a data structure, which acts as a basis for the generation of a theoretical model. Finally, this model is discussed and used as a basis for interim conclusions about the presence of a mechanism that may be responsible for exerting influence upon the phenomenon of interest.

## 4.2 Sentiment Analysis

The first stage in analysing the free text data was an assessment of the overall polarity of responses for each of the data display formats, using *sentiment analysis*:

“Sentiment analysis is the task of identifying positive and negative opinions, emotions, and evaluations” (Wilson *et al*, 2005, p.347).

Typical approaches utilise lexicons of positive and negative words or phrases that tag words with their *a priori* polarity, before classifying a document based on the number of such features present (Pang and Lee, 2004). Some methods also attempt to identify the strength of polarity (Nasukawa and Yi, 2003). In either case, analysis provides an indication of whether the content exhibits positive, negative (and sometimes neutral) polarity.

However, a limitation of using pre-constructed lexicons is that this approach may not take overall context into account (Wilson *et al*, 2005); furthermore, the complex structures of some sentences may distort a dominant sentiment (Nasukawa and Yi, 2003). This means it can be difficult to build an effective framework for automated or algorithmic-based sentiment analysis (Hatzivassiloglou and McKeown, 1997; Turney, 2001; 2002).

Furthermore, although most sentiment analysis is conducted at the document level, it is sometimes necessary to conduct sentence-level or even phrase-level analysis (Wilson *et al*, 2005). This consideration is highly applicable here due to the nature of the data, as the entries consist largely of short passages commenting upon specific performance information formats. For these reasons, it was concluded that the best way to ensure the sentiment polarity of each entry was correctly assessed was to conduct manual analysis of all items individually.

Therefore a template was constructed comprising a column for each of the performance information types, with adjacent columns for sentiment polarity coding (see *Appendix 'D'*). The condition ‘positive’ was assigned numeric value 1, ‘neutral’ 2 and ‘negative’ 3. (The optional commentary entries were not coded for sentiment polarity as their purpose was to inductively capture themes for subsequent analysis).

Each respondent was assigned to a row, whereby his or her responses were recorded across each of the categories (3,853 rows, 12,008 individual entries in total). A further column documented occurrences whereby a respondent provided commentary on all four performance information types (1,379 rows), along with that particular permutation of polarity codes (e.g. '1/3/3/2').

Coding criteria were determined as follows:

1. Responses must describe the effect of the use of that particular data display type (either on respondents or their force). Therefore, general comments such as 'performance management is wrong' were not coded for sentiment analysis (although still retained for subsequent analysis at the Gioia stage).
2. Coding was limited to 'positive', 'neutral' and 'negative', as it was considered unnecessary and potentially problematic to try to add additional 'strong' and 'moderate' categories for positive and negative polarity.
3. 'Neutral' was defined as 'where overall polarity is neither positive nor negative, either because the respondent states there is little or no effect (e.g. 'none'), or because a response is balanced with pros and cons' (e.g. 'this can help as well as hinder').
4. Whilst some sentiment analysis approaches utilise a 'both' category (see Wilson *et al*, 2005), in this case where a field contained both positive and negative commentary, either this was coded in line with the stronger sentiment polarity, or as 'neutral' if the content was equally balanced.

#### **4.2.1 Inter-rater Reliability**

In order to imbue rigour and enable firm conclusions to be drawn from the analysis, *inter-rater* (or *inter-coder*) reliability tests were conducted (see Cohen, 1960; Lavrakas, 2008; Hallgren, 2012). Two raters independently coded a sample of 50 rows (comprising 155 individual text entries), selected by identifying a starting point within the main data set using a random number generator. For anonymity, the raters were assigned labels IR1 and IR2, with pre-existing coding from the main data set being allocated the moniker OC.

A Mann-Whitney U test was then conducted to assess whether the sample of 155 entries was representative of the remainder of the coded entries ( $N = 11,027$ ) this indicated there was no significant difference between the two samples  $p = 0.700$ , providing confidence that the sample selected for inter-rater reliability testing was reflective of the overall data set.

Subsequent analysis was conducted on the three sets of 155 entries, to establish if there were any significant differences between them. A Kruskal-Wallis test found there were none ( $p = 0.617$ ), with pairwise Mann-Whitney U tests confirming this: OC v IR1 ( $p = 0.369$ ), OC v IR2 ( $p = 0.890$ ), IR1 v IR2 ( $p = 0.297$ ).

Inter-rater reliability tests using the Kappa statistic (see Cohen, 1960) was then performed to determine consistency among raters. For OC v IR1, agreement was reached in 83.9% of cases<sup>16</sup>, with Kappa = 0.734 (95% CI = 0.652 and 0.816),  $p = <0.001$ ; reflecting ‘substantial’ consistency (Landis and Koch, 1977). For OC v IR2, agreement was reached in 89.1% of cases, with Kappa = 0.817 (95% CI = .736 and .898),  $p = <0.001$ ; this falls into the highest category of consistency (i.e.  $>0.81$ ). For IR1 v IR2, agreement was reached in 85.16% of cases, with Kappa = 0.756 (95% CI = .672 and .84),  $p = <0.001$ ; this also falls into the ‘substantial’ range’.

Additionally, a test for Fleiss’ Kappa (see Fleiss, 1971; Fleiss and Cohen, 1973) was conducted to measure the overall level of agreement between all three raters. This established Fleiss’ Kappa = 0.769 (95% CI = .702 to .836),  $p = <0.001$ , confirming a high level of agreement between all three raters, towards the top end of the ‘substantial’ category. The output of these tests therefore confirms there is a high degree of inter-rater reliability, which assures rigour and enables strong inferences to be drawn from the analysis of sentiment codes relating to the main data set.

---

<sup>16</sup> Note: Hallgren (2012) advises percentages alone should not be relied upon to report inter-rater agreement as they overlook chance agreement; the Kappa statistic adjusts for this and is therefore a more reliable measure. Percentage values are only reported here alongside other statistics for illustrative purposes.

#### 4.2.2 Main Data Set

Descriptive statistics for the main data set are shown in Table 4.1, below.

Table 4.1: Frequency table (all categories)

	Binary comparisons	League tables	Numerical targets	SPC charts
<b>Total rows</b>	3,853	3,853	3,853	3,853
<b>Completed</b>	3,241	3,173	3,361	2,233
<b>Empty</b>	612	680	492	1,620
<b>Coded</b>	2,987	3,056	3,119	2,020
<b>Uncoded</b>	254	117	242	213
<b>Of coded:</b>				
<b>Positive (1)</b>	268	132	191	819
<b>Neutral (2)</b>	1,248	814	1,052	847
<b>Negative (3)</b>	1,471	2,110	1,876	354
<b>Positive %</b>	9.0%	4.3%	6.1%	40.6%
<b>Neutral %</b>	41.8%	26.6%	33.6%	41.9%
<b>Negative %</b>	49.2%	69.1%	60.2%	17.5%

This pattern of responses indicates lower levels of positive sentiment are associated with the first three formats than with SPC; this pattern is also apparent in reverse when the breakdown of negative responses is examined. Additional statistics at Table 4.2. corroborate this.

Table 4.2: Additional statistics (all categories)

	Binary comparisons	League tables	Numerical targets	SPC charts
<b>N</b>	2,987	3,056	3,119	2,020
<b>Skewness</b>	-0.625	-1.328	-0.968	0.382
<b>Std error of skewness</b>	0.045	0.044	0.044	0.054
<b>Chi-Square</b>	822.678 $p = <0.001$ $w = 0.52$ (large effect)	1976.995 $p = <0.002$ $w = 0.80$ (large effect)	1365.669 $p = <0.001$ $w = 0.66$ (large effect)	227.751 $p = <0.001$ $w = 0.33$ (medium effect)

Table 4.2, shows that the first three formats exhibit negative skewness (indicating negative sentiment polarity), whilst SPC exhibits positive skewness. This is supported by the results from the Chi-square ‘goodness of fit’ statistics. Overall, the analysis finds a tendency for respondents to associate the use of binary comparisons, league tables and numerical targets with negative effects, as well as a tendency to associate the use of SPC charts with positive effects.

### 4.2.3 Permutations

Where a respondent completed all four fields, the overall combination of sentiment codes<sup>17</sup> was recorded in a separate column. A matrix was constructed containing all 81 permutations (i.e. 1/1/1/1 to 3/3/3/3), along with descriptive statistics displaying the count and percentage for each combination; this may be viewed at *Appendix 'E'*. This produced 1,379 complete entries, with only the following six combinations attracting a percentage >4%:

- 3/3/3/1 (15.6%, n = 215)
- 3/3/3/3 (12.7%, n = 175)
- 3/3/3/2 (10.7%, n = 147)
- 2/2/2/2 (8.3%, n = 115)
- 2/3/3/2 (5.9%, n = 82)
- 2/3/2/2 (5.9%, n = 81)

A 'goodness-of-fit' Chi-Square test produced output of 5951.862 ( $p = <0.001$ ), indicating that the pattern of responses is significantly different from what could be expected through chance alone;  $w = 2.08$ , which is categorised as a very large effect.

The strongest combination indicates negativity towards the first three performance information formats, with positive sentiment polarity towards SPC. When combined with 3/3/3/2, a total of 26.3% of respondents associate negative effects with the first three formats, whilst citing either positive or neutral effects for SPC charts. This suggests respondents who assessed all four formats often made a distinction between the effects of SPC chart usage and other formats. A separate group (12.7% of responses) ascribed negative sentiment polarity to all four performance information types; potential reasons for this will be explored during subsequent analysis.

Overall, there was a tendency for respondents to discriminate between SPC charts and the other formats, associating positive or neutral effects with SPC, whilst citing negative effects for the others. Taken in isolation, these patterns cannot be used to conclusively claim SPC charts are consistently held in higher regard; however, the results certainly seem to indicate there are strong patterns in respect of the direction of polarity, depending on the particular format.

---

<sup>17</sup> i.e 1= 'positive', 2 = 'neutral', 3 = 'negative'.

#### **4.2.4 Sentiment Analysis: Summary and Conclusions**

The analysis offers an insight into respondents' experiences of the effects of different police performance information formats. Primarily, it found:

1. Binary comparisons, league tables and numerical targets were all associated with negative effects. This was confirmed from multiple perspectives (i.e. frequency count, percentage breakdown and skewness).
2. SPC charts were less likely to be associated with negative effects and more likely to produce positive effects, when compared to the other formats. This too, was confirmed from multiple angles, as above.

Overall, sentiment analysis indicates the use of the comparatively simplistic formats tends to produce more negative effects than SPC charts; it also found SPC charts are associated with more positive effects in comparison. This provides a starting point for consideration of a relationship between certain performance information formats and either positive or negative effects; it also begins to open up the possibility of there being a mechanism responsible for producing events of a particular polarity, depending on the type of data display used.

These prospects will be explored in greater depth during the forthcoming thematic analysis, as well as the experimental component phase of the research, to establish if the findings of sentiment analysis correlate with the other facets of this study.

### **4.3 Thematic Analysis**

#### **4.3.1 Introduction**

The next stage of analysis goes beyond assessing the polarity most strongly associated with particular performance information formats, to identifying prominent themes and patterns within the free text data. It also aims to establish whether any observed patterns reflect the types of behavioural dysfunction and unintended consequences recounted within the literature. Its ultimate objectives are to assess the likely presence of mechanisms and explore whether data display could affect their operation, thereby influencing the likelihood, nature or extent of dysfunction.

### 4.3.2 Thematic Analysis: The Gioia Method

To fulfil these objectives, *thematic analysis* was conducted using an approach developed by Gioia and colleagues (Gioia and Chittipeddi, 1991; Gioia *et al*, 1994, Corley and Gioia, 2004; Gioia *et al*, 2012). Thematic analysis may be defined as, “...a search for themes that emerge as being important to the description of the phenomenon” (Fereday and Muir-Cochrane, 2006, p.82; see also Daly *et al*, 1997). This is a recursive process, which involves moving back and forth between the entire data set, coded extracts of data, and the analysis produced (Braun and Clarke, 2006).

Thematic analysis may employ inductive, deductive, or hybrid approaches (Braun and Clarke, 2006; Fereday and Muir-Cochrane, 2006). However, it is acknowledged that when data are considered in the light of extant literature and existing theory, a purely inductive approach may not be practicable, as the research process may transition to a more abductive form of research (Gioia *et al*, 2012; see also Alvesson and Kärreman, 2007). This involves, “...cycling between emergent data, themes, concepts, and dimensions and the relevant literature” (Gioia *et al*, 2012, p.21).

Furthermore, as there is no universally accepted ‘boilerplate’ (i.e. template) for determining quality or writing up qualitative research (Pratt, 2009), such approaches are often subject to criticism that ‘anything goes’ (Antaki *et al*, 2003; Laubschagne, 2003). To counter these accusations, it is necessary to demonstrate rigour throughout the process, in order to promote credibility and integrity (Koch, 1994); in particular, when conducting thematic analysis, it is essential to demonstrate how themes are generated from the raw data (Fereday and Muir-Cochrane, 2006).

To enable this, Gioia *et al* (2012) propose a methodology for ensuring systematic conceptual and analytical discipline, designed to imbue rigour into the research and analytical processes. It involves firstly identifying initial concepts from within the data and grouping them into categories for ‘first-order’ analysis (i.e. an analysis using informant-centric terms and codes). This stage utilises *open coding* (Locke, 2001) to recognise and capture important moments within the raw data (Boyatzis, 1998). Corley and Gioia (2004) recommend using *in-vivo* (Strauss and Corbin, 1990) or ‘first order’ (Van Maanen, 1979) codes (i.e. language used by informants) whenever possible; indeed, an established method of strengthening validity is to directly quote participants’ own words (Rice and Ezzy, 1999; Patton, 2002).

The next stage moves from open coding to *axial coding* (Strauss and Corbin, 1990; Pratt *et al*, 2006) and ‘second-order’ analysis (i.e. using researcher-centric concepts, themes, and dimensions), to identify relationships, similarities and differences among the categories, thereby reducing them to a more manageable number (Gioia *et al*, 2012). The final stage involves distilling the first order concepts and second order themes into *aggregate dimensions*, which enables the construction of a *data structure* and ultimately a theoretical model (Gioia *et al*, 2012, pp. 20-21).

This tiered approach clearly marks out a traceable path between raw data and the theoretical model. Whilst first order concepts may produce a rich narrative, they do not necessarily suggest an obvious theoretical framework; however, second order themes enable the identification of patterns in the data and underlying explanatory dimensions. They may also generate perspectives that are relevant beyond the boundaries of the organisation being studied; furthermore, it is here that the data begin to engage with theory (Gioia and Chittipeddi, 1991).

The generation of a data structure, described as the ‘pivotal’ step in the entire research approach (Gioia *et al*, 2012, p.20) offers a visual representation of the progression from raw data to themes and theoretical propositions. This emergent data structure, supported by exemplar quotes from respondents, provides richness, integrity and rigour, thus enabling the production of strong theory.

### **4.3.3 Application of the Gioia Method**

In order to gather free text data for analysis, survey respondents were presented with four performance information formats used in UK police forces (a binary comparison table, a league table, a numerical target and a SPC chart). Each stimulus was accompanied by a single open question, such as:

“If league tables are used in your organisation, do you believe this has any effect? If so, please describe”.

The question was worded in this way to enable respondents to record any type of effect associated with that particular format, be it positive, negative, or neutral. Furthermore, respondents were given the option of making open comments about any aspect of police performance management in a separate field. The purpose of this design was to identify common themes across the performance information

types, as well as demarcate any patterns which appear to be most strongly associated with any particular format. Responses were exported onto an Excel master sheet, which can be inspected at *Appendix 'F'*.

After manually coding the 13,135 entries, 1,351 (10.3%) potentially useful representative quotes were identified for further analysis, as they appeared to reflect emergent patterns and themes within the broader data set. These entries were then arranged by category (i.e. performance information type) and analysis was commenced utilising the Gioia method, with each category being addressed separately, in order to preserve any indigenous features.

The first stage of this process involved printing each of the 1,351 quotations onto individual pieces of paper bearing a unique reference number derived from the master sheet of 13,135 entries, reviewing each one in turn and then manually grouping them into potential first order concepts on a large surface. As each theme began to solidify, it was assigned a provisional title; some of these themes continued to emerge, whilst others waned or became subsumed elsewhere. This process eventually resulted in the identification of clear first order concepts, as well as potential second order themes.

The output was then transferred onto another Excel spreadsheet, whereby the first order concepts were listed and referenced using the numbering system described, so that each quotation could be traced directly to its source. Further examination of these concepts then took place and they were arranged into second order themes, before being distilled into aggregate dimensions. This resulted in the construction of a combined data structure for the binary comparisons, league tables and numerical targets categories (due to significant consistency between them), as well as a separate data structure for the SPC charts category. The process was recursive, iterative and organic; it felt as though the themes 'suggested themselves'.

Consequently, it is proposed that the starting point for analysis was broadly inductive, beginning with an open question resulting in first order concepts being provisionally identified, clustered, rearranged and finally designated. This was particularly the case for the binary comparisons category, as it was the first to undergo analysis; however, for subsequent categories, it is accepted that prior knowledge of the themes identified in the first category would be present in the

consciousness of the researcher. Therefore, although subsequent categories were analysed in the same fashion, it is suggested that the analytical process shifted from an inductive to an abductive disposition as it proceeded.

#### **4.3.4 Data Structure and Discussion**

This analysis identified several strong themes and patterns, uncovering apparent relationships between many of these themes, as well as potential influences operating amongst them. The two separate data structures generated by thematic analysis are presented at Figures 4.1 and 4.2 below. The first is an amalgam of the binary comparisons, league tables and numerical targets categories, as it was found that the concepts unearthed were remarkably consistent; the second pertains to the SPC charts theme, where first order concepts fit within identical second order themes to those present within the primary data structure, but tend to be of opposing polarity.

In line with Gioia's method, first order concepts utilised informant-centric terms, second order themes were defined using researcher-centric concepts, whilst aggregate dimensions cluster these themes and concepts. Overall, three aggregate dimensions were identified, being founded upon ten distinct second order themes; these in turn were derived from 31 first order concepts that emerged during the paper-based stage of analysis.<sup>18</sup>

Following presentation of these data structures, each of the aggregate dimensions is discussed in turn, with reference to the underpinning themes and concepts; this narrative is supported by representative quotes from respondents. As the focus of the thesis is on the relationship between data display and behavioural dysfunction, it is natural that these two themes feature most prominently in this discussion.

---

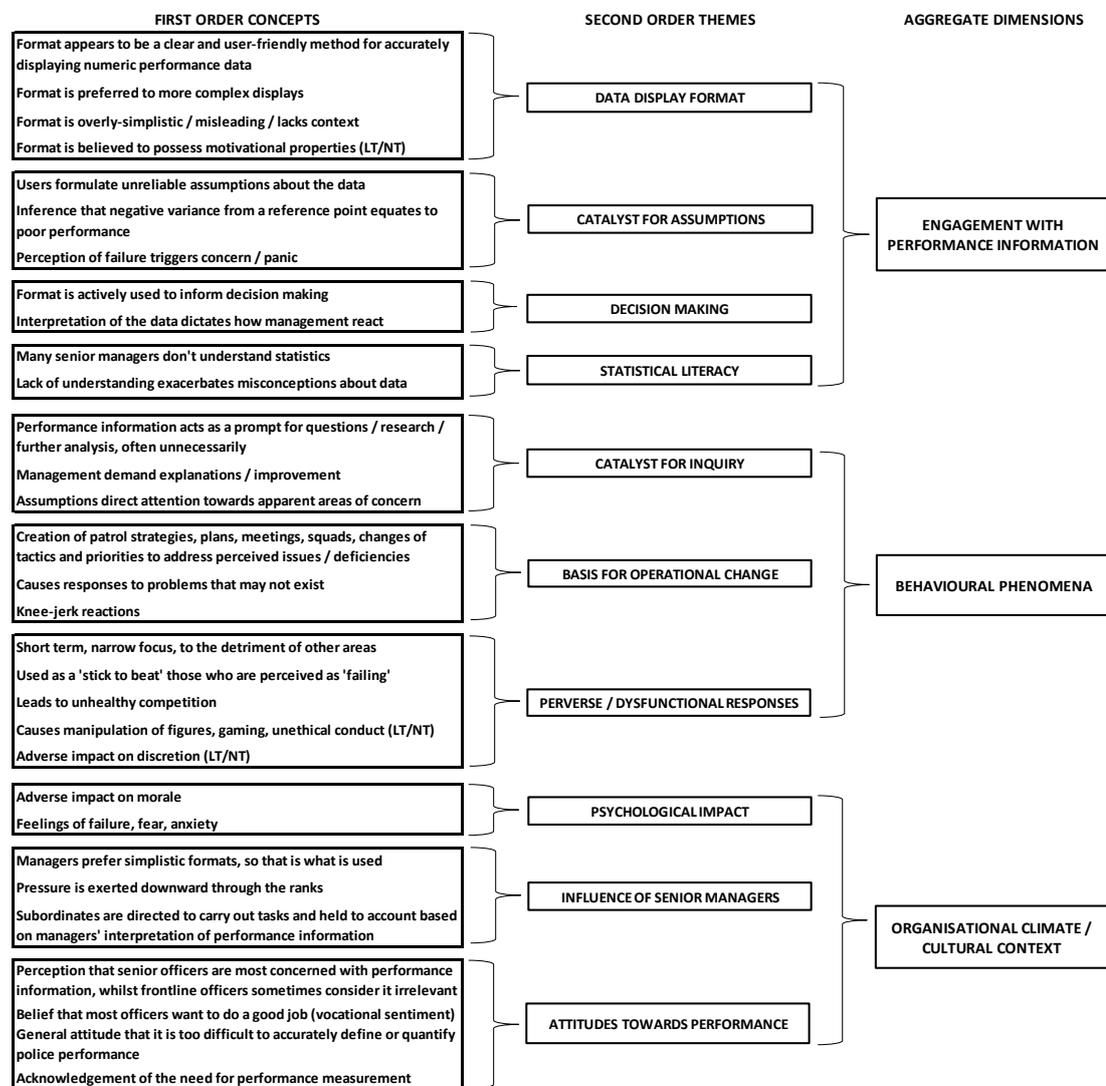
<sup>18</sup> Notes:

1. '(LT/NT)' next to certain first order concepts in the data structure signifies these concepts only emerged strongly in the League Tables and Numerical Targets categories.
2. Although the polarity of some first order concepts are opposing (e.g. 'Format appears clear and user-friendly' / 'Format is overly-simplistic and misleading'), they have been clustered together under the relevant second order theme (e.g. 'Data Display Format'), as the data clearly fit within that particular theme regardless of polarity; supporting narrative also incorporates comments provided from each perspective to ensure symmetry.

Next, a theoretical model is presented, which highlights relationships between themes, as well as a pathway between data display and behavioural dysfunction; it also postulates the existence of a causal mechanism which may partly explain the phenomena of interest. Interim conclusions are then presented regarding data display and its capacity to act as a catalyst for dysfunction; the subsequent chapter then explores this prospect further, utilising statistical methods to analyse quantitative data generated during the experimental component of the study.

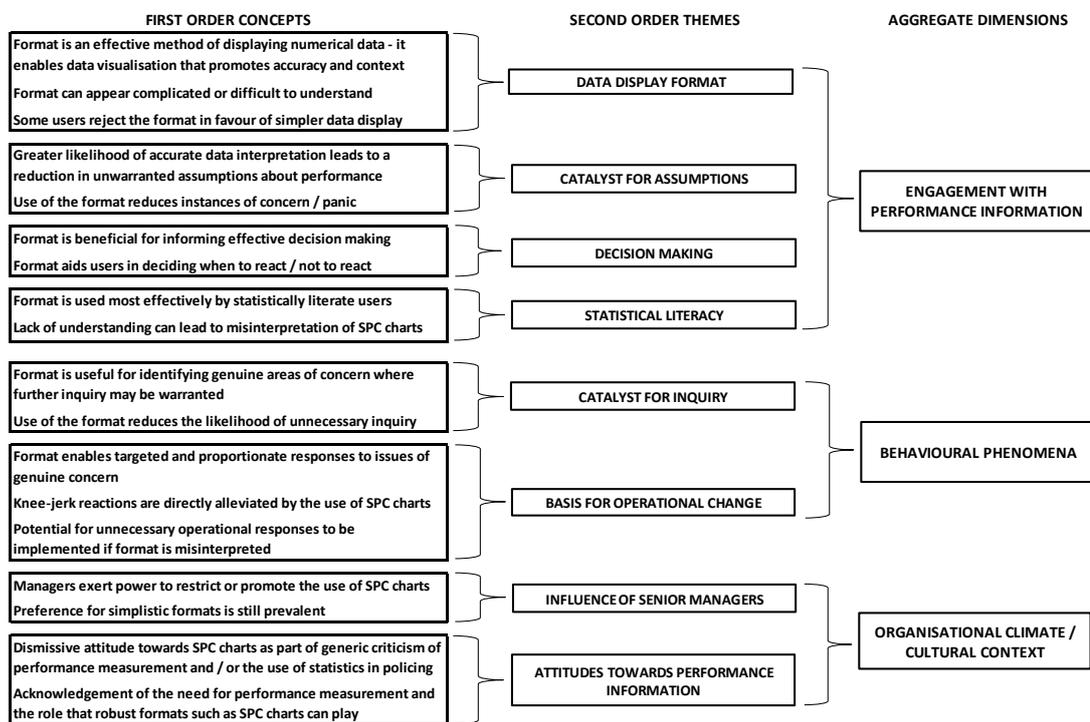
The primary data structure is presented at Figure 4.1:

*Figure 4.1: Data Structure: Binary Comparisons / League Tables / Numerical Targets*



In respect of the data structure for the SPC charts category, this comprised eight second order themes derived from 18 first order concepts; many of these components aligned with those present within the main data structure, whilst others were indigenous to the SPC category. As can be seen below, there were insufficient data to construct second order themes relating to psychological impact or behavioural dysfunction; therefore the SPC data structure features fewer second order themes than the previous one.

Figure 4.2: Data Structure: SPC charts



### **4.3.5 Aggregate Dimension: Engagement with Performance Information**

#### **4.3.5.1 Introduction**

This section focuses on the initial interaction between performance information and its users (typically police managers). It explores the thoughts, interpretations and assumptions of users when confronted with police performance information in varying formats and examines how different forms of data display may influence decision-making. It also considers how pre-existing knowledge of statistical concepts may affect these early interactions with the data.

#### **4.3.5.2 Second Order Theme: Data Display Format**

Whilst some of the subsequent themes to be discussed are comprised almost entirely of first order concepts exhibiting overwhelmingly negative sentiments, those that combine to produce the ‘Data Display Format’ theme are mixed. For instance, there is a clear split between respondents who believe that binary comparisons, league tables and numerical targets are useful performance information formats that offer clarity and accuracy, versus those who argue they are excessively simplistic and potentially misleading.

In contrast, whilst the discussion about the design and use of SPC charts also exhibits divergent viewpoints, these tend to focus on the perceived complexity of the format versus its usefulness as a more reliable form of data display. Personal preferences also feature as a strong first order concept across all performance information categories. Representative quotes can be viewed in Table 4.3, toward the end of this section.

In respect of binary comparisons, those expressing positive sentiments cite user-friendliness and simplicity of presentation, arguing the format is easy to read and understand, suggesting “...it puts statistics in a digestible format and highlights changes in figures simply” (BC551). There is also a firm belief that the format is capable of portraying data accurately, including the depiction of trends and changes.

However, others argue binary comparisons are inherently superficial and misleading, asserting the format “...does not provide any context or clarity and overly simplifies the presentation of data to a point that it has little meaning” (BC3202). League

tables, too, are accused of being crude and incapable of accurately depicting differences between peer groups, whilst numerical targets attract criticism for lacking depth and being founded on arbitrary reference points.

Others praised league tables and numerical targets for providing ‘clarity and focus’, claiming they also possess motivational properties. For example, one respondent insists a target, *“Spurs teams and individuals to perform and better themselves”* (NT2273), whilst another suggests, *“League tables clearly show the best and worst performing teams. Serves as a motivational tool to improve where possible”* (LT232).

A handful of respondents even seemed unable to envisage performance measurement without numerical targets. For instance, one respondent asks, *“Without a target what do we work to?”* (NT3472), whilst another reflects, *“We TRY not to use figure driven targets but how else do you gauge performance?”* (NT2489). This could indicate an inability to discriminate between this format and others, or possibly mean some respondents felt they needed a target as a reference point.

There was also a recurrent theme of reference-dependent formats being favoured even to the extent of respondents being dismissive of more reliable data display formats; this suggests personal preference for simplicity of presentation may come at the expense of statistical rigour. For example, one respondent asserts a target *“...demonstrates in simple terms areas of success or failure. Better than graphs and pie charts”* (NT422).

Such preferences may partly explain why these formats remain popular in UK police forces despite their limitations. Their visual appearance seems to promise the “...clarity, precision and efficiency” that Tufte (2001, p.13) argues for, yet due to their simplistic nature they tend to mislead performance information users. One respondent asserts, *“If you present the data in a certain way, this can distort the facts”* (FT3068).

In respect of the SPC charts category, respondents either praised the reliability of the format or rejected it for appearing too complicated. Those who highlighted its strengths cited statistical robustness, clarity and accuracy, arguing the format assists quick identification of patterns, trends and outliers. One officer argues it enables

*“...better presentation of performance data, a truer account of the bigger picture” (SPC2305), although another points out data are “...more commonly presented as a ratio between previous month and current month or as a comparison between the same period in the previous year” (SPC2723).*

Others disliked the format for appearing complicated and unappealing to the eye, insisting it appears *“...jumbled and incomprehensible” (SPC2457)* and difficult to interpret. Some not only rejected SPC for its apparent complexity, but expressed firm preferences for visually simpler forms of data display: *“Personally I prefer data to be numerical with a comparison year-on-year to dictate what action is necessary” (SPC2254)*, was a typical viewpoint.

These recurrent patterns highlight some of the key issues relating to data display formats commonly used to present police performance information. Fundamentally, initial engagement between performance information and its users invites interpretation; therefore if a format is overly simplistic, its design may adversely affect the quality of interaction with the data, especially as *“...statistics can appear better / worse dependent upon the format in which they are presented” (FT680).*

Table 4.3 contains additional typical statements regarding the use of different formats; the table is assigned a marker - ‘T1’ – the purpose of which is to enable the data pertaining to this theme to be cross-referenced with components of the theoretical model presented towards the end of this chapter. Subsequent themes are labelled in a similar fashion.

**Table 4.3: Second Order Theme 'Data Display Format' (T1)**

<b>Category</b>	<b>Representative Quotes</b>
<i>Binary Comparisons</i>	<p>“Easy to read and understand. Paints the picture clearly” (BC2734).</p> <p>“Accurate representation of crime trends” (BC1200).</p> <p>“Personally I like this format as it clearly shows the difference between the two months figures. This then shows the increase or decrease and the percentage of difference. Simple and effective” (BC1341).</p> <p>“It suggests there is an increase, when actually it may be completely normal” (BC3036).</p> <p>“When shown like this, data can be misconstrued” (BC554).</p>
<i>League Tables</i>	<p>“Good visual aid - the tables you have used have been clear and easy to follow” (LT1805).</p> <p>“They give a false impression of which team is performing well or not” (LT2564).</p> <p>“It presents data inaccurately as there is no context to it” (LT662).</p> <p>“It is too simplistic a tool to demonstrate a team’s performance” (LT89).</p>
<i>Numerical Targets</i>	<p>“Very clear and concise - no misinterpretation” (NT2787).</p> <p>“Easy to understand using a few figures instead of lots of confusing charts” (NT977).</p> <p>“Targets are important to raise standards and clearly define our priorities and objectives” (NT3640).</p> <p>“It is a simplistic approach to complex issues” (NT1327).</p> <p>“Numerical targets are picked out of the air (e.g. Achieve 90% of X... why not 88% or 92%?)” (NT1465).</p> <p>“Who sets the targets? Based on what? It seems to street officers someone sticks a pin in a chart and says ‘that will do’” (NT99).</p>
<i>SPC Charts</i>	<p>“It’s a better visual representation as you can see trends and identify where something is outside of the norm” (SPC3664).</p> <p>“SPC charts give a much more accurate and useful representation of data which understands natural variation” (SPC3670).</p> <p>“Visual representation in a graph form can be more impactful than raw data, and when set against a time line adds context” (SPC3355).</p> <p>“Visually clearer and more able to see a pattern emerging” (SPC664).</p> <p>“Very busy and difficult to read” (SPC1225).</p> <p>“In this format I believe it would be more difficult to understand - a numerical approach is far easier to understand in my opinion” (SPC1056).</p>

In summary, whilst some respondents cite design deficiencies and warn of misplaced faith in binary comparisons, league tables and numerical targets, others seem content to utilise them as a basis from which to draw inferences, based purely on variance from isolated reference points. Some respondents also warn that the more simplistic formats can be misleading, whilst suggesting the use of SPC charts promotes accurate interpretation and greater understanding of the data.

Ultimately, regardless of whether a particular data display format may be considered reliable or otherwise, its visual appearance potentially occupies a central role in shaping the nature of the interaction between performance information and user at the initial point of engagement. Therefore, data display seems well-placed to affect assumptions about performance, by influencing user's initial interpretations. The following theme explores this notion further.

#### **4.3.5.3 Second Order Theme: Catalyst for Assumptions**

This next theme moves beyond exploring performance information users' beliefs about (and understanding of) different data display formats, to examining what happens as a consequence of their interaction with them. It looks specifically at the assumptions that arise as a result of what the data appear to convey, as well as how this impacts on perceptions about performance itself.

Overall, with regard to binary comparisons, league tables and numerical targets, there were strong patterns of users formulating unsafe assumptions following interpretation of the data. There was also an overriding theme of negative variance from reference points leading to assumptions about poor performance and elevated levels of concern. Conversely, although some respondents expressed reservations about the perceived complexity of SPC charts, others suggested they help to moderate adverse reactions. Representative quotes can be viewed in Table 4.4, toward the end of this section.

Specifically in respect of binary comparisons, negative variance from the historical reference point gave rise to a false assumption of an increase in crime and the associated belief that this was due to police failings; for league tables, the dominant inference was that lower positions equate to poor performance, whilst negative variance from a numerical target was also widely interpreted to reflect failure. In

contrast, the contextualised nature of SPC data enabled users to assess the content more accurately, thereby preventing unwarranted assumptions about perceived trends or unsubstantiated areas of concern.

Once again, there was a noticeable split in how respondents reacted to the various data display formats. For example, some respondents were supportive of league tables, believing they “...*help identify weak team members or weak supervisors*” (LT1084) and “...*show who is and isn't working*” (LT148). Others however, questioned whether the format is an appropriate starting point for making assumptions about performance and described typical effects of its use:

*“There is a belief that being at the bottom is a failure and there is great pressure placed on those thought to be ‘not doing enough’. The general attitude is that if people are at the bottom they are simply not trying enough. Explanations are disregarded as being mere excuses” (LT2723).*

Similarly, whilst some respondents were confident that numerical targets help “...*identify areas of poor performance and where intervention tactics are required*” (NT3796), others point out the format “...*provokes senior officers to judge the target as a pass or fail*” (NT150), causing an “...*assumption that current performance is poor when this may not be the case*” (NT653). The chain of thought from interpretation to assumptions about performance was consistently evident; it is also noticeable how these reference-dependent formats directly frame users' thought processes and influence their perceptions.

Although the free text question enquired about the use of different data display formats in respondents' own forces, some respondents commented specifically on their interpretation of stimuli used within the survey instrument. This provided an insight into the assumptions triggered by these particular examples; for instance, those who provided such commentary alongside the binary comparisons stimulus were consistent in their assumption there was a verifiable trend depicting a rise in crime, indicating poor performance.

Respondents confirmed “...*when presented in this format the message is that crime IS increasing and inferred that we are not carrying out our basic function of preventing crime*” (BC214). Even those who harbour reservations about whether the

format is robust enough to be able to interpret the difference as a 'rise', or whether it is appropriate to draw such conclusions about performance, recount this is exactly what happens in police forces:

*"The assumption would be that this is an increase to be concerned about without taking into consideration the mean across a period of time, the normal range the figures move between and what this actually means. There would be remonstrations about poor performance and drives to improve" (BC3468).*

One of the consequences of these assumptions is that the perception of failure triggers concern, having *"...an effect on the Senior Management Team who run around like headless chickens if they see an increase in anything" (BC2635)*. Another respondent stated, *"It spreads panic among the innumerate" (NT3211)*. When such unwarranted assumptions involve judgments about perceived failure, it also seems entirely likely there could be damaging effects on those individuals or teams unfairly labelled as such.

Conversely, it appears the pattern of hasty or unreliable assumptions is reversed where SPC charts are used to present data. Whilst some users express reservations such as, *"They are not easy or straightforward for people to understand or draw inferences from" (BC1945)*, others argue the format *"...allows consideration of patterns and trends and leads to a greater understanding, which then allows you to tackle any issues effectively and provide a proportionate response" (SPC3302)*.

Some respondents also commented specifically on the SPC chart stimulus used in the survey instrument, with a typical statement being *"...it shows a stable picture. It is more informative than a small snapshot of information" (SPC3535)*. By displaying data in this format, this also appears to moderate undue concern of the type strongly associated with the other three forms of data display. For instance, one respondent observes, *"I just look to see that the figures are in the 'normal' range. I see the increases and decreases as natural so it doesn't concern me" (SPC1859)*.

Therefore, whilst it appears the use of SPC charts leads to a reduction in unwarranted assumptions and undue concern, it is notable that the other three formats all seem to behave as a catalyst for erroneous assumptions and associated elevated levels of

concern. Furthermore, such assumptions occur as a direct consequence of the flawed interpretations that routinely arise due to the limitations of these particular formats; one respondent muses, “*It is too easy to make assumptions on limited information*” (FT3820), whilst another offers an observation that is absolutely central to the notion of data display affecting perceptions:

*“Interesting how a different way of presenting the same information can change perceptions about performance” (FT2867).*

Table 4.4 shows additional representative statements relevant to this second order theme:

<b>Table 4.4: Second Order Theme ‘Catalyst for Assumptions’ (T2)</b>	
<b>Category</b>	<b>Representative Quotes</b>
<i>Binary Comparisons</i>	<p>“It is clearly demonstrating that crime has increased and that would be a concern” (BC3543).</p> <p>“Displayed in this format it appears crime is rising so we are performing poorly” (BC3251).</p> <p>“It creates a sense of panic; no one takes a long view or examines underlying trends” (BC3802).</p>
<i>League Tables</i>	<p>“It identifies teams which could be working harder” (LT1661).</p> <p>“The bottom team will always be perceived as underperforming” (LT3619).</p> <p>“The only effect is to highlight the team performing the worst, regardless of whether it falls within the ‘normal range’ of performance results” (LT1433).</p>
<i>Numerical Targets</i>	<p>“Identifies outliers and trends and is useful in this capacity” (NT3801).</p> <p>“If below the target the message is quite clear - You are not doing your job!” (NT214).</p> <p>“It has an effect of panicking insecure senior officers who do not understand statistics” (NT2855).</p>
<i>SPC Charts</i>	<p>“Clearer, more accurate and understandable picture” (SPC3515).</p> <p>“Better use of data as it shows fluctuations of crime but that they are all in the normal parameters” (SPC3578).</p> <p>“It actually shows some context to the data. Although there are spikes it shows that, on the whole, figures remain stable” (SPC1022).</p>

Overall, it appears that the choice of data display format directly influences the nature of users' interactions with performance information, causing a propensity to seek meaning from initial interpretation, regardless of whether the format is sufficiently reliable. Whether chosen formats happen to be either superficial or robust, consequent assumptions then become the foundation for decision-making and potential action.

#### **4.3.5.4 Second Order Theme: Decision-making**

This theme relates to the effect of performance information on decision-making and how the data display format can influence this process. It is the stage where users begin to formulate choices as a consequence of their assumptions about the data. As discussed within the literature review, many commentators assert that the primary function of performance information is to inform decision-making (Moynihan, 2008; Lavertu and Moynihan, 2012); it is therefore no surprise that this is exactly what occurs following initial engagement, interpretation and formulation of assumptions, regardless of the reliability of the format.

Two recurrent first order categories combine to produce this theme; firstly, there is universal agreement that the formats subject of this study are actively used to make operational decisions in police forces, and secondly, that the interpretations and assumptions that arise from their use dictate how managers react. Often this takes the form of deciding that perceived failure requires correction, leading to the types of inappropriate (and even perverse) responses that are explored within subsequent themes. Representative quotes can be viewed in Table 4.5, toward the end of this section.

Respondents confirmed, *"Information displayed in this fashion is frequent"* (BC1863) and *"...such basic data are used to drive senior officers' operational decisions and activity"* (BC2600). Such observations verify that the design of the stimuli used during the experimental phase of the study is indeed representative of the types of performance information used within the operational environment. Although some respondents warned that *"...without further information re the performance figure, decisions could be made for problems that don't exist"* (NT2621), there was broad agreement that even incomplete or potentially unreliable performance data usually provoked a decision to act.

Others acknowledged that the appropriate decision may in fact be to ‘do nothing’, although respondents recounted that managers are usually reluctant to refrain from reacting, even though “...data is [sic] often presented in a format that is misleading and untrustworthy” (FT2696). (This prospect will be tested during the experimental phase of the research, to assess whether performance information users do indeed tend to act, even when confronted with potentially unreliable data display formats). One respondent observed, “When presented with figures in isolation there is never a ‘do nothing’ option, even though the resultant effect may not be effective or necessary” (FT3624).

Respondents also reported that managers’ (mis)interpretation of the data and their consequent formulation of assumptions ultimately dictates how they decide to respond. For example, one observes that the binary comparison format “...directs senior officers’ focus and their decision-making regards operational resourcing / responses” (BC2425), whilst another suggests there is “...pressure to meet targets which can influence operational decisions” (NT1576).

Although managers attract particular criticism for using unreliable data display formats as a basis for decision-making, it was evident that decisions made by officers at all levels can be affected by their interpretation of performance information. For instance, one respondent asserts league tables “...can be destructive, negatively influencing decision-making and causing rifts between individual teams” (LT2517). Another claims the use of numerical targets “...result in perverse decisions and tactics, as well as wasted time, effort and money” (NT2336).

Conversely, respondents reported that the SPC chart format is beneficial for aiding effective decision-making, as “...it enables a long term view and avoids concentrating on short term natural fluctuation - allows more informed decision-making” (SPC2869). Others noted that as the SPC chart stimulus used in the experimental phase depicted a stable data set, this enabled them to feel comfortable in deciding *not* to react, contrary to the experience of the majority of users who envisaged adverse issues when interacting with the other three formats.

Table 4.5 shows additional typical statements relevant to data display and decision-making:

**Table 4.5: Second Order Theme ‘Decision-making’ (T3)**

<b>Category</b>	<b>Representative Quotes</b>
<i>Binary Comparisons</i>	<p>“Crime figures have been presented in this way for years” (BC2660).</p> <p>“This is exactly how we do it...” (BC2709).</p> <p>“Basic stats like this fail to show the full meaning of the figures but decisions are still based on snap shot figures such as above” (BC3378).</p> <p>“Does not give a detailed picture of what is really occurring and this leads to inappropriate activity from poor decision-making” (BC3775).</p>
<i>League Tables</i>	<p>“It can push people to make decisions they might otherwise not make” (LT1632).</p> <p>“Influences certain officers’ decisions and results in situations as with Team 2 at [Redacted] where discretion is rarely shown and the decision to arrest is made simply to generate another crime arrest to be collated at the end of the month when team figures are compared” (LT1563).</p>
<i>Numerical Targets</i>	<p>“It has a negative effect as officers chase the target rather than making decisions with integrity” (NT2284).</p> <p>“Causes poor decisions by management” (NT3612).</p> <p>“It causes decisions to be made to satisfy the meeting of the target” (NT973).</p>
<i>SPC Charts</i>	<p>“The sample range is large enough so that informed decisions can be made” (SPC2650).</p> <p>“It clearly displays a larger amount of data from which more informed decisions and plans can be made” (SPC551).</p> <p>“Control levels add useful context to inform action or lack of necessity thereof” (SPC3521).</p>

Overall, the consequences of using unreliable forms of data display as a basis for deciding whether to take action are that poor choices are made; as a result of what is perceived, users envisage issues which may not be present and decision-making is adversely affected. Managers may then elect to instigate further inquiry or initiate operational responses, intended to address perceived issues (these conditions are explored in due course). This arises as a direct consequence of the assumptions triggered by interaction with performance information, the nature of which are ultimately shaped by the design of the chosen data display format.

#### 4.3.5.5. Second Order Theme: Statistical Literacy

Although not necessarily sequential (unlike other themes discussed herein), the final theme in this dimension is that of statistical literacy and its implications for data presentation and use. In particular, respondents reported that many senior managers are untrained in the use of statistics and therefore ill-equipped to interpret data. This lack of understanding leads to unfounded beliefs about the capability and reliability of some data display formats and exacerbates misconceptions about perceived patterns observed within data. It also influences choices about preferred formats.

Conversely, where users are statistically aware, this assists in the selection of robust formats and acts as a condition that prevents individuals from engaging with data presented in unreliable formats, thereby potentially moderating adverse effects downstream. Representative quotes can be viewed in Table 4.6, toward the end of this section.

Overall, the lack of statistical awareness was considerable; some respondents even deemed themselves statistically literate but exhibited profound misunderstanding about the capability of certain formats. For example, one respondent claims binary comparisons “...make it easy to understand where we are and what we need to do - I am in charge of Local Performance so my understanding may be better” (BC3504). Similarly, another respondent displayed an erroneous belief about statistical significance, asserting “...in terms of stats anything above or below 2 to 3% is significant” (BC1841).

Respondents consistently bemoaned what they perceive as senior managers’ inability to grasp basic statistical concepts, describing their use of performance data as “...simplistic and reactionary...” (BC2723), where “...the vast majority of managers clearly do not know anything about the interpretation of statistics” (FT1733). One summed up the situation as follows:

*“Most statistics we see seem to presented / interpreted at senior management level by people with no depth of understanding about data reliability, statistics, error, variation, standard deviation etc” (NT1788).*

Whilst being critical of senior managers’ lack of statistical awareness, many respondents recognised this as a consequence of little or no training. One officer

observed, “...performance is managed by police officers who have moved into senior ranks but who may have had little academic or vocational background in understanding how to manage performance” (FT2568). Another simply stated, “Managers need more training in statistical analysis” (FT2167).

Respondents warn that such a fundamental lack of understanding routinely causes misconceptions about data and affects how managers respond. For example, one argues “...there is often no understanding of basics (mean, median, mode, standard deviation) which means often very senior officers react in the wrong way to statistical presentations” (FT3204). The consequences are that if “...officers / staff are not able to fully interpret the data it could lead to misconceptions about performance” (NT2446).

In respect of SPC charts, although those familiar with the format assert it “...aids better understanding” (SPC1323) due to being “...a very quick and effective method of presenting statistics” (SPC1711), those lacking statistical awareness sometimes struggle to interpret the data and are therefore reluctant to use such charts. Again, lack of training is cited as an obstacle to their use, as “...people don’t always have the knowledge of statistics to be able to interpret charts of this kind” (SPC3490); this can lead to a situation where users “...tend to only read the last part of the graph to see if crime is up or down” (SPC1339). This suggests that even where reliable formats are available, if users are unable to interpret them properly, this could cause disengagement or misuse.

It was clear from responses that the issue of statistical literacy transcends the individual data display formats under scrutiny and reflects a fundamental problem within policing. This was evident in the commentary provided, which consistently described a staggering lack of awareness of basic statistical concepts amongst senior managers, leading to a preference for the use of ‘blunt’ tools and reluctance to engage with more reliable formats; one respondent claims, “In general, statistical understanding by police officers is dire” (FT3205).

Furthermore, pre-existing statistical awareness appears to be a condition capable of moderating adverse cognitive reactions to performance data; respondents who recognised the limitations of simplistic forms of data display were unwilling to engage with such formats and this could potentially halt the sequence of events

whereby users formulate unsafe assumptions that impair decision-making. Conversely, those who possess insufficient knowledge to question whether particular designs are of sufficient depth were prone to actively engaging with unreliable formats.

Lack of statistical literacy therefore seems to reinforce the types of unwarranted assumptions made by performance information users, as well as potentially inhibit the use of reliable forms of data display. Additionally, it does nothing to prevent defective decision-making, as users are not equipped to intervene in their own thought processes by questioning the validity of the data display format.

It may well be the case that training in basic statistical concepts would present an opportunity to mitigate the poor design and application of police performance information, as the knowledge gained could encourage the selection and use of reliable data display formats. It may also help prevent potentially damaging interactions with overly simplistic formats, as users would be aware of their limitations and disengage from the outset.

Table 4.6 shows additional typical statements regarding the topic of statistical literacy and its relevance to the use of different formats:

**Table 4.6: Second Order Theme ‘Statistical Literacy’ (T4)**

<b>Category</b>	<b>Representative Quotes</b>
<i>Binary Comparisons</i>	<p>“Not all our SLT [<i>senior leadership team</i>] understand how to interpret performance data correctly and so would see this at face value” (BC3051).</p> <p>“In my experience the police use of statistics is damaging because of the level of ignorance shown in data collection, processing and interpretation” (BC1863).</p> <p>“Misleading presentation of short term trends can paint an inaccurate long-term picture and lead to inappropriate policy responses” (BC2306).</p>
<i>League Tables</i>	<p>“Generally a negative effect on the teams involved, with management unable to see this and not accepting there is a ‘normal’ range - someone always has to be at the bottom, but that does not mean there is a problem” (LT3190).</p> <p>“It has an effect on those individuals who fail to appreciate that statistics and charts are open to many different interpretations other than that which appears to be the obvious” (LT1502).</p>
<i>Numerical Targets</i>	<p>“There is never an explanation for the reason a particular target is set, and you are then challenged if you fail to achieve this arbitrary target. Where the target set is based on an average there is no recognition of the fact that with an average half of the numbers must be below the average!” (NT3737).</p> <p>“Managers fail to understand that figures fluctuate over time and that over the longer term, tend to balance out” (NT2996).</p> <p>“Used without thought by people who are not qualified to understand” (NT55).</p>
<i>SPC Charts</i>	<p>[SPC charts are] “...distributed to, and expected to be understood by police officers - not analysts, business managers, statisticians or accountants; but cops. Cops who have been skilled in law enforcement, community engagement, investigations, intelligence - not performance management” (SPC2550).</p> <p>“It is not presented with an explanation or guidance on how to interpret it so we often revert to type i.e. we have gone up / down compared to last month” (SPC3737).</p>

In summary, these observations indicate statistical literacy is an absolutely critical factor when it comes to selecting and using data display formats; if users are unaware of limitations, this exacerbates the risk of unsuitable formats being preferred, along with the likelihood of misinterpretation, leading to unwarranted assumptions and impaired decision-making. Overall, statistical literacy is found to be a pivotal condition affecting the design and use of performance information.

## **4.3.6 Aggregate Dimension: Behavioural Phenomena**

### **4.3.6.1 Introduction**

This section examines the types of behavioural outcomes that occur following engagement with performance information; as a consequence of the ways that particular data display formats affect users' thought processes, interpretations and assumptions, what actions do they tend to undertake and what is likely to occur as a result?

### **4.3.6.2 Second Order Theme: Catalyst for Inquiry**

The first theme in this area considers how users' interpretation of performance information prompts questions or further analysis into perceived issues. In circumstances where the data display format is robust enough to enable reliable assumptions to be made, this would be an entirely appropriate next step; however where interpretations and assumptions are made following engagement with unreliable formats, such incentive for inquiry could arise as a result of misinterpretation, meaning that it may not actually be appropriate to intervene.

In respect of binary comparisons, league tables and numerical targets, responses were once again split between those who hold the belief that they are meaningful data display formats, versus those who argue they are not a legitimate starting point for action, as it is likely the assumptions that trigger such action could be flawed. Meanwhile, those who commented on SPC charts were insistent this format affords the necessary depth to identify when and where it is appropriate to delve further and when it is not. Representative quotes can be viewed in Table 4.7, toward the end of this section.

Respondents recount that in the case of the first three formats, engagement consistently “...triggers questions / further explanation” (BC3774). Some suggested binary comparisons “...easily highlight an increase in crime figures, worthy of further analysis to identify cause and solution options” (BC2327), observing, “...questions are asked about why there has been a rise. Answers are expected” (BC3766).

Similar thought processes are activated during interactions with league tables and numerical targets; respondents note that league tables are used “...to prompt further enquiries to find out the underlying reasons behind differences in performance” (LT1149), whilst numerical targets are considered to be “...a clear indicator; it shows an issue and ensures questions are asked as to ‘Why?’” (NT2836).

Although these formats act as “...a prompt to ask for further research and information” (NT3787), there tended to be an emphasis on perceived failure, which in turn led to demands for improvement. For instance, one respondent describes a situation where, “...managers look at the stats and look for improvement without understanding the picture” (BC1445). In some cases, “...teams are told: ‘If Team X can achieve that then what is going wrong on your team?’ (LT315). Another officer recalls an example where failure to achieve a numerical target by even the narrowest of margins led to an adverse reaction:

*“We once were told that we missed a target by 0.1%. Management demanded improvement and have threatened to discipline those who do not contribute to performance” (NT459).*

Therefore, as a consequence of the assumption that negative variance from a reference point equates to failure, managers routinely believe there is an imperative for inquiry and focus their attention on apparent areas of concern. The danger here is that use of unreliable data display formats may cause users to fixate on inconsequential (or even non-existent) issues to the exclusion of unobserved matters of genuine concern. In other words:

*“It triggers managers to believe that there has been a marked fall in performance and to respond by taking some step or other to address a perceived performance problem” (BC3261).*

Such management behaviour was identified within the literature review as being a precursor to dysfunction; examples of managers demanding explanations or improvement for apparent deficiencies were found to be a likely catalyst for adverse behavioural responses and unintended consequences (Eterno and Silverman, 2012). This finding therefore opens up the possibility there may be a connection between certain data display formats and particular antecedents to dysfunction. (This prospect

will be explicitly tested in the experimental component of this study, to ascertain whether data display acts as a direct influencing factor upon the likelihood of such reactions).

In the case of SPC charts, although similar curiosity about performance is triggered by engagement, the comparative richness of the format enables users to focus upon genuine issues of concern, thereby preventing unwarranted inquiry. As discussed in the previous themes, this is likely the result of better interpretation, accurate assumptions and more rounded decision-making; the common factor being that whilst engagement with performance information acts as a catalyst for certain cognitive responses, the appearance of the data display format dramatically shapes the nature and direction of these responses. As summarised by one respondent:

*“This is a more balanced and honest approach. It shows the normal number and range. It shows balanced performance and indicates peaks where further research may be necessary to deal with high crime areas / times”*  
(SPC3088).

Therefore, regardless of the format used, by this point a chain reaction is well underway. Where the data display format is insufficiently robust, this consistently leads to misplaced belief that further analysis or explanation, improvement or action are required, along with unwarranted attention towards minor or groundless concerns. Conversely, where the format is reliable, this enables users to be discriminating about when, how, or if intervention is required. Either way, the user is now at the stage whereby they may choose to implement operational changes; the implications of this will be explored in the forthcoming themes.

Table 4.7 shows additional typical statements demonstrating how performance information acts as a catalyst for inquiry:

**Table 4.7: Second Order Theme ‘Catalyst for Inquiry’ (T5)**

<b>Category</b>	<b>Representative Quotes</b>
<i>Binary Comparisons</i>	<p>“It shows that crime is increasing, so prompts us that we need to find out why, and then put measures in place to change the pattern” (BC1037).</p> <p>“It will cause officers to research the reasons (or perceived reasons) for the increase and prepare to offer an explanation at the monthly performance meeting” (BC3797).</p> <p>“Senior managers will demand an improvement” (BC2367).</p>
<i>League Tables</i>	<p>“It may initiate questions as to why one area or team is different to another (LT3841).</p> <p>“Look at performance of that team and ask questions why they are not performing as good as the others” (LT1162).</p> <p>“Team at the bottom is perceived as failing and there is an expectation that improvement will be made” (LT3746).</p>
<i>Numerical Targets</i>	<p>“Numerical targets are set and questions asked if not achieved” (NT3163).</p> <p>“People can clearly see that improvement is required” (NT2481).</p> <p>“One is expected to reach these targets and if not, questions are asked as to why one has not reached them” (NT431).</p>
<i>SPC Charts</i>	<p>“It can give you an indication of direction of travel to enable to you ask questions” (SPC3779).</p> <p>“It gives us an understanding and then allows the question of ‘why’ to be asked before taking action” (SPC3225).</p> <p>“Mature discussions about the current situation. Consideration of reasons for variation and focus on the root causes” (SPC3301).</p>

In summary, respondents’ experiences indicate that the interaction between performance information and its users consistently leads to considerations about whether to initiate further inquiry; furthermore, that there is a propensity for doing so where users envisage perceived deficiencies. Although not at the extreme end of perversity, where unreliable data display formats are used as the basis for instigating inquiry, such intervention may be considered dysfunctional (as well as being an antecedent to wider dysfunction); however, this tendency was found to be significantly moderated where the chosen format was sufficiently reliable.

#### **4.3.6.3. Second Order Theme: Basis for Operational Change**

In addition to the unnecessary inquiry prompted by engagement with unreliable data display formats, respondents also reported that a vast array of unwarranted or disproportionate operational responses tend to be implemented. These include the creation of patrol strategies, written plans and meetings, the inception of new squads, as well as sudden changes of tactics and priorities. This consistently arises when users' interpretations and assumptions cause a belief that action is required to rectify apparent deficiencies.

Such unnecessary operational responses were found to be strongly associated with the use of binary comparisons, league tables and numerical targets; respondents repeatedly cited 'knee jerk' reactions to data when these formats are used. In contrast, the use of SPC charts consistently produced more targeted and restrained deployment of tactics and resources, as well as a reduction in disproportionate or unwarranted operational responses. Representative quotes can be viewed in Table 4.8, toward the end of this section.

Respondents reported that the first three formats were habitually *"...used by senior officers to determine where resources need to be"* (BC1234). Consequently *"...patrol strategies and taskings would be implemented"* (BC1072), *"...staffing levels are changed, priorities are moved, specific areas targeted"* (NT3245). Some respondents suggested such responses were appropriate, as performance information presented in these formats appears to enable managers to be *"...aware of where to concentrate resources to address increases and introduce reduction plans"* (BC3413).

Others observed that the perceived deficiencies envisioned as a consequence of viewing data in such formats are often accepted without question, ultimately leading to the belief there is a valid mandate for action. For example, commenting on the binary comparison stimulus, one officer stated, *"It will lead to an operational plan to combat the increase"* (BC3000), whilst another observed *"...resources and activity are shifted to target the fall in performance"* (NT3245).

However, many were concerned that these formats *"...portray an image of crime increasing and may lead to action being taken when the reality is that crime may be*

*stable*” (BC2604), thereby promoting “...reactions to a perceived trend that does not exist, wasting resources” (BC734). Respondents reported this causes “...frenzied activity...” (LT3583) and “...much management fuss. Taskings. Quality checks. Dip samples. Extra meetings. Action plans. Posters. Many hours discussion to identify the problem” (NT3733); this not only absorbs capacity, but also leaves other areas exposed.

It is quite possible those who implement such operational changes do so with good intentions in the belief there is a genuine problem requiring prompt resolution, however, if assumptions about performance arise from the use of unreliable data display formats, then there is no mandate to respond in this way. Additionally, managers directing unnecessary remedial action or operational activity is reflective of antecedents to dysfunction discussed in the literature review (Home Office, 2015).

A recurrent observation that also emerged was the accusation that senior managers’ responses to data presented in these formats were akin to ‘knee-jerk’ reactions. Respondents consistently spoke of “...a massive knee jerk reaction involving a redeployment of resources” (BC2876) as being the norm. This may be a consequence of simplistic data display formats being limited to portraying variance from isolated or unstable reference points, leading to a binary narrative about ‘improvement’ or ‘deterioration’, or ‘good’ or ‘bad’ performance, with no provision for gradation, scale, or overall context.

Conversely, SPC charts appear to present an opportunity for largely avoiding such unnecessary reactions, by reframing users’ perceptions about data content and thereby stunting the faulty assumptions that lead to inappropriate operational responses. Respondents point out the format “...gives a much clearer picture of the data that is being captured and this allows the organisation to direct resourcing and solutions accurately” (SPC2242). Some even suggest it specifically alleviates the likelihood of “...knee-jerk action to resolve a problem which may not be actually present” (SPC3203), by encouraging “...better understanding and comprehension of the overall picture” (SPC1483).

Therefore, the SPC format appears to offer the dual benefits of not only tending to prevent inappropriate reactions, but also promoting swift identification of genuine issues of concern and the inception of targeted and proportionate responses.

However, there is a proviso relevant to statistical literacy; those unfamiliar with SPC may misinterpret the charts, thereby leading to inappropriate operational responses, regardless of the robustness of the format itself. As explained by one respondent:

*“Because many of our senior leaders lack a basic grasp of statistical analysis, such a presentation can lead to misinterpretation and knee-jerk changes of business processes in an attempt to tackle a problem that does not exist” (SPC3671).*

Therefore, care must be taken regarding the use of SPC charts, as if users are ill-equipped to interpret them, this can also lead to unnecessary operational changes (albeit on a smaller scale than with the other formats). Consequently, it is imperative that users are sufficiently equipped to identify and utilise reliable data display formats, in order to prevent misdirection of resources and organisational capacity being wasted.

Table 4.8 shows additional typical statements regarding the way performance information acts as a spur for operational changes:

**Table 4.8: Second Order Theme ‘Basis for Operational Change’ (T7)**

<b>Category</b>	<b>Representative Quotes</b>
<i>Binary Comparisons</i>	<p>“It would show the areas for improvement or change of tactics needed” (BC1081).</p> <p>“Managers initiate responses to problems that are not there” (BC3043).</p> <p>“Encourages knee-jerk reactions, then back-slapping when regression to the mean occurs and crime drops!” (BC2301).</p>
<i>League Tables</i>	<p>“Usually results in a knee jerk operation to target the perceived problem resulting in other areas being neglected” (NT1898).</p> <p>“The effect is that there is often a lot of activity driven with the production of action plans etc without an evidence base” (LT3668).</p> <p>“Knee jerk response to deployment of staff and resources with limited analysis of the problem or effective methods of dealing with them” (LT3486).</p>
<i>Numerical Targets</i>	<p>“Yes, this would increase focus, performance would be able to be measured and operations and other initiatives could be introduced to strive in order to achieve the target” (NT1580).</p> <p>“It sucks all resources into one area for a period of time, causing a spike elsewhere, then all resources are directed there, repeat <i>ad infinitum</i>” (NT480).</p> <p>“Abstraction of resources to perceived performance issues where there may well not actually be an issue” (NT3060).</p>
<i>SPC Charts</i>	<p>“Allows patterns to be identified enabling resources to be better targeted” (SPC1262).</p> <p>“Can identify trends and assist with resource deployment” (SPC460).</p> <p>“I believe that this method puts into perspective fluctuations in crime levels and prevents ‘knee jerk’ reactions” (SPC1549).</p> <p>“Can be helpful, but senior management so unused to seeing this have no acceptance of ‘normal range’ and feel compelled to ‘do something’” (SPC3190).</p>

As a final point, the types of operational responses discussed within this theme may not necessarily be perverse in the same way as deliberate gaming, but where unwarranted action is taken, this still represents an inappropriate (and therefore dysfunctional) reaction; furthermore, such responses may also act as an antecedent to broader dysfunction and unintended consequences.

#### **4.3.6.4 Second Order Theme: Perverse / Dysfunctional Responses**

Whilst the previous theme examined dysfunction at the ‘thinner end of the wedge’, this theme explores respondents’ experiences of more serious behavioural dysfunction and its relationship with different performance information formats. The survey instrument produced extensive free text content closely reflecting the types of perversity recounted within the literature. Of paramount importance here, however, is whether certain types of behavioural dysfunction are most strongly associated with the use of particular data display formats; where other variables remain constant, do certain formats appear to exacerbate or moderate such behaviour?

Overall, the first order concepts within this theme highlight consistent forms of dysfunction associated with the use of binary comparisons, league tables and numerical targets, albeit to varying intensities. Furthermore, outright perversity was found to be most extensively associated with the use of numerical targets, although league tables were also found to trigger similar patterns of behaviour. In some cases this bordered on unethical and potentially unlawful conduct. Representative quotes can be viewed in Table 4.9, toward the end of this section.

Analysis also found that the patterns of behaviour associated with these three formats were strongly demarcated from those connected with the use of SPC charts. Indeed, the only adverse behavioural tendencies associated with SPC chart usage involved occasional unnecessary operational responses arising as a result of misinterpretation, as discussed within the previous theme. There was no commentary provided regarding gaming or other perverse use at all; this does not, of course, suggest SPC is immune from being associated with dysfunction, however, it does give an indication as to the general disposition of effects usually connected with its use.

In respect of binary comparisons, although respondents provided examples of perverse behaviour, this theme was not as pronounced as it was for league tables and numerical targets; dysfunction tended to be limited to short-termism, improper use of performance information by managers and the likelihood of triggering unhealthy competition. (These conditions were also evident for the other formats, along with additional, more extreme, reactions, as discussed below).

One recurrent notion was that of managers using performance information as a ‘stick to beat’ those perceived as failing. For example, referring to binary comparisons, one respondent observes “...when presented in this format it is used as a stick to beat perceived poorer performing teams” (BC2520), whilst another recounts similar use of numerical targets: “I have seen such data used as a stick to beat otherwise highly motivated employees” (NT3038).

A further type of reaction (common across these three formats) was in respect of unhealthy internalised competition being triggered in response to assumptions about performance. In an effort to “...get the reds changed to greens by whatever means” (FT3625), teams and individuals vigorously implemented the types of unnecessary operational responses discussed previously, “...causing districts to compete against each other and focus locally and not on the bigger picture” (BC3667).

Therefore, it appears from respondents’ experiences that whilst binary comparisons can indeed trigger perversity, their greatest influence on behavioural dysfunction is manifested in the inappropriate operational responses enacted by senior managers as a result of misinterpretation. Furthermore, respondents reported little in the way of beneficial outcomes; it seems the best case scenario when binary comparisons are presented is where users recognise their limitations and disengage prior to taking action (i.e. the best ‘response’ is not to respond).

Overall, perversity associated with league tables and numerical targets was more extensive. In respect of league tables, a prominent symptom was unhealthy competition, with individuals “...focussing on things that move you up the table rather than focussing on what matters” (LT3476). Whilst some respondents insisted league tables “...can motivate by creating healthy competition” (LT1481), others argued the format “...engenders a dangerous and unhealthy level of competitiveness” (LT1483), which “...creates acrimony, distrust and unwelcome behaviour via unhealthy competitive practices” (LT1727).

Respondents consistently cited negative effects when data are presented in league tables, creating an adversarial climate where individuals and teams engage in a range of perverse practices, damaging organisational cohesion and interdepartmental cooperation. Examples include ‘cherry picking’, a narrow focus on measured dimensions to the exclusion of areas not pertinent to the league table, adverse impact

upon discretion and even manipulation of figures. One respondent summarised the main effects as follows:

*“They have three effects - mid ranging teams tend to sit back and relax, safe in the knowledge they are neither the best nor the worst. The top ranked teams tend to gloat to their colleagues, deriding their poor standing in the league tables. They also strive to continue being top and thus - potentially - break the rules to do so, or at very least adopt gaming tactics (as an example, a pub fight would have multiple separate crimes recorded dependent on how many detections or arrests can be attained). Finally, bottom ranked teams end up chasing arrests / detections / whatever is being ranked and again, this leads to dysfunctional behaviours and gaming tactics” (LT3248).*

Similar types of behaviour were observed in respect of numerical targets, where, if anything, the nature and extent of perversity was even more extreme than with league tables. Respondents reported that targets became the *de facto* purpose for activity, resulting in unhealthy competition, over-zealousness, primacy of quantity over quality, manipulation of figures and other unethical practices. These behaviours closely match those recorded within the literature and according to one respondent, arise because, *“...the target becomes the most important thing; not the victim, integrity, honesty, but the target” (NT3177).*

Overall, respondents were emphatic that *“...people will alter their behaviour to game the system...” (NT644)* when numerical targets are present. Examples included *“...withholding intelligence from other teams...” (NT3716)*, *“...chasing easy to achieve results at the expense of longer term hard-to-accomplish solutions” (NT2205)* and *“...a ‘no discretion’ way of working” (NT1901)*. Respondents also recounted instances of target-driven behavioural dysfunction that were quite simply bizarre:

*“Our force contact team prides itself on contacting victims every hour to advise them that no patrols are available to attend an incident reported hours previously – when in fact all this does is serves to remind the caller that we have not met their expectation. Ironically the fact the call is made is counted as a positive indicator for the department” (NT3723).*

A particular area of concern that emerged was the strong association between numerical targets and “...*ethically questionable behaviour*” (NT1811), including “...*unethical detections, bad practice and corruption*” (NT449), as well as “...*misconduct in an attempt to meet what often appears to be an arbitrarily set target*” (NT987). Officers cited “...*unlawful stop and searches, arrests that were inappropriate...*” (FT1522) and claimed that crimes were routinely misrecorded, downgraded, or generally ‘massaged’; some even alleged that a minority of “...*officers break the law to meet the targets*” (LT2405), leading to “...*a culture of manipulated figures*” (NT2073). For example:

*“Our organisation has recently started throwing people in to offices to frantically ‘no crime’ as much as they can... reviewing all crime reports and trying to bin as many as they can insisting they are not necessarily crimes. My personal favourite was the smashed double glazed window which they informed me was not necessarily vandalism as ‘it could have been a bird strike’. Yes, yes of course it could. How silly of me not to think of that and inform the irate complainer of same” (FT382).*

These types of behaviour are reflective of cases of gaming, misreporting and perversity recorded in the literature. Furthermore, it is notable that of all data display formats examined, such dysfunction is most strongly associated with numerical targets. This could suggest that the format is prone to triggering particularly extreme adverse reactions due to the aspirational nature of the reference point; therefore the experimental phase of the research will further explore whether this relationship with dysfunction is more pronounced with numerical targets than other formats.

Furthermore, both the free text responses and extant literature suggest league tables and numerical targets are more strongly associated with behavioural dysfunction than binary comparisons; it may be that whilst binary comparisons depict variance capable of triggering unnecessary remedial action, unlike league tables and numerical targets, they are not necessarily explicitly designed to induce competition. League tables and numerical targets may also exert greater motivational power to instigate dysfunction as they are often accompanied by sanction and reward.

In respect of SPC charts, respondents did not report a relationship between their use and perversity at all; this may be because those who understand how to interpret this

form of data display tend to use it appropriately, thus mitigating adverse behavioural tendencies. Alternatively, users who are not statistically aware may be reluctant to engage, meaning its use tends to be limited to occasions whereby users are proficient at interpretation and able to assess whether operational responses are necessary. Although there is the possibility some users may interpret SPC charts incorrectly and initiate unnecessary operational changes, it seems this is the extent of undesirable behavioural outcomes associated with the format.

Table 4.9 shows additional representative statements regarding the ways in which the use of different performance information formats appears to influence behavioural dysfunction:

**Table 4.9: Second Order Theme ‘Perverse / Dysfunctional Responses’ (T7)**

<b>Category</b>	<b>Representative Quotes</b>
<b>Binary Comparisons</b>	<p>“Focus on short term changes, which may not be statistically significant” (BC27).</p> <p>“This results in the ‘daily beatings’ at DMM [<i>Daily Management Meeting</i>] (BC699).</p>
<b>League Tables</b>	<p>“Whilst many officers refer to the ‘healthy’ competition that are league tables, it is clear that creating an adversarial system whereby one team attempts to out-do another causes dysfunction. Rather than focusing on the task, process or issues at hand, the team become concerned with their position in a table” (LT3649).</p> <p>“It causes officers to worry about performance and begin to concentrate on crimes which tick boxes and ignore other types of incidents” (LT1268).</p> <p>“It can encourage officers to be over-zealous, taking away the officers’ discretion in some respects and alienating the public” (LT1536).</p> <p>“It has a very negative effect - increases gaming and manipulation of figures” (LT3660).</p>
<b>Numerical Targets</b>	<p>“The focus of management is on achieving and exceeding the targets as opposed to providing a quality service” (NT535).</p> <p>“Officers when given these targets tend to ‘cherry pick’ jobs, easy shoplifters etc... They try and steer clear of more complex and usually more serious jobs... Detect 12 shoplifting offences, great!!! Detect one serious robbery, ‘Oi, my office - your detections are low!’” (NT2122).</p> <p>“It can lead to perverse incentives e.g. resources targeted to improve figures rather than quality of service. (‘Low hanging fruit’)” (NT3263).</p> <p>“I have 6 months of 30 years to serve and am constantly pestered to bring in 2 arrests a month” (FT487).</p> <p>“We have all witnessed figures being manipulated for show; burglaries turned into criminal damage, GBHs [<i>Grievous bodily harm</i>] made into ABHs [<i>Actual bodily harm</i>]. We’ve all been told stop search isn’t used for performance, but we’ve all been told in quiet corners that we had better do more if we want that course we’re after” (FT286).</p> <p>“This results in ‘massaging’ of figures; e.g. a theft (where the victim can’t be sure of seeing the theft) is a ‘lost property report’” (NT50).</p> <p>“Green is good and red is bad, we can only have X burglaries today, record it as something else” (NT3182).</p> <p>“At present we are being asked to search more people and make 20% of them positive searches for weapons/drugs/alcohol. The only way that can be done is by selectively recording searches” (NT1923).</p>
<b>SPC Charts</b>	<p><i>No data provided by respondents to indicate there is a significant relationship between SPC charts and perversity or dysfunctional behaviour.</i></p>

To summarise, the examples provided within this theme demonstrate remarkably consistent patterns of behavioural dysfunction and unintended consequences, strongly associated with particular data display formats. These patterns (along with the absence of such when SPC charts are used) occur even where other variables remain constant (e.g. performance information users, or organisational climate); this suggests the data display format itself could well be a major influencing factor affecting behavioural outcomes.

### **4.3.7 Aggregate Dimension: Organisational Climate / Cultural Context**

#### **4.3.7.1 Introduction**

This section explores the wider organisational climate and cultural context, as well as other influences which emerged as themes relevant to the design and use of police performance information. It considers the wider impact of the practices already discussed, along with the influence of senior managers, as well as the attitudes of police officers towards performance information and statistics in general. It is important to note that unlike some of the themes explored within the previous dimensions, the areas discussed herein are neither sequential nor unidirectional; the overall context influences, and is influenced by, micro-level activity.

#### **4.3.7.2 Second Order Theme: Psychological Impact**

This theme considers the psychological effects found to be most strongly associated with the various data display formats, assessing their impact on the prevailing organisational climate, as well upon as the individuals operating within it. With the exception of SPC charts, these effects were found to be overwhelmingly negative. The following discussion identifies implications for officers and the overarching organisational climate, arising as a consequence of the behaviours that tend to ensue as a result of the use of binary comparisons, league tables and numerical targets. Representative quotes can be viewed in Table 4.10, toward the end of this section.

A prominent implication was the “...*destructive impact on morale...*” (LT2060), which occurs when these formats are used. Again, this effect can be traced to misinterpretation of data and consequent belief that apparent variance reflects poor performance, resulting in flawed assumptions and / or inappropriate behavioural responses. For example, respondents stated that use of binary comparisons

“...lowers morale, makes personnel feel they are under pressure...” (BC145), whilst league tables “...damage morale, are divisive, cause friction...” (LT1846) and numerical targets “...can be demoralising, as when you work hard and see poor results it can be disheartening” (NT889).

This suggests there is an association between these formats and low morale, triggered by the assumptions, decisions and actions of managers that occur when they are used. Conversely, whilst there was minimal commentary on any links between the use of SPC charts and psychological effects, one officer suggested they provide, “*Greater context, keeping morale*” (SPC972). Without proposing that the SPC format is a morale *booster*, it certainly appears to limit the deleterious effect on morale found to be strongly associated with the other three formats.

This finding on morale was interlinked with testimony of there being a damaging emotional impact on officers, many of whom reported experiencing “...an atmosphere of fear and concern...” (LT2094), “...a feeling of failure” (BC1043) and “...anxiety...” (BC173), as a result of the assumptions and behaviours that tend to be generated by use of these formats. Respondents consistently described a climate of negativity, blame and pressure, which left officers feeling demotivated and even humiliated as a direct consequence of managers’ faulty engagement with data and subsequent erroneous perceptions about performance.

Again, these experiences were strongly and exclusively tied to the use of binary comparisons, league tables and numerical targets; the common factor being that these formats consistently mislead users into believing there are issues that require remedial action, ultimately causing “...resentment, anxiety, stress” (NT401). Furthermore, the conditions generated seem entirely capable of encouraging perverse or otherwise dysfunctional responses, which in turn could perpetuate psychological damage. In respect of SPC charts, respondents did not report any of the adverse conditions associated with other formats, either in respect of there being a negative psychological impact on individuals, or upon the overall organisational climate.

Table 4.10 shows additional typical statements regarding the associations between certain performance information and psychological effects on officers:

**Table 4.10: Second Order Theme ‘Psychological Impact’ (T8)**

<b>Category</b>	<b>Representative Quotes</b>
<i>Binary Comparisons</i>	<p>“It has a negative impact on morale. Supervisors review the figures and assume that there has been a dip in performance despite the fact that this may be a natural fluctuation in figures due to any number of possible factors, none of which have to do with the performance of the team” (BC2920).</p> <p>“Fear” (BC3833).</p> <p>“Presenting in such a way demotivates and leads to despondency as it gives an indication of poor performance and an expectancy of higher demands on resources to remedy the problem” (BC2959).</p>
<i>League Tables</i>	<p>“It doesn’t help team morale, as however well you are performing someone will always be at the bottom of the table” (LT1256).</p> <p>“It names and shames and demotivates / demoralises” (LT3173).</p> <p>“These tables undermine self-confidence, engender competitiveness and prevent cohesion between teams” (LT2103).</p>
<i>Numerical Targets</i>	<p>“This reduces morale and appears to lay the blame for crime at the feet of the officers” (NT2878).</p> <p>“This is the worst. Arbitrary and unachievable targets are set. Then officers at the ground level are blamed for not achieving them” (NT160).</p> <p>“Negative effect, feeling of failure if targets not met” (NT1897).</p> <p>“It takes away the human aspect of dealing with the public as officers are in fear of under achieving and being disciplined” (NT344).</p>
<i>SPC Charts</i>	<p><b><i>No data provided by respondents to indicate there is a significant association between SPC charts and strong psychological effects.</i></b></p>

As a final point, it is notable that the adverse effects described seem to occur even where managers believe they are acting in the best interests of the service; it is not simply a question of aggressive application, but one of there being a strong tendency to misinterpret certain formats, leading to unwarranted assumptions, poor decisions and misguided reactions. In turn, these assumptions and responses contribute to the overwhelmingly negative effect on officers’ psychological well-being that is observed when these data display formats are used.

#### **4.3.7.3 Second Order Theme: Influence of Senior Managers**

A further theme that emerged within this aggregate dimension was the influence of senior managers on subordinates, as well as the overall climate and cultural context. By nature of their seniority, managers are in a position to shape the norms

surrounding the use of performance information, along with preferences about formats. Coupled with a generally woeful lack of statistical literacy, senior managers' power to exert their will regarding preferred data display formats has led to particular formats either becoming firmly embedded or completely rejected.

There was also a pattern of pressure being cascaded down the ranks and subordinates being held to account based on managers' perceptions about performance. This was found to be a multilevel condition where managers at various ranks replicated and reinforced behaviour. Challenge was relatively rare, and when it did occur, was generally frowned upon and unsuccessful in facilitating change. The overall effect of this combination was a climate where the *status quo* prevails and reliable forms of data display remain marginalised due to personal preferences for simplistic formats. Representative quotes can be viewed in Table 4.11, toward the end of this section.

Respondents reported managers exhibited strong affinity for binary comparisons, league tables and numerical targets (albeit to varying degrees); therefore, these formats assumed primacy in police performance management systems. For instance, despite the known limitations of binary comparisons, the format is ubiquitous as “...it presents the data in a way that management wish it to be seen” (BC3204). This situation persists because “...senior managers prefer the comfort blanket of using the old charts and methods to monitor performance” (FT3319).

The effect is that if a senior manager favours simplistic formats, that is what is used; this even applies when junior officers (or even analysts) suggest reliable formats are utilised instead. For example, one respondent states, “We are trying to eradicate them, but unfortunately supervisors and managers continue to cling to them” (LT3843). Another, who had “...spent years in performance management roles...” attempted to challenge the use of unreliable formats, but became “...very unpopular, especially with certain individuals in very senior positions who preferred to wallow in ignorance” (FT3038).

The influence of senior managers may partly explain why the use of simplistic forms of data display persists in UK policing despite more robust alternatives being readily available; where a user opting for ‘style over substance’ happens to be a senior manager, it is difficult to dissent. Furthermore, where valid concerns about such formats are simply dismissed by bosses, this also seems entirely capable of

contributing towards the types of psychological damage discussed in the previous theme.

In addition to the (perhaps unwitting) impact caused by the issue of preference, respondents also recounted overt and deliberate pressure being exerted downward through the ranks. This arose due to misconceived assumptions about perceived performance deficiencies, leading to “...concerns which are cascaded down for improvement and explanation” (BC2234), as well as intense pressure for “...performance to be enhanced and this is then fed to lower ranking officers through the management sliding scale” (BC2059).

Again, use of certain formats seems to be a primary catalyst for this pressure, arising due to misinterpretation and unfounded assumptions. It is also notable that pressure is often replicated at each level; therefore, unless a relatively junior individual is prepared to challenge upwards, the use of unreliable formats becomes normalised at every level. This behaviour directly contributes towards the types of behavioural dysfunction and psychological damage discussed previously; respondents talk of a “...chain effect of concern and stress from the higher management down to the supervisors of the team” (NT2978), as well as “...pressure on PCs to use unethical / unprofessional practices to meet targets” (LT2397).

Furthermore, as a result of managers’ assumptions, subordinates are directed to carry out tasks and held to account via reporting mechanisms. This perpetuates a culture of command and control, with the requirement to report upon daily activity to senior managers, which “...affects every officer in how they are tasked every day” (NT2609) and where “...teams are brought together to account for their activity and to answer why there has been an increase” (BC3308).

Once again, directives for such activity and associated accountability can be firmly traced to managers’ initial engagement with performance information. Here it can be seen how the influence of senior managers can initiate unnecessary operational changes and even perverse behavioural responses on the part of subordinates. Furthermore, the ‘command and control’ culture of the police service does not appear to lend itself to an environment where even highly unreliable data display formats can be easily challenged or replaced.

This latter point is illustrated by the power wielded by senior managers in respect of sidelining the use of SPC charts. Although this form of data display is generally associated with more beneficial effects than the other three formats, respondents reported that some senior managers were reluctant to engage with SPC; this may well be partly responsible for this format being underused within sections of the police service.

However, respondents did not report any association between SPC charts and the downward pressure or aggressive ‘holding to account’ recorded elsewhere, yet found the format is still “...largely ignored due to reliance on up / down comparisons” (SPC1875). Even in cases where SPC charts are used, “...it is still very much secondary to the traditional methods of presentation” (SPC3802). This demonstrates the dominance of the other three formats in police performance management systems and the reluctance of some senior managers to ‘let go’ of these traditional formats.

Table 4.11 shows additional typical statements regarding the influence of senior management:

**Table 4.11: Second Order Theme ‘Influence of Senior Managers’ (T9)**

<b>Category</b>	<b>Representative Quotes</b>
<i>Binary Comparisons</i>	<p>“It is a meaningless binary comparison. This is constantly pointed out but still the same figures are produced” (BC3543).</p> <p>“Shows an increase which puts pressure on senior management who then put pressure on lower ranks” (BC1109).</p> <p>“SLT [<i>Senior Leadership Team</i>] task Inspectors to task Sgts to improve performance” (BC2362).</p>
<i>League Tables</i>	<p>“These are used and again fed through the rank process from the higher rank down. This then places pressure on supervisors for their team to improve their figures” (LT2059).</p> <p>“Team D might as well pick up their p45’s as the bosses will crucify them” (LT1110).<sup>19</sup></p> <p>“Bottom teams literally bullied or as it’s called, ‘held to account’ - leading to performance manipulation” (LT3131).</p>
<i>Numerical Targets</i>	<p>“Over the recent months we have tried to move away from targets, but senior managers find it difficult because this is how they are used to measuring performance” (NT3590).</p> <p>“Again, this just causes senior managers to cascade pressure down the ranks causing those at the bottom to be put under intense scrutiny” (NT1111).</p> <p>“When target rate is not being met, Command officers will seek explanations from Area Commanders who in turn seek explanations from the lower ranks” (NT2237).</p>
<i>SPC Charts</i>	<p>“It is used but it is used alongside the traditional year on year or month on month comparison data. This means any benefit to using the above approach is lost as more attention is paid to the binary comparison” (SPC3797).</p> <p>“Senior staff don’t like the upper and lower boundaries” (SPC3744).</p> <p>“I was first taught a version of SPC in 2006 and tried to introduce it into my force to prevent the demand for a knee jerk reaction to a variation in crime occurrence and outcomes... needless to say whilst I used to defend my position in performance meetings as a district commander there was no interest in it” (FT3657).</p>

Overall, the influence of senior managers is found to be a significant factor in perpetuating the use of unreliable formats and a major obstacle to embedding the use of more robust forms of data display; this is despite the recurrent and well-documented issues associated with simplistic formats. Analysis of the free text data also exposes a direct link between the use of certain formats and behavioural

<sup>19</sup> Note: This is a reference to the specific experimental stimulus within the survey instrument.

dysfunction, on account of managers formulating erroneous assumptions then engaging in conduct likely to act as antecedents to dysfunction, such as exerting pressure upon subordinates, or even outright bullying.

#### **4.3.7.4 Second Order Theme: Attitudes towards Performance Information**

The final theme in this section explores police attitudes towards performance information. The free text entries revealed a perceived split between senior and frontline officers, as well as general disdain towards performance information. There was also extensive commentary on the difficulties associated with defining and quantifying police 'performance', as well as evidence of a strong vocational sentiment amongst officers. Finally, although some respondents were dismissive of performance measurement *per se* (and statistics in general), others acknowledged the need for accurate and meaningful performance information. Representative quotes can be viewed in Table 4.12, toward the end of this section.

There was a strong perception that senior officers are most concerned with performance information, whilst frontline officers are often disengaged. This split led to claims “...*senior managers care about figures and the lower ranks care about the people*” (FT1472), causing “...*division between management and the workers*” (NT143). This perception was coupled with a sense of general ambivalence towards performance information of any type. For example, one respondent suggested “...*people are sick up to the back teeth with pie charts, spread sheets and graphs*” (FT2307), whilst another quipped, “*99% of statistics are pointless*” (NT392).

These viewpoints reflect a stance whereby performance information seems to be perceived as a homogenous entity regardless of its format; some respondents were equally derisory towards robust forms of data display as they were in respect of simplistic formats. Commentary on SPC charts was particularly limited and quite polemic; respondents tended to either extol their benefits, or include them in generic criticism of performance information.

Indeed, there was a view that not only was performance information unimportant, but that managers' constant focus on it was an unwelcome hindrance to undertaking the functions of policing. However, it transpired that such disdain was often coupled with a strong vocational sentiment. Comments such as, “*Most officers just want to*

*catch and convict criminals and do their best for victims. Figures are almost irrelevant” (BC3275) and “...it is not my job to be concerned about statistics or figures. My job is to be out on the frontline trying to make a difference and serve the community” (BC766) were typical.*

Such passion for wanting to help people and do a good job was evident in many comments bemoaning the apparent irrelevance of performance data. This may provide a little context to some of the generic dislike of statistics; it appears that some officers do not make the connection between the effective use of meaningful performance information formats and performance itself.

Others emphasised difficulties in defining or quantifying police performance and in particular the suitability of using crime data as a performance measure. This echoes discussions in the literature (see Bayley, 1994; Home Office, 1990; 1997; Coleman and Moynihan, 1996) and may also partly explain the general disdain towards performance information evident in responses. As stated by one respondent, “*We provide help and comfort to people in need at a time in their life when they really need help and that is immeasurable” (SPC1737). Others observed “...many areas of policing do not count towards figures” (LT1620), highlighting the consequent disposition towards “...easily measurable aspects of police work” (LT1823).*

In respect of the suitability of using crime figures as a performance indicator, respondents insisted “*...police do not necessarily control the level of crime, but respond to it” (BC556), citing the unpredictability of crime rates and the multiple external variables affecting them. Some observed that proactivity can even result in an increase in crime, “...for instance, when officers search suspects and find knives / drugs etc” (NT658). Consequently, whilst some senior managers ascribe performance success or failure to apparent changes in crime rates, many frontline officers tended to disregard them as a valid measure of performance. One respondent also highlighted the impotence of crime reduction targets on crime reduction:*

*“Targets have no effect on whether a burglar will go offending tonight or not” (NT397).*

However, despite many respondents dismissing the relevance of crime data (or performance information generally), others offered nuanced observations regarding

the benefits of effective performance measurement, as well as the importance of thoughtful data presentation and use. For example, one officer asserts “...*crime figures have their place but we need to be a bit smarter about how they are presented*” (FT1516), whilst another suggests “...*statistics can be very helpful in the job we do as long as they are presented in a clear and understandable way*” (FT494).

Others acknowledged “...*performance needs to be measured as the public have a right for the police to be held accountable...*” (FT3611) and “...*the use of statistics and performance figures in policing can be really useful in predicting demand*” (FT3234). Overall, this suggests an appreciation for the use of well-presented data in understanding performance and informing decision-making, as well as for promoting legitimacy. Taken in conjunction with respondents’ views on crime data, this indicates an acceptance that whilst crime figures may not be a performance indicator *per se*, they can still be a useful source of information, which, if presented in meaningful formats, could in fact be extremely useful for decision makers.

The crux of these observations is that many respondents appreciate the need for performance measurement and analysis, but that relevant measures and reliable data display formats should be used. Some even suggested this could help officers focus on the job at hand and prevent them from reacting to perverse incentives:

*“Performance is about doing the right thing in the right way at the right time to achieve an appropriate result, not to come first in a league table”*  
(FT2666).

Table 4.12 displays additional representative statements regarding general attitudes towards performance information:

**Table 4.12: Second Order Theme ‘Attitudes towards Performance Information (T10)**

<b>Category</b>	<b>Representative Quotes</b>
<i>Binary Comparisons</i>	<p>“Meaningless at ground level - statistics satisfy the ‘bean counters’” (BC1519).</p> <p>“Too many variables affect crime figures - officers on the ground know this and pay little if no attention to said figures” (BC1102).</p> <p>“Police work cannot be measured in so many areas - how do you measure helping an old lady into bed after she has taken a fall? Yet this type of help is so valuable and fundamental to the image of the great British bobby, it makes us human and not just a uniform, it is priceless but immeasurable” (BC2312).</p>
<i>League Tables</i>	<p>“We are the police. We should not be driven by figures or statistics” (LT331).</p> <p>“This is really only of concern to officers in management positions. It does not affect how officers in a front line role deal with what they are presented with on a day to day basis” (LT826).</p> <p>“Complete waste of time and an extremely poor management tool. The challenge is to encourage performance by doing the job that is expected of you!” (LT2839).</p> <p>“We all come to work to do a good job and work hard - not think about performance indicators” (LT2024).</p>
<i>Numerical Targets</i>	<p>“Officers have no interest in statistics” (NT2413).</p> <p>“To us they are irrelevant and a bane of our working life, we turn up, do the best we can, catch criminals and protect the public” (NT1876).</p> <p>“Crime by its very nature is unpredictable. Some crimes are easier to detect than others” (NT1922).</p> <p>“If I have evidence, I will charge someone and the crime will be detected. I can’t make evidence appear out of thin air. If someone phones at 3am and says their window is smashed, there is no CCTV, there is nothing learned from house to house, nobody is traced in the vicinity and there is no opportunity for forensic evidence etc... then I'm not going to detect the crime, am I?” (NT382).</p> <p>“I think it is useful to measure performance generally but setting arbitrary targets for individual and team performance is damaging and does not reflect the nature of the work we do” (NT870).</p> <p>“Important to have knowledge of how well we are performing but completely wrong to create targets” (NT2549).</p>
<i>SPC Charts</i>	<p>“I’m not interested in this kind of thing. I just like locking up bad people who need to be taken off the street” (SPC1436).</p> <p>“A complete switch off and little use to operational staff. I have seen data displayed in this manner and have no idea how it is to be interpreted” (SPC2698).</p> <p>“I don’t care. I just want to get home in one piece” (SPC2119).</p>

In summary, although there is a degree of negativity towards statistics in general, officers tend to exhibit a strong desire to do a good job, but are irked by managers' apparent obsession with, and misuse of, performance information. Reservations are also held about the types of formats commonly used and whether crime data should be used as a measure of performance at all. Some respondents, however, exhibited an understanding of the benefits of effective performance measurement, which was reflected in their observations.

#### **4.4 Theoretical Model**

As a result of using data structures to categorise the free text entries, notable relationships and potential causal influences are identified amidst the concepts, themes and dimensions. In particular, whilst acknowledging there are a multitude of factors capable of affecting behaviour, it becomes apparent there is a relationship between data display and behavioural dysfunction. The data structures also expose intermediate behaviours displayed by managers that can act as antecedents to dysfunction, such as demanding improvements for perceived deficiencies and exerting pressure on subordinates.

The conceptual themes generated within the data structures were therefore used to construct an empirically-grounded theoretical *process model* (see Besharov, 2014), which explicates these links and highlights influencing factors that appear to contribute towards the phenomenon of interest. Crucially, the components of the model can be traced via the data structures directly to the data produced by the survey instrument, thereby assuring solid provenance.

##### **4.4.1 Theory Generation and Objectives of the Model**

In constructing the theoretical model, due regard was paid to established principles guiding the generation of strong theory (see, for example, Bacharach, 1989). Such theory may be defined as "...a group of logically organized laws or relationships that constitutes explanation in a discipline" (Heinen, 1985, p.414); it should also identify underlying mechanisms and provide an explanatory account for the phenomenon of interest (Foss, 2010; 2014).

Furthermore, strong theory is more than merely descriptive; it should possess explanatory (and predictive) utility, thereby enabling observers to understand the

features of a phenomenon (Cook and Campbell, 1979; Kerlinger and Lee, 2000; Gelso, 2006; Udo-Akang, 2012). It should be also characterised by methodological rigour and be practically useful (Hodgkinson and Rousseau, 2009; Hodgkinson and Starkey, 2011). Ultimately, it should provide answers to the *why* question (Kaplan, 1964; Merton, 1967; Sutton and Staw, 1995). Fundamentally:

“The overarching goal of theoretical research...is to develop a set of reliable claims that identify valid causal relationships between a phenomenon and its precedents” (Oxley *et al*, 2010, p.379).

Therefore, the theoretical model aims to explicate such relationships in a systematic and rigorous manner, compatible with the critical realist interpretation of causation. To do so, it organises the 10 second order themes into three main areas; firstly, it depicts those that act as antecedents to the choice of preferred data display format and exert influence over its use. Next, it displays those themes most pertinent to the engagement phase, where the interaction between performance information and user occurs, leading to the activation of a proposed mechanism. This interaction then triggers a sequence of events that leads to commonly-observed effects, including the phenomenon of interest - behavioural dysfunction.

Themes within the model are tagged with reference markers (‘T1’ to ‘T10’) that correspond to the relevant data tables and associated narratives from the previous section, thereby directing the reader to supporting data for these themes and relationships. Some themes are treated as constructs (e.g. ‘Statistical Literacy’), whilst others are presented as potential influences of or upon constructs (e.g. ‘Catalyst for Assumptions’). As the model is derived from both the primary data structure and its supplementary ‘SPC charts’ counterpart, where second order themes were absent from the latter data structure, they have been ‘greyed out’, indicating the use of this particular format is not strongly associated with these themes.

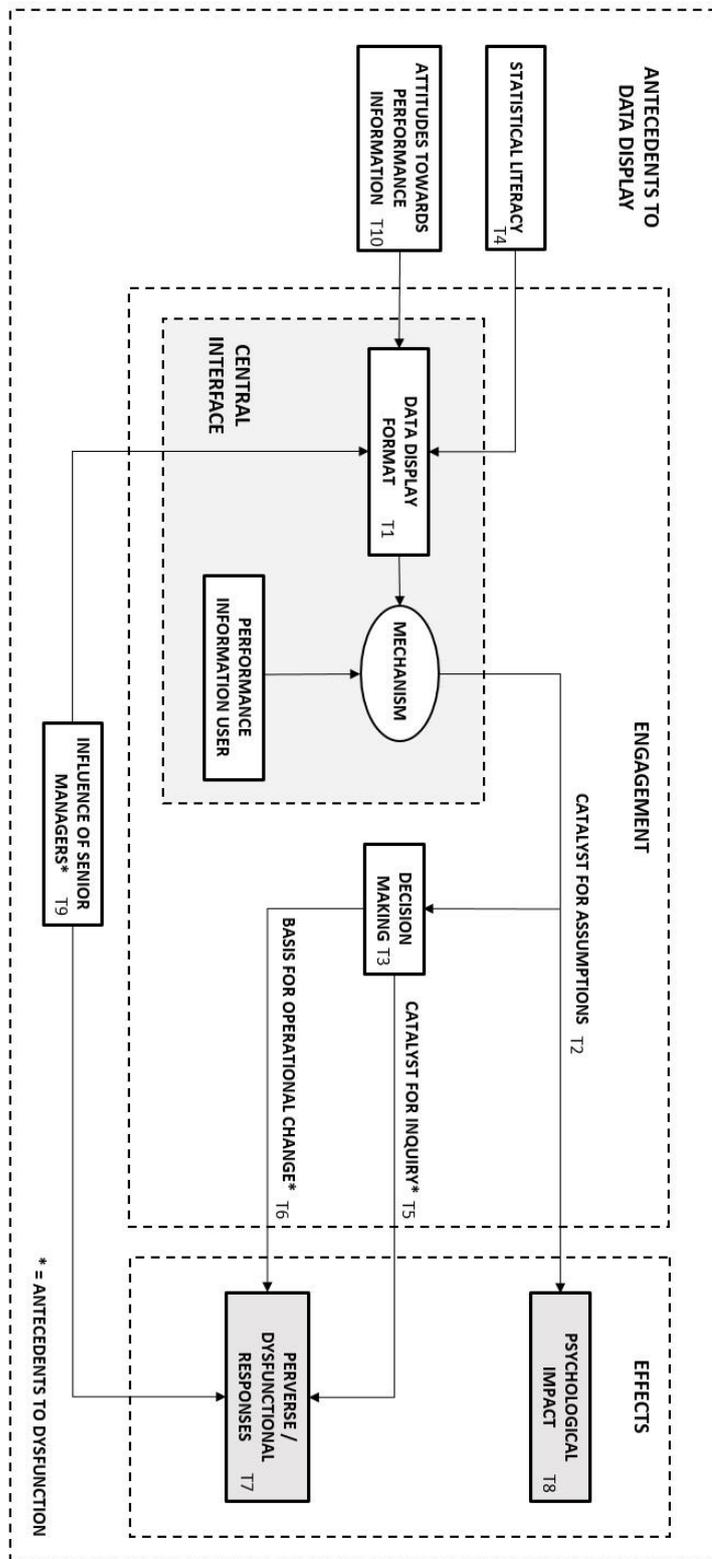
The primary objective of the model is to offer a simple visual illustration of relationships between the identified themes within the data structure. It is important to emphasise this model is drawn exclusively from themes that emerged from the data set, so may well omit other factors that were not the focus of this study; therefore it is not intended to be an exhaustive explication of all variables capable of affecting behavioural dysfunction, but a straightforward graphical representation of

relationships between the interlinked themes identified during analysis. Furthermore, the layout of the themes is designed to assist ease of navigation, rather than signify any kind of hierarchy amongst them.

It is also necessary to state that although the arrows depict the most likely direction of causal influence, there is no attempt to ascribe precise weightings or argue there may not be a degree of reciprocal effect (or overlap) between them. Similarly, it is accepted there may be relationships between themes that are not explicitly depicted within the model (e.g. it is possible that ‘Statistical Literacy’ may have an effect on ‘Attitudes to Performance Information’, or *vice versa*). The reason any such potential associations have been omitted is to focus on the relationships that emerged most strongly from the data and produce a parsimonious model that provides a provisional explanation for the phenomenon of interest.

The model is presented in Figure 4.3, followed by commentary on its operation. This narrative explores relationships between second order themes, as well as their effects on thought processes and behaviour; specifically, it acts as a guide for the sequence of events depicted in the model, examining how the use of particular data display formats seems to either exacerbate or moderate behavioural dysfunction. Finally, it provisionally postulates the existence of a causal mechanism, providing the basis for an explanatory account that explicates the relationship between data display and behavioural dysfunction.

Figure 4.3: Theoretical Process Model – Influence of Data Display on Behavioural Dysfunction<sup>20</sup>



<sup>20</sup> Note: The proposed mechanism is depicted as an oval in the diagram to differentiate it from the data-derived themes.

#### **4.4.2 Commentary and Explication**

As data display is central to this study, the narrative first examines how different formats affect engagement with performance information, before tracing the pathway triggered by activation of the proposed mechanism to the phenomenon of interest. It also addresses those themes that influence the design and use of particular formats ('Antecedents to Data Display'), as well as the effects of their use. Interim conclusions are located after this account.

Although the model incorporates all themes from the three aggregate dimensions, its structure is not arranged by them. The focal point is labelled 'Central Interface', which is situated within the 'Engagement' domain; it is here that performance information and users combine via the proposed mechanism to produce a series of events. Fundamentally, it is posited that the chosen data display format (T1) affects the operation of this mechanism; this prospect will be extensively tested in the forthcoming experimental component of this study.

Essentially, it appears that during the engagement between performance information and users, this interaction consistently acts as a catalyst for assumptions (T2), accurate or otherwise; in the case of binary comparisons, league tables and numerical targets, as such assumptions are affected by the unreliable nature of these formats, they are often unwarranted and lead to unnecessary concern about perceived problems, which may not exist. This is the very crux of the issue of data display; the chosen format appears to directly affect the nature of engagement and thereafter influence individuals' choices and behaviour.

Consequently, whilst reference-dependent formats may appear user-friendly and visually pleasing, they typically reduce complex data sources to simplistic representations of 'improvement' or 'deterioration', 'good' or 'bad' etc; this drives assumptions about 'success' or 'failure' in equally simplistic terms and where data are taken at face value, users may envisage changes, trends or trajectories which may be non-significant or even non-existent. Thus, negative variance from the reference point leads to explicit or implicit urges to improve, generating pressure to rectify perceived deficiencies; remedial action is then instigated, whether or not this is necessary. Conversely, positive variance can lead to issues of genuine concern within data sets being obscured by these formats and therefore overlooked.

Whilst binary comparisons trigger an aspiration to attain or exceed the temporal reference point, it appears that league tables and numerical targets each possess even greater power to instigate behavioural changes. League tables produce strong incentives to climb to higher positions and avoid being considered inferior to peers; this occurs despite the unstable nature of such rudimentary ranking methods. Similarly, the targets used within respondents' organisations provide a precise aspirational 'pass' / 'fail' reference point, which if not attained, causes feelings of failure and induces pressure to address perceived deficiencies.

Conversely, where SPC charts are used to present data, users tend to interpret the content accurately and react in a different way to how they do when engaging with reference-dependent formats; the natural consequence of presenting data in this way is that users formulate fairer assumptions about performance and thereby concentrate efforts towards those areas which truly warrant intervention. SPC charts also clearly inform users when data are statistically unremarkable, thereby reassuring them that tactical changes or remedial activities are not required.

However, there is a caveat relating to the use of more complex formats; although they may be more reliable, users require a degree of pre-existing statistical literacy (T4) to engage effectively; otherwise, there is a risk of misinterpreting even these formats and activating a similar chain of events to that triggered by unreliable designs. There is also the wider issue of choosing formats; decision makers must possess the statistical wherewithal to recognise and select robust designs, especially as many seem predisposed to reject certain formats on account of them appearing complicated or visually displeasing. For this reason, statistical literacy is positioned within the model as one of the major antecedents of data display.

Statistical literacy also looms large as a significant factor in enabling users to distinguish between reliable and unreliable data display formats; it can prevent misconceptions arising following engagement with superficial formats, as well as mitigate misinterpretation of robust formats. Ultimately, statistical literacy appears to be a primary condition capable of limiting behavioural dysfunction, as the adverse chain of events is halted where users are statistically aware and therefore reject unreliable formats from the outset.

However, where simplistic reference-dependent forms of data display are used, the ‘thinking’ stage is followed by an ‘action’ stage, whereby users consistently misinterpret data, formulate unwarranted assumptions and enact particular types of adverse tendencies; furthermore, these occurrences are strongly aligned to specific data display formats. At the ‘thin end’ of this behavioural ‘wedge’, users commission further research into perceived issues to ascertain whether action is required; however, as the initial stimulus was unreliable, this further analysis represents unnecessary activity, which absorbs capacity from within the system. Therefore, whilst not at the more extreme end of perversity, such unwarranted reactions may be considered intrinsically dysfunctional.

Alternatively (or additionally), users demand explanations and / or improvement for perceived failings (T5), based on the assumptions triggered by their interpretation of the data. Where the data display format is unreliable, this too may be classed as an inappropriate and potentially damaging response, as subordinates scramble to provide explanations for issues which may not exist. Indeed, if there is no significant issue present, a valid explanation is unlikely to be found; much less a basis to hypothesise about potential causes or solutions. In contrast, robust formats encourage focused inquiry where there is a valid basis for such, as well as promote consideration about reasons for identified patterns within the data.

Furthermore, managers often initiate a variety of operational tactics intended to combat potentially non-existent issues (T6), often at the expense of valid matters of concern, which remain unseen and therefore unaddressed. Although likely well-intentioned, these actions exhaust finite resources and absorb capacity unnecessarily; subordinates are particularly frustrated at what they perceive to be ‘knee-jerk’ reactions. Such responses, along with managers’ demands for improvement and mistaken assumptions, are consistently identified as being antecedents to dysfunction.<sup>21</sup> Conversely, the use of SPC charts tends to promote targeted operational responses, as well as a reduction in ‘knee-jerk’ reactions.

In addition to unnecessary operational deployments, simplistic formats were most strongly associated with more extreme forms of behavioural dysfunction and perversity (T7), even to the extent of unethical or unlawful acts. This closely mirrors

---

<sup>21</sup> These ‘Antecedents to Dysfunction’ are clearly marked with an asterix on the model.

instances recorded within the literature, including deliberate falsification of figures, unlawful searches, gaming and other acts designed to enhance standing in league tables or attain numerical targets. Conversely, there was nothing within the free text data to suggest a link between SPC charts and outright perversity; indeed, the commentary suggests their use actively mitigates such conduct by halting the described sequence of events. Therefore, the 'Perverse / Dysfunctional Responses' theme within the model is greyed out for the SPC category.

This theme completes the path from the 'Data Display Format' theme to the phenomenon of interest and although there seems little doubt that improper application of any type of performance information could instigate dysfunction, the steps between these two themes trace a consistent path whereby engagement with particular formats tends to produce such behaviour; this occurs even where performance information users act with good intentions. It is also noticeable that the misguided assumptions formulated by managers act as a direct catalyst for those behaviours that may be considered as antecedents to dysfunction.

Indeed, the influence (and behaviour) of senior managers was identified as a major antecedent (T9), both in respect of affecting the choice of formats and as an intermediate condition responsible for triggering broader dysfunction. As senior managers determine which data display formats are preferred, this can either have positive or negative implications for performance information use, depending on their selection. Furthermore, where subordinates express reservations about unreliable formats, these concerns are often dismissed by managers.

A further particularly strong theme was adverse psychological impact (T8) on officers when binary comparisons, league tables or numerical targets are used. This most routinely takes the form of low morale, as well as feelings of worthlessness and failure that arise as a consequence of senior managers' flawed assumptions and questionable decision-making (T3); perceived failure also tends to be emphasised and is often accompanied by explicit or implicit personal criticism and sanctions. Such conditions lead to a negatively-charged organisational climate, characterised by concern and stress.

Additionally, for those charged with carrying out tasks in response to managers' interpretation of data, this toxic climate is exacerbated by constant downward

pressure to explain and address perceived deficiencies, based solely on managers' unfounded assumptions, which have ultimately been caused by engagement with faulty data display. Such pressure, along with the managers' directives to conduct unnecessary operational activities, causes psychological damage and is identified as a primary antecedent to behavioural dysfunction.

In respect of SPC charts however, the free text data did not find any link to the types of adverse psychological effects associated with the other three formats. Indeed, the limited narrative in this area was of opposing polarity to the other categories, even highlighting a small beneficial effect on psychological wellbeing. This suggests that the use of more robust forms of data display has advantageous implications for the organisational climate, as they do not tend to be associated with the types of adverse psychological effects caused by the use of simplistic formats. For these reasons, the 'Psychological Impact' theme within the model is greyed out for the SPC category.

With regard to attitudes towards performance information (T10), it was found that officers exhibited a strong sense of vocation, although many displayed negativity towards statistics in general. Furthermore, perceptions of a split between senior and frontline officers were rife, with respondents reporting that managers' reactions to data are a constant cause of disaffection. None of this is conducive to an environment where statistical literacy and the use of reliable data display formats are encouraged; it perpetuates a state of disinterest (and even hostility) towards performance information and prolongs the use of unreliable formats, due to officers being unable or unwilling to challenge their use. Such attitudes combine to constitute one of the main antecedents affecting the choice and use of data display formats.

#### **4.4.3 Critical Realist Perspectives**

From a critical realist viewpoint, the second order themes depicted in the model may be considered as *entities*; of particular relevance are 'Data Display Format' and 'Performance Information User', as it seems highly likely that there is indeed a mechanism operating at the nexus between these two entities. It also appears that the relationship between the two is *necessary and contingent* for the proposed mechanism to be activated. Each possesses internal structures, as well as powers and liabilities to act; however if no interaction occurs between them, the mechanism

remains dormant and the effects described towards the right hand side of the model are not produced.

Activation of this mechanism appears to trigger cognitive processes whereby data display directly influences how users interpret performance data, leading to assumptions about performance and thereby affecting decision-making. In isolation, performance information does not possess an innate power to act (unlike its users), however when engagement occurs, the proposed mechanism produces a series of *events*, ultimately leading towards or away from behavioural dysfunction, depending on the data display format used.

In respect of the overarching organisational climate and cultural context, it is important to recognise that this encompasses all components of the model, yet can also be affected by them. The notion of the overarching context exerting influence on the behaviour of actors, whilst acknowledging their agency and ability to influence the structure within which they operate, is reflective of the notion of *Structure and Agency*; influence is not unidirectional, as each condition can act upon the other, affecting behaviour at both the macro and micro levels.

The overriding context therefore creates conditions that influence actors and affect the type of data display used; in turn, the chosen format appears to impact upon the operation of the mechanism, which then contributes to dysfunction. Consequently, although performance information users possess utility, data display appears to be a key factor in producing observed effects – this prospect will be further tested in the forthcoming experimental component of the study, which attempts to isolate data display from other conditions likely to exert influence on the phenomenon of interest, in order to assess its innate powers and liabilities.

#### **4.5 Summary and Interim Conclusions**

In summary therefore, the analysis highlights recurrent patterns and themes which suggest there is a substantial relationship between data display and behavioural dysfunction; the theoretical model depicts the interplay between these themes, demonstrating how data display acts as a pivotal factor affecting the operation of a postulated mechanism situated at the interface between performance information and users. If the data display format is unreliable, it is therefore here that it impairs this

interaction and causes problems downstream; furthermore, it seems that a primary condition capable of preventing this sequence from being triggered is pre-existing statistical literacy.

The analysis therefore indicates the use of simplistic reference-dependent forms of data display could well be a powerful ‘upstream’ catalyst for dysfunction. It also finds that the use of statistically robust formats, such as SPC charts, appears to directly affect the polarity of the mechanism’s operation, leading to more preferable outcomes. Furthermore, the analysis confirms the types of behavioural dysfunction and unintended consequences recounted by respondents closely matches that reported within the literature.

Overall, the model provides a platform from where to commence deeper exploration of the operation of this postulated mechanism. The interim findings to this point certainly indicate that data display seems capable of directly contributing toward behavioural dysfunction, as well as producing intermediate behaviours on the part of managers that act as antecedents to dysfunction. These phenomena can each be traced to where performance information users interact with data display, causing the mechanism to produce effects; this mechanics of this interaction and its impact on the phenomenon of interest will be further tested in Chapter Five.

## Chapter Five

### Quantitative Analysis and Findings

#### 5.1 Introduction

This chapter discusses the results of a series of psychometric micro-experiments designed to assess respondents' interpretation of, and reactions to, depictions of hypothetical numeric police performance information. These are presented using a range of data display formats (i.e. binary comparisons, league tables, numerical targets, SPC charts and contextualised peer comparison charts); each of the visual stimuli depict comparable data sources (e.g. crime rates) and are directly representative of how such data are presented in UK police forces.<sup>22</sup>

The narrative will discuss each question in turn, comparing and contrasting data produced by the tests, in order to highlight any notable patterns or tendencies. Primarily, this quantitative analysis aims to test the model and explore the properties of the postulated mechanism that seems to operate at the nexus between performance information and users. By ascertaining whether the output of the micro-experiments reflects themes and concepts identified during the qualitative analysis, the tests aim to establish if there is a significant relationship between particular data display formats and the likelihood, nature, or extent of behavioural dysfunction.

It is acknowledged that responses to the tests could be affected by external factors, such as the latent influences of operational norms in respondents' forces, or their levels of statistical literacy. For these reasons, the results of the experiments were cross-checked against variables configured to assess potential influencing factors, in order to isolate the direct effects of data display as much as possible. Ultimately, the results of these supplementary tests indicate that such conditions exerted minimal or no influence on responses, thereby enhancing the validity of the results. Details of these additional tests are reported at section 5.6.2 and thereafter.

---

<sup>22</sup> As stated previously, the use of contextual peer comparison charts is comparatively rare in UK police forces; however, their inclusion in the experiments is useful for assessing how respondents react in comparison to the other, more widely-used, data display formats.

The provenance to the variables tested derives from the literature review and is outlined in Chapter Three, however they also reflect recurrent behaviours identified during thematic analysis; for instance, respondents reported a tendency for managers to demand explanations for perceived deficiencies, as well as initiate disproportionate or unwarranted operational responses.<sup>23</sup> This strengthens validity and provides additional assurance that the micro-experiments are properly focused towards assessing the presence and operation of the proposed mechanism.

At the end of this chapter, an updated version of the theoretical process model will be presented, highlighting those areas where the quantitative analysis further assesses the themes and relationships identified during Chapter Four. This will be accompanied by a summary of the test results and a brief discussion about their implications.

## **5.2 Question 1: Interpretation of the Stimulus**

The first question in each thematic block of micro-experiments assesses respondents' interpretation of the data contained within respective visual stimuli. It seeks to establish whether users perceive trajectories or noteworthy differences between groups, as well as ascertain their perceptions about performance. Subsequent questions assess respondents' levels of concern arising as a result of their initial assumptions, as well as how likely they would be to engage in various behavioural responses.

### **5.2.1 Binary Comparisons**

The first thematic block relates to binary comparisons. As each block features an identical analytical approach, the content pertaining to the binary comparison stimulus involves enhanced commentary; results from subsequent blocks are reported in a more condensed fashion. After the statistics are reported for each test, there is a brief interpretation of those particular results in relative isolation; a more comprehensive summary of each section's findings is presented at appropriate junctures thereafter.

---

<sup>23</sup> In the author's professional experience, the behaviours described in the variables are also consistent with reactions observed in UK police forces when certain formats are used.

As previously discussed, binary comparisons involve the use of temporal reference points to depict apparent trends or trajectories and remain one of the most common forms of data display used within UK policing. The stimulus used for the binary comparison tests is reproduced below.

*Figure 5.1: Binary comparison stimulus*

Crime Figures			
Last month	This month	Difference	Percentage difference
1,598	1,745	147	8.4%

From the author’s professional experience, it is commonplace for inferences to be drawn from the same type and level of information as that depicted in the stimulus. In this case, the most obvious assumption might be that crime is increasing (even though the difference may not be significant or part of a trend); therefore, the analysis will explore whether respondents do indeed tend to draw this conclusion, and if so, how they react as a consequence. Key questions are therefore:

1. Will respondents tend to assume there is a trajectory and therefore select the ‘Crime is increasing’ option?
2. Will the group of respondents who select ‘Crime is increasing’ tend to enter higher scores in respect of their level of concern?
3. Will this group of respondents tend to enact disproportionate / unwarranted behavioural responses?<sup>24</sup>

Question 1, associated response options and frequencies are presented in Table 5.1.

*Table 5.1: Binary comparison stimulus (Question 1)*

Q1: “In respect of the crime rate, which of the following does the table appear to indicate?” Crime is:		
Response	Frequency	Percent
Increasing	4,806	89.6%
Decreasing	41	0.8%
Stable	189	3.5%
Don’t know	330	6.1%
<b>TOTAL</b>	<b>5,366</b>	<b>100.0%</b>

<sup>24</sup> The questions have been grouped together here in order to demonstrate their sequential nature; however each one is discussed in turn by stimuli type before moving to the next.

The results show 89.6% of respondents (4,806 of 5,366) interpreted the stimulus to mean crime is increasing. A ‘goodness-of-fit’ Chi-Square test produced output of 11960.827 ( $p = <0.001$ ), indicating that this pattern of responses is significantly different from what could be expected through chance alone. The appropriate effect size estimate, Cohen’s  $w$  (see Cohen, 1988; Tapanes, 2008) was  $w = 1.49$ , which is categorised as a very large effect.<sup>25</sup> Overall therefore, the analysis identifies a strong tendency for respondents to misinterpret the stimulus in the expected direction.

### 5.2.2 League Tables

The second experimental stimulus exposed to respondents was a league table. Although this format purports to depict relative performance between peers by using ranks as reference points, its construction is vulnerable to limitations, meaning it can be unsafe to draw conclusions about relative performance. Nevertheless, league tables are also a common form of data display used to present police performance information. The stimulus used in the tests is presented below.

*Figure 5.2: League table stimulus*

Position	Team
1	Team 'C'
2	Team 'A'
3	Team 'E'
4	Team 'B'
5	Team 'D'

The most obvious interpretation of the stimulus might be that Team ‘D’ is performing poorly; therefore, the analysis will explore whether respondents do indeed tend to draw this conclusion, and if so, how they react as a consequence. Key questions are:

1. Will respondents tend to assume Team D’s position in the league table equates to poor performance and therefore select the ‘Performance is poor’ option?

---

<sup>25</sup> The Chi-Square statistic and effect size are included within the text here in order to provide an explanation of their function; thereafter, they will be incorporated into data tables.

2. Will the group of respondents who select ‘Performance is poor’ tend to enter higher scores in respect of their level of concern?
3. Will this group of respondents tend to enact disproportionate / unwarranted behavioural responses?

Question 1, associated response options and frequencies are presented in Table 5.2.

*Table 5.2: League table stimulus (Question 1)*

Q1: “In respect of Team ‘D’s position in the table, which of the following appears to most accurately describe the team’s performance?” Performance is:		
Response	Frequency	Percent
Good	41	0.8%
Poor	2,423	45.1%
Acceptable	281	5.2%
Don’t know	2,629	48.9%
<i>TOTAL</i>	<i>5,374</i>	<i>100.0%</i>

A Chi-Square test produced output of 11960.827 ( $p = <0.001$ ,  $w = 0.88$ ), indicating a large effect. Overall, the analysis identifies two strong tendencies; respondents either opted for ‘Don’t know’ or misinterpreted the stimulus in the expected direction.

### 5.2.3 Numerical Targets

Earlier sections of this thesis discussed the concept of numerical targets acting as aspirational reference points and how variance from such reference points might influence assumptions about performance. Therefore, respondents were presented with such a target alongside the current performance level and asked to provide an interpretation, in order to ascertain whether this data display format does indeed tend to frame assumptions. The stimulus used in the tests is presented below.

*Figure 5.3: Numerical target stimulus*

Detected Crimes	
Target	Performance
25%	11.3%

In this case, the most obvious interpretation of the stimulus might be that performance is poor; therefore, the analysis will explore whether respondents do

indeed tend to draw this conclusion, and if so, how they react as a consequence. Key questions are:

1. Will respondents tend to assume that variance from the target equates to poor performance and therefore select the ‘Performance is poor’ option?
2. Will the group of respondents who select ‘Performance is poor’ tend to enter higher scores in respect of their level of concern?
3. Will this group of respondents tend to enact disproportionate / unwarranted behavioural responses?

Question 1, associated response options and frequencies are presented in Table 5.3:

*Table 5.3: Numerical target stimulus (Question1)*

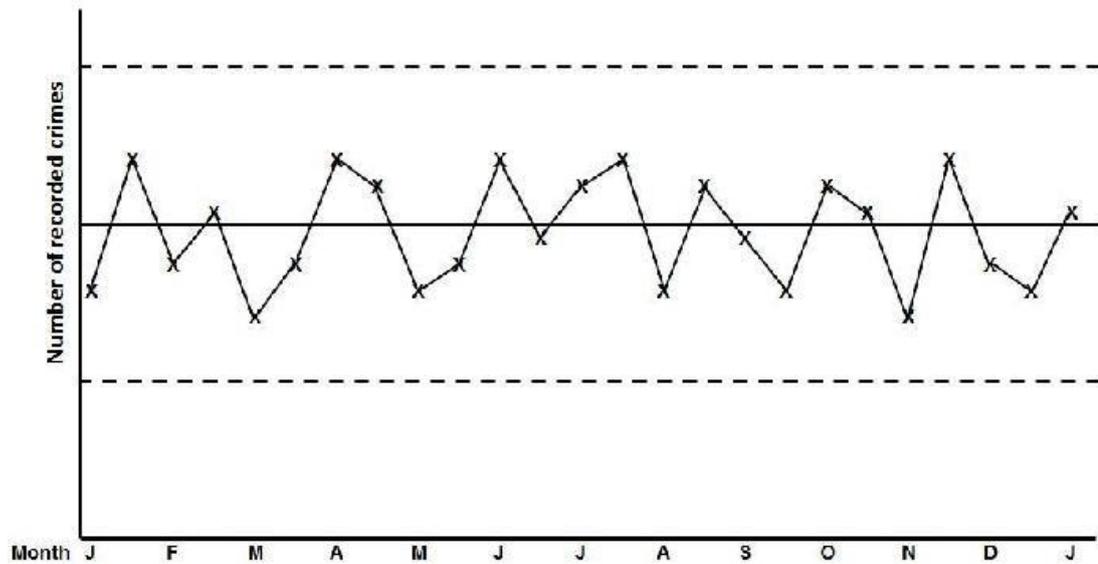
Q1: “In respect of performance against the target, which of the following statements appears to be most accurate?” Performance is:		
Response	Frequency	Percent
Good	90	1.7%
Poor	4,208	78.4%
Acceptable	515	9.6%
Don't know	557	10.4%
<i>TOTAL</i>	<i>5,370</i>	<i>100.0%</i>

A Chi-Square test produced output of 8254.460 ( $p = <0.001$ ,  $w = 1.24$ ), indicating a very large effect; therefore, the analysis identifies a strong tendency for respondents to misinterpret the stimulus in the anticipated direction.

#### **5.2.4 SPC Charts**

It was previously suggested that the use of SPC methodology may help promote accurate interpretation of numeric performance information; consequently, the analysis examines whether the use of SPC does indeed lead to greater accuracy of interpretation. The stimulus used in the tests is presented below.

Figure 5.4: Statistical process control chart stimulus<sup>26</sup>



Although the stimulus is visually more complex than the previous data display formats, it is speculated the most obvious interpretation might be that crime is stable (even for respondents unfamiliar with SPC); therefore, the analysis will explore whether respondents do indeed tend to draw this conclusion, and if so, how they react as a consequence. Key questions are:

1. Will respondents tend to accurately interpret the data and therefore select the 'Crime is stable' option?
2. Will the group of respondents who select 'Crime is stable' tend to enter lower scores in respect of their level of concern?
3. Will this group of respondents tend to be less likely to enact disproportionate / unwarranted behavioural responses?

Question 1, associated response options and frequencies are presented in Table 5.4:

<sup>26</sup> The stimulus is deliberately simplified; it omits numeric values so as to encourage respondents to make an assessment based on its general appearance (i.e. visual data pattern), rather than volume of recorded crimes.

Table 5.4: SPC Stimulus (Question1)

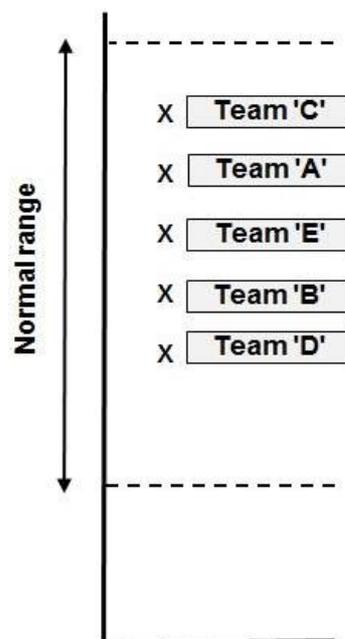
Q1: "In respect of the crime rate, which of the following does the chart appear to indicate?" Crime is:		
Response	Frequency	Percent
Increasing	333	7.4%
Decreasing	156	3.5%
Stable	3,689	82.3%
Don't know	305	6.8%
TOTAL	4,483	100.0%

A Chi-Square test produced output of 7863.171 ( $p = <0.001$ ,  $w = 1.32$ ), indicating a very large effect; consequently, the analysis identifies a strong tendency for respondents to accurately interpret the stimulus in the expected direction.

### 5.2.5 Contextualised Peer Comparisons

In a similar fashion to the manner in which SPC charts promote accurate data interpretation, it was previously suggested that equivalent methods of presenting comparative peer performance information may do the same. Although the contextualised peer comparison methodology proposed herein is not routinely used within UK police forces, the format was considered worthy of assessment during the experimental phase of this study, to examine whether this is indeed the case. The stimulus used in the tests is presented below.

Figure 5.5: Contextualised peer comparison chart stimulus



Again, although the stimulus is visually more complex than others tested, it is speculated that respondents may well draw the appropriate conclusion (i.e. Team ‘D’s performance is not significantly different from the peer group), and therefore select the ‘Performance is acceptable’ option. This is despite Team ‘D’ being in the same position as it was within the league table stimulus (i.e. ‘bottom’). Consequently, the analysis will examine whether respondents do indeed tend to draw this conclusion, and if so, how they react as a consequence. Key questions are:

1. Will respondents tend to accurately interpret the chart and therefore select the ‘Performance is acceptable’ option?
2. Will the group of respondents who select ‘Performance is acceptable’ tend to enter lower scores in respect of their level of concern?
3. Will this group of respondents tend to be less likely to enact disproportionate / unwarranted behavioural responses?

Question 1, associated response options and frequencies are presented in Table 5.5:

*Table 5.5: Contextualised peer comparison stimulus (Question 1)*

Q1: “In respect of Team ‘D’s position in the chart, which of the following appears to most accurately describe the team’s performance?” Performance is:		
Response	Frequency	Percent
Good	315	7.1%
Poor	593	13.3%
Acceptable	2,985	66.9%
Don’t know	571	12.8%
<i>TOTAL</i>	<i>4,464</i>	<i>100.0%</i>

A Chi-Square test produced output of 4216.233 ( $p = <0.001$ ,  $w = 0.94$ ), indicating a large effect. Overall, therefore, the analysis identifies a strong tendency for respondents to accurately interpret the stimulus in the expected direction.

### **5.2.6 Question 1: Summary of Findings**

With the exception of the league table stimulus, respondents tended to interpret the stimuli in the anticipated direction; this suggests performance information users accept at face value whatever data display appears to convey, whether or not a particular format may be considered reliable; this tendency evidently extends to

engagement with formats that depict variance from isolated reference points and applies even where insufficient information is provided to draw any meaningful conclusions.

Furthermore, in the cases of the first three stimuli, this occurred even though such scant information was provided that the appropriate response should have been 'Don't know'. It was also notable that respondents tended to accurately interpret the latter two stimuli, even though they were more visually complex. In the case of the league table stimulus, a large proportion of respondents selected 'Don't know', whilst a similar proportion recorded an interpretation in the expected direction for the stimulus; potential reasons for this will be explored in due course.

The results also reflect output of the thematic analysis; robust data display formats were found to promote greater accuracy of interpretation, whilst simplistic formats consistently misled users. The findings also confirm that the perception of negative variance from a reference point does indeed cause performance information users to envisage deficiencies. This suggests the interaction between performance information and users is pivotal in affecting their subsequent cognitive processes; a prospect first encountered during the thematic analysis.

Finally, these experiments demonstrate users tend to formulate unreliable assumptions when interacting with certain formats; this also reflects the findings of the thematic analysis and indicates that data display could exert a direct influence on the operation of the proposed mechanism. Subsequent tests will assess the nature and strength of potential effects on performance information users' behaviour and whether their reactions may ultimately contribute towards behavioural dysfunction.

### **5.3 Question 2: Effect of Interpretation on Levels of Concern**

The previous tests were designed to solely assess respondents' interpretation of the various data display formats and their propensity to draw conclusions even when presented with limited information. Subsequent tests aim to explore how respondents react as a consequence of their assumptions about performance, beginning with an assessment of whether their interpretations induce feelings of concern. Data were captured on a 5-point Likert-type scale; again, the question and statistics pertaining to each format are presented in turn.

### 5.3.1 Binary Comparisons

Question 2, associated response options and frequencies for the binary comparison stimulus are presented in Table 5.6:

Table 5.6: Binary comparison stimulus (Question 2)

Q2: "As a result of the information contained in the table, how likely is it that you would feel concerned about the crime rate?"		
Response	Frequency	Percent
Very unlikely	182	3.5%
Unlikely	1,156	22.0%
Don't know	532	1.1%
Likely	2,719	51.7%
Very likely	667	12.7
TOTAL	5,256	100.0%

Skewness was found to be -0.600 (weighting towards higher levels of concern) and a Chi-Square test produced output of 3,772.096 ( $p = <0.001$ ,  $w = 0.85$ ), signifying a large effect. Furthermore, a Pearson's Chi-Square test produced output of 896.011 ( $p = <0.001$ ), indicating there is a significant relationship between the variables in Questions 1 and 2. The appropriate measure of effect size for this test is Cramer's  $V$  statistic, which was found to be 0.238, indicating a small-to-medium effect.<sup>27</sup>

A new variable was then computed to enable groups to be compared using the Mann-Whitney U test for independent samples. Respondents who selected the 'Increasing' option in Question 1 were placed into one group, whilst those who chose any of the remaining options were coded into a separate group. The objective was to establish if a significant difference exists between the two groups; i.e. are those who interpret the stimulus to mean crime is increasing more likely to record higher levels of concern than other respondents?

This test produced statistically significant results ( $p = <0.001$ ,  $r = 0.27$ ) with the group who believed crime was increasing displaying a higher mean rank (2,757) than the group who chose other options (1,501); therefore, members of this group were significantly more likely to experience higher levels of concern than other respondents.

<sup>27</sup> 1. For all Cramer's  $V$  effect size statistics, the  $p$  value is identical to that reported within the relevant test. 2. The Pearson's Chi-Square statistic and effect size are accompanied by an explanation of their function; thereafter, they will simply be reported alongside the other results.

### 5.3.2 League Tables

Question 2, associated response options and frequencies for the league table stimulus are presented in Table 5.7:

Table 5.7: League table stimulus (Question 2)

Q2: "As a result of the information contained in the table, how likely is it that you would feel concerned about Team 'D's performance?"		
Response	Frequency	Percent
Very unlikely	215	3.1%
Unlikely	604	11.5%
Don't know	1,554	29.7%
Likely	2,064	39.5%
Very likely	794	15.2%
TOTAL	5,231	100.0%

Skewness was found to be -0.486 (weighting towards higher levels of concern) and a Chi-Square test produced output of 2,144.730 ( $p = <0.001$ ,  $w = 0.64$ ), indicating a large effect. A Pearson's Chi-Square test produced output of 2,547.806 ( $p = <0.001$ ,  $V = 0.403$ ), denoting a significant relationship between the variables in Questions 1 and 2, with a medium-to-large effect.

A Mann-Whitney U test was then conducted to establish if a significant difference exists in respect of levels of concern between those who believed Team 'D's performance was poor and other respondents. This test produced statistically significant results ( $p = <0.001$ ,  $r = 0.58$ ) with the group who believed performance was poor displaying a higher mean rank (3,531), compared to the group who chose the other options (1,861). Therefore, this group was significantly more likely to exhibit higher levels of concern than other respondents.

### 5.3.3 Numerical Targets

Question 2, associated response options and frequencies for the numerical target stimulus are presented in Table 5.8.

Table 5.8: Numerical target stimulus (Question 2)

Q2: "As a result of the information contained in the table, how likely is it that you would feel concerned about performance?"		
Response	Frequency	Percent
Very unlikely	155	3.0%
Unlikely	519	9.9%
Don't know	409	7.8%
Likely	2,591	49.5%
Very likely	1,558	29.8%
<i>TOTAL</i>	<i>5,232</i>	<i>100.0%</i>

Skewness was -1.106 (weighting towards higher levels of concern) and a Chi-Square test produced output of 3,943.566 ( $p = <0.001$ ,  $w = 0.87$ ), denoting a large effect. A Pearson's Chi-Square test produced output of 2,068.058 ( $p = <0.001$ ,  $V = 0.363$ ), indicating there is a significant relationship between the variables in Questions 1 and 2, with a medium effect.

A Mann-Whitney U test was then conducted to establish if a significant difference exists in respect of levels of concern between those who believed performance was poor and other respondents. This test produced statistically significant results ( $p = <0.001$ ,  $r = 0.45$ ) with the group who believed performance was poor displaying a higher mean rank (2,953) than the group who chose the other options (1,355). Therefore, members of this group were significantly more likely to experience higher levels of concern than other respondents.

### 5.3.4 SPC Charts

Question 2, associated response options and frequencies for the SPC chart stimulus are presented in Table 5.9.

Table 5.9: SPC stimulus (Question 2)

Q2: "As a result of the information contained in the chart, how likely is it that you would feel concerned about the crime rate?"		
Response	Frequency	Percent
Very unlikely	548	12.3%
Unlikely	2,606	58.6%
Don't know	582	13.1%
Likely	651	14.6%
Very likely	62	1.4%
<i>TOTAL</i>	<i>4,449</i>	<i>100.0%</i>

Skewness was 0.826 (weighting towards lower levels of concern); a Chi-Square test produced output of 4,382.096 ( $p = <0.001$ ,  $w = 0.99$ ), signifying a large effect. A Pearson’s Chi-Square test produced output of 1,263.020 ( $p = <0.001$ ,  $V = 0.308$ ), indicating there is a significant relationship between the variables in Questions 1 and 2, with a medium effect.

A Mann-Whitney U test was then conducted to establish if a significant difference exists in respect of levels of concern between those who believed crime was stable and other respondents. This test produced statistically significant results ( $p = <0.001$ ,  $r = 0.33$ ) with the group who believed crime was stable displaying a lower mean rank (2,053) than the group who chose the other options (3,030). Therefore, members of this group were significantly more likely to experience lower levels of concern than other respondents.

### 5.3.5 Contextualised Peer Comparisons

Question 2, associated response options and frequencies for the contextualised peer comparison stimulus are presented in Table 5.10.

*Table 5.10: Contextualised peer comparison stimulus (Question 2)*

Q2: “As a result of the information contained in the chart, how likely is it that you would feel concerned about Team ‘D’s performance?”		
Response	Frequency	Percent
Very unlikely	449	10.1%
Unlikely	2,303	52.1%
Don’t know	680	15.4%
Likely	834	18.9%
Very likely	158	3.6%
<b>TOTAL</b>	<b>4,424</b>	<b>100.0%</b>

Skewness was 0.650 (weighting towards lower levels of concern) and a Chi-Square test produced output of 3,135.143 ( $p = <0.001$ ,  $w = 0.84$ ), indicating a large effect. A Pearson’s Chi-Square test produced output of 3,553.790 ( $p = <0.001$ ,  $V = 0.517$ ), denoting a significant relationship between the variables in Questions 1 and 2, with a large effect.

A Mann-Whitney U test was then conducted to establish if a significant difference exists in respect of levels of concern between those who believed Team ‘D’s

performance was acceptable and other respondents. This test produced statistically significant results ( $p = <0.001$ ,  $r = 0.38$ ) with the group who believed performance was acceptable displaying a lower mean rank (1,902) than the group who chose the other options (2,844). Therefore, those in this group were significantly more likely to exhibit lower levels of concern than other respondents.

### **5.3.6 Question 2: Summary of Findings**

The tests in this section produced significant findings which demonstrate a strong relationship between respondents' initial interpretation of the data and subsequent levels of concern. Again, regardless of the reliability of a particular data display format, respondents formulate assumptions as a result of what it purports to depict; this appears to be directly responsible for causing either substantially elevated or reduced levels of concern.

For the binary comparisons, league tables and numerical targets stimuli, respondents consistently recorded elevated levels of concern due to being misled by these formats; conversely, as the SPC chart and contextualised peer comparison stimuli depicted stable patterns, the majority of respondents were able to recognise this and therefore did not exhibit undue concern.

However, it is pertinent to acknowledge that a stable data set depicting crime rates should not necessarily assuage concern about crime – for example, the crime rate could be so high that its sheer volume induces concern, or the type of crime might be so alarming that even a stable pattern would invite disquiet (this prospect is touched upon in Section 7.5). Nevertheless, the aim of these tests is to specifically assess how respondents react to data display and it is apparent that where patterns do not depict deterioration or elements of apparent significance, associated concern is mitigated.

These findings therefore not only confirm a relationship between data display and interpretation, but also establish a further strong association between users' perceptions about performance and consequent levels of concern. The output of the tests further supports the likelihood of there being a mechanism operating with contrasting polarities, the direction of which seems largely dependent on the type of data display format being utilised.

## 5.4 Question 3: Behavioural Responses

This section examines whether users' interpretations, assumptions and levels of concern impact on their behavioural responses, and if so, whether they tend to engage in conduct likely to act as antecedents to dysfunction. As the thematic analysis found that the types of adverse behaviours listed in Question 3 were strongly associated with particular data display formats, the forthcoming tests aim to establish a) whether there is indeed a substantial link between certain formats and greater likelihood of dysfunctional responses, and b) whether the operation of the proposed mechanism may be a primary factor in producing such reactions.

### 5.4.1 Binary Comparisons

Respondents were therefore exposed to the binary comparison stimulus once more and asked:

*Q3. "As a result of the information contained in the table, how likely is it that you would respond as follows?"*

They were then provided with a list of behaviours and invited to record responses on a series of 5-point Likert-type scales, as below.

1. Do nothing.<sup>28</sup>
2. Ask for an explanation about performance.
3. Communicate an expectation there should be an improvement.
4. Initiate an operational response (e.g. commission further research, change tactics, deploy resources).

Where performance information users engage with unreliable forms of data display, the latter three behaviours have been shown to act as antecedents to dysfunction. Therefore during these tests, such responses are designated as 'disproportionate / unwarranted behavioural responses', or 'DUBR'. Results for the four individual variables pertaining to the binary comparison stimulus are provided in in Table 5.11, followed by detailed analysis of the aggregated DUBR variable (N = 4,917).

---

<sup>28</sup> The variable "Do nothing" was reverse-coded to enable aggregation of the variables pertaining to this question.

Table 5.11: Question 3 variables data table (Binary Comparisons)<sup>29</sup>

Variable	Median	Skewness	Spearman's rho
Do nothing	2 (unlikely)	0.928	-0.479 ( $p = <.001$ )
Ask for an explanation	4 (likely)	-0.938	0.420 ( $p = <.001$ )
Demand improvement	4 (likely)	-0.402	0.526 ( $p = <.001$ )
Operational response	4 (likely)	-0.460	0.460 ( $p = <.001$ )

The table shows each central tendency leans towards DUBR. Furthermore, the Spearman's *rho* statistic indicates each variable correlates strongly with the 'level of concern' variable, and in the expected direction. In respect of the aggregated variable, skewness was -0.759 (weighting towards higher levels of DUBR) and a Chi-Square test produced output of 3,717.634 ( $p = <0.001$ ,  $w = 0.87$ ), indicating a large effect. The results of analysis of associations between DUBR and output of previous tests also show significant relationships between variables at Questions 1 and 3, as well as those at 2 and 3, as below:

Table 5.12: Question 3 aggregated variable statistics table (Binary Comparisons)

<b>Pearson's Chi-Square (variables in Questions 1 and 3)</b>	387.045 ( $p = <0.001$ , $V = 0.162$ ) (small effect size)
<b>Pearson's Chi-Square (variables in Questions 2 and 3)</b>	3683.679 ( $p = <0.001$ , $V = 0.433$ ) (medium-to-large effect size)
<b>Spearman's rho (variables in Questions 2 and 3)</b>	0.609 ( $p = <0.001$ ) (strong positive correlation)

A Mann-Whitney U test was then conducted to establish whether those who interpreted the stimulus to mean crime was increasing responded significantly differently to other respondents when presented with the options in Question 3. This test produced statistically significant results ( $p = <0.001$ ,  $r = 0.22$ ), with the group who believed crime was increasing displaying a higher mean rank (2,563) than the group who chose the other options (1,546). Therefore, members of this group were significantly more likely to enact DUBR than other respondents.

<sup>29</sup> The median is reported, as these are the results of non-parametric testing.

A further Mann-Whitney U test was conducted to establish whether there was a significant difference between those who selected ‘Very unlikely’ or ‘Unlikely’ at Question 2 (likelihood of concern), versus those who chose ‘Likely’ or ‘Very likely’, when presented with the options in Question 3. The results were statistically significant ( $p = <0.001$ ,  $r = 0.55$ ) with a large effect size; those who exhibited higher levels of concern displayed a higher mean rank (2,645) than others (1,095). Therefore, respondents who recorded higher levels of concern exhibited the greatest tendency to initiate DUBR.

### 5.4.2 League Tables

Respondents were presented with the same options as previously, and asked:

*Q3. “As a result of the information contained in the table, and specifically in relation to Team ‘D’, how likely is it that you would respond as follows?”*

Statistics for individual variables are produced in Table 5.13 (N = 4,927).

*Table 5.13: Question 3 variables data table (League tables)*

<b>Variable</b>	<b>Median</b>	<b>Skewness</b>	<b>Spearman’s rho</b>
Do nothing	2 (unlikely)	0.935	-0.469 ( $p = <0.001$ )
Ask for an explanation	4 (likely)	-1.257	0.437 ( $p = <0.001$ )
Demand improvement	3 (Don’t know)	-0.423	0.610 ( $p = <0.001$ )
Operational response	4 (likely)	-0.565	0.377 ( $p = <0.001$ )

The table shows each central tendency is weighted towards DUBR. Additionally, the Spearman’s *rho* statistic shows each variable correlates strongly with the ‘level of concern’ variable at Question 2, and in the anticipated direction. In respect of the aggregated variable, skewness was -0.778 (weighting towards higher levels of DUBR) and a Chi-Square test produced output of 4,404.24 ( $p = <0.001$ ,  $w = 0.95$ ), indicating a large effect. Further tests for associations between DUBR and output of previous tests also show significant relationships between variables, as below:

Table 5.14: Question 3 aggregated variable statistics table (League tables)

<b>Pearson's Chi-Square (variables in Questions 1 and 3)</b>	1151.81 ( $p = <0.001$ , $V = 0.279$ ) (small-to-medium effect size)
<b>Pearson's Chi-Square (variables in Questions 2 and 3)</b>	4222.822 ( $p = <0.001$ , $V = 0.463$ ) (medium-to-large effect size)
<b>Spearman's rho (variables in Questions 2 and 3)</b>	0.622 ( $p = <0.001$ ) (strong positive correlation)

A Mann-Whitney U test was then conducted to establish whether those who interpreted the stimulus to mean Team 'D's performance was poor responded significantly differently to other respondents. This test produced statistically significant results ( $p = <0.001$ ,  $r = 0.43$ ), with the group who believed performance was poor displaying a higher mean rank (3,126) than the group who selected the other options (1,915). Therefore, members of this group were significantly more likely to initiate DUBR than other respondents.

As previously, a further Mann-Whitney U test was conducted to establish whether there were significant differences in how different groups from Question 2 responded to Question 3. The results were statistically significant ( $p = <0.001$ ,  $r = 0.52$ ) with a large effect size; those who exhibited higher levels of concern displayed a higher mean rank (2,014) than others (754). Again, therefore, respondents who recorded higher levels of concern exhibited the greatest tendency to initiate DUBR.

### 5.4.3 Numerical Targets

Respondents were presented with the same options as previously, and asked:

*Q3. "As a result of the information contained in the table, how likely is it that you would respond as follows?"*

Statistics for individual variables are produced in Table 5.15 (N = 4,927).

Table 5.15: Question 3 variables data table (Numerical targets)

Variable	Median	Skewness	Spearman's rho
Do nothing	2 (unlikely)	1.457	-0.521 ( $p = <0.001$ )
Ask for an explanation	4 (likely)	-1.531	0.516 ( $p = <0.001$ )
Demand improvement	4 (likely)	-0.938	0.574 ( $p = <0.001$ )
Operational response	4 (likely)	-1.110	0.458 ( $p = <0.001$ )

Again, all central tendencies are weighted towards DUBR; furthermore, the Spearman's  $\rho$  statistic indicates each variable correlates strongly with the 'level of concern' variable, and in the anticipated direction. With regard to the aggregated variable, skewness was -1.244 (weighting towards higher levels of DUBR) and a Chi-Square test produced output of 4,424.04 ( $p = <0.001$ ,  $w = 1.00$ ), indicating a large effect. Additional tests for associations between DUBR and output of previous tests also show significant relationships between variables, as below:

Table 5.16: Question 3 aggregated variable statistics table (Numerical targets)

<b>Pearson's Chi-Square (variables in Questions 1 and 3)</b>	1084.425 ( $p = <0.001$ , $V = 0.271$ ) (small-to-medium effect size)
<b>Pearson's Chi-Square (variables in Questions 2 and 3)</b>	4261.809 ( $p = <0.001$ , $V = 0.465$ ) (medium-to-large effect size)
<b>Spearman's rho (variables in Questions 2 and 3)</b>	0.647 ( $p = <0.001$ ) (strong positive correlation)

A Mann-Whitney U test was conducted to establish whether those who interpreted the stimulus to mean performance was poor responded significantly differently to others. This test produced statistically significant results ( $p = <0.001$ ,  $r = 0.38$ ), with the group who believed performance was poor displaying a higher mean rank (2,737) than those who selected other options (1,387). Therefore, members of this group were significantly more likely to enact DUBR.

A further Mann-Whitney U test was then conducted to establish whether there were significant differences in how different groups from Question 2 responded to Question 3. The results were statistically significant ( $p = <0.001$ ,  $r = 0.45$ ) with a large effect size; those who exhibited higher levels of concern displayed a higher

mean rank (2,508) than others (742); therefore respondents who recorded higher levels of concern exhibited the greatest tendency to initiate DUBR.

#### 5.4.4 SPC Charts

Respondents were presented with the same options as previously, and asked:

*Q3. “As a result of the information contained in the chart, how likely is it that you would respond as follows?”*

Statistics for individual variables are produced in Table 5.17 (N = 4,738).

*Table 5.17: Question 3 variables data table (SPC charts)*

Variable	Median	Skewness	Spearman's rho
Do nothing	3 (Don't know)	0.117	-0.469 ( $p = <0.001$ )
Ask for an explanation	3 (Don't know)	0.006	0.492 ( $p = <0.001$ )
Demand improvement	3 (Don't know)	0.123	0.509 ( $p = <0.001$ )
Operational response	3 (Don't know)	-0.108	0.447 ( $p = <0.001$ )

The results are noticeably less skewed than the output observed for the previous data display formats. Respondents still displayed a slight reluctance to ‘Do nothing’ (a condition identified during thematic analysis) and a tendency to ‘Initiate an operational response’ (albeit with much greater restraint than previously observed); however, central tendencies were found to be either in the expected direction, or markedly more balanced than previously. Furthermore, the Spearman’s *rho* statistic confirms that each variable correlates strongly with the ‘level of concern’ variable.

In respect of the aggregated variable, skewness was -0.144 (marginal weighting towards higher levels of DUBR) and a Chi-Square test produced output of 2,806.181 ( $p = <0.001$ ,  $w = 0.80$ ), indicating a large effect. Additional tests for associations between DUBR and output of previous tests also confirm significant relationships between variables, as below:

Table 5.18: Question 3 aggregated variable statistics table (SPC charts)

<b>Pearson's Chi-Square (variables in Questions 1 and 3)</b>	661.828 ( $p = <0.001$ , $V = 0.224$ ) (small effect size)
<b>Pearson's Chi-Square (variables in Questions 2 and 3)</b>	3796.792 ( $p = <0.001$ , $V = 0.466$ ) (medium-to-large effect size)
<b>Spearman's rho (variables in Questions 2 and 3)</b>	0.593 ( $p = <0.001$ ) (strong positive correlation)

A Mann-Whitney U test was conducted to establish whether those who interpreted the stimulus to mean crime was stable responded significantly differently to others. This test produced statistically significant results ( $p = <0.001$ ,  $r = 0.22$ ), with the group who believed crime was stable displaying a lower mean rank (2,065) than those who selected other options (2,778). Therefore, members of this group were significantly less likely to enact DUBR.

A further Mann-Whitney U test was then conducted to establish whether there were significant differences in how different groups from Question 2 responded to Question 3. The results were statistically significant ( $p = <0.001$ ,  $r = 0.52$ ) with a large effect size; those who exhibited lower levels of concern displayed a lower mean rank (1,636) than others (3,112); therefore respondents who recorded lower levels of concern were significantly less likely to initiate DUBR.

#### 5.4.5 Contextualised Peer Comparisons

Respondents were presented with the same options as previously, and asked:

*Q3. "As a result of the information contained in the chart, and specifically in relation to Team 'D', how likely is it that you would respond as follows?"*

Statistics for individual variables are produced in Table 5.19 (N = 4,372).

Table 5.19: Question 3 variables data table (Contextualised peer comparisons)

Variable	Median	Skewness	Spearman's rho
Do nothing	3 (Don't know)	0.146	-0.490 ( $p = <0.001$ )
Ask for an explanation	4 (likely)	-0.289	0.556 ( $p = <0.001$ )
Demand improvement	3 (Don't know)	-0.012	0.601 ( $p = <0.001$ )
Operational response	3 (Don't know)	-0.070	0.486 ( $p = <0.001$ )

Again, central tendencies were more balanced than those pertaining to the first three data display formats, although respondents still displayed a slight reluctance to ‘Do nothing’ and a tendency to ‘Initiate an operational response’ (albeit with comparative restraint). As previously, the Spearman’s *rho* statistic indicates each variable correlates strongly with the ‘level of concern’ variable.

For the aggregated variable, skewness was -0.226 (marginal weighting towards higher levels of DUBR) and a Chi-Square test produced output of 2,571.030 ( $p = <0.001$ ,  $w = 0.77$ ), denoting a large effect. Additional tests for associations between DUBR and output of previous tests also show significant relationships between variables, as below:

*Table 5.20: Question 3 aggregated variable statistics table (Contextualised peer comparisons)*

<b>Pearson’s Chi-Square (variables in Questions 1 and 3)</b>	1421.217 ( $p = <0.001$ , $V = 0.329$ ) (medium effect size)
<b>Pearson’s Chi-Square (variables in Questions 2 and 3)</b>	4602.534 ( $p = <0.001$ , $V = 0.513$ ) (large effect size)
<b>Spearman’s <i>rho</i> (variables in Questions 2 and 3)</b>	0.650 ( $p = <0.001$ ) (strong positive correlation)

A Mann-Whitney U test was conducted to establish whether those who interpreted the stimulus to mean Team ‘D’’s performance was acceptable responded significantly differently to others. This test produced statistically significant results ( $p = <0.001$ ,  $r = 0.23$ ), with the group who believed performance was acceptable displaying a lower mean rank (1,982) than those who chose other options (2,605). Therefore, members of this group were significantly less likely to enact DUBR.

A further Mann-Whitney U test was then conducted to establish whether there were significant differences in how different groups from Question 2 responded to Question 3. The results were statistically significant ( $p = <0.001$ ,  $r = 0.61$ ) with a large effect size; those who exhibited lower levels of concern displayed a lower mean rank (1,461) than others (2,944); therefore respondents who recorded lower levels of concern were significantly less likely to initiate DUBR.

### **5.4.6 Question 3: Summary of Findings**

The tests in this section identify strong tendencies that indicate the degree and polarity of behavioural responses are firmly linked to the type of data display format used. Significant relationships were identified between variables, which suggest performance information users' initial interpretation of data directly influences subsequent choices and outcomes. The apparent effect on decision-making echoes what was discovered during thematic analysis and confirms data display appears to exert a direct influence on users' thought processes and behaviour.

Notably, considerable differences in the extent of disproportionate / unwarranted behavioural responses were observed when users engaged with different formats; in respect of binary comparisons, league tables and numerical targets, respondents consistently recorded high levels of adverse behavioural tendencies, whereas with SPC charts and contextualised peer comparisons, this effect was comprehensively moderated.

Furthermore, the contrasting polarity amidst the results reflects that of the sentiment analysis, which found substantial alignment between negative effects and the use of reference-dependent formats, alongside more positive effects where statistically robust forms of data display were employed. Similarly, thematic analysis identified recurrent themes involving antecedents to dysfunction when simplistic formats were used; the experiments found these exact same behaviours were significantly more likely to occur when respondents engaged with these formats.

### **5.5 Regression Analysis**

In order to further assess the apparent relationships between the identified tendencies at Question 3 and the patterns observed amongst other variables, regression analysis was then conducted (see Field, 2013, pp.760-813).<sup>30</sup> Binary logistic regression was considered the most appropriate method, as the outcome variable is conceptualised as a categorical dichotomy, enabling assessments to be made about which of two groups an individual is likely to belong to.

---

<sup>30</sup> The regression analysis conducted for the binary comparisons stimulus is reported in a narrative fashion, in order to provide explanation and transparency regarding the procedure; output for the other data display formats is tabulated.

Furthermore, this type of regression is suited to the non-parametric data generated by the survey instrument, as it makes no assumptions about the distribution of the predictor variables<sup>31</sup> and is capable of handling categorical predictors, such as the nominal descriptors in Question 1 (Brace *et al*, 2009).

Firstly, a dependent variable was generated by recoding the aggregate variable for 'behavioural response' from Question 3, as follows:

1. The 'Don't know' option was removed.
2. Values between 4 and 10 inclusive were recoded as 1; values between 11 and 16 inclusive were recoded as 2.<sup>32</sup>

The dependent variable is therefore conceptualised as respondents' tendency to enact disproportionate / unwarranted behavioural responses (DUBR).

Individual categorical variables were computed for each of the four options in Question 1 (i.e. 'Increasing', 'Decreasing', 'Stable', and 'Don't know'). This was done in order to assess the strength of each condition as a predictor for disproportionate / unwarranted behavioural responses. These variables were placed alongside the 'level of concern' variable produced at Question 2, and binary logistic regression was conducted against the dependent variable 'DUBR'.

### **5.5.1 Binary Comparisons**

Following the style of previous commentary on the quantitative analysis, the narrative pertaining to the first stimulus tested is more descriptive than subsequent blocks, whereupon results are reported relatively parsimoniously. With regard to the binary comparison stimulus, a total of 3,334 cases were analysed and the model was found to significantly predict behavioural responses, with a medium effect size (omnibus Chi-Square = 1001.406,  $p = <0.001$ , Nagelkerke's  $R^2$  approximation =

---

<sup>31</sup> Whilst being fully cognisant of the critical realist perspective on 'prediction', the terms 'predictor variable' and 'dependent variable' are used here simply as terms integral to regression analysis.

<sup>32</sup> This procedure produced a new variable with a range between 4 and 16. The rationale for removing the 'Don't know' option was to enable generation of a binary grouping variable that categorises a greater or lesser tendency to enact disproportionate / unwarranted behavioural responses – it would therefore be inappropriate for the mid-point of the original scales in Question 3 (i.e. 'Don't know') to contribute towards a dependent variable specifically designed to establish likelihood of *action*.

0.387).<sup>33</sup> A Hosmer and Lemeshow test confirmed the model fits the data well (Chi-Square = 2.869,  $p = 0.412$ ); this statistic, which is a test of the *null hypothesis* that the model is a good fit, was significant due to the  $p$  value being well in excess of 0.05 (see Brace *et al*, 2009, p.332; Field, 2013, pp.765, 876).

The model correctly predicted 87.6% of respondents likely to enact disproportionate / unwarranted behavioural responses and 68.2% of those unlikely to do so. Overall, 82.9% of predictions were accurate. Therefore, the model is considered highly effective at predicting when respondents are likely to enact DUBR, whilst being slightly less accurate at predicting when they are unlikely to do so.

Of the categorical variables derived from Question 1, only ‘Crime is increasing’ was found to be a significant predictor (Wald = 18.125,  $p = <0.001$ ), along with the ‘level of concern’ variable (Wald = 672.038,  $p = <0.001$ ). Furthermore, as the Wald statistic indicates whether the  $b$  coefficient for each predictor is significantly different from zero (Field, 2013, p.784), it can be concluded that these two variables appear most useful in predicting the DUBR dependent variable.

The model underwent two further iterations, following the principles of parsimony (Field 2013, p.768); non-significant predictors were stripped away to produce the optimal model; this confirmed that the variables ‘Crime is increasing’ and ‘level of concern’ reliably predicted the DUBR dependent variable. Again,  $N = 3,334$ , with the model significantly predicting outcomes, and a medium effect size (omnibus Chi-Square = 995.805,  $p = <0.001$ , Nagelkerke’s  $R^2$  approximation = 0.386). A Hosmer and Lemeshow test confirmed a good fit (Chi-Square = 3.045,  $p = 0.385$ ).

This iteration of the model correctly predicted 87.7% of those respondents likely to enact DUBR and 68.1% of those unlikely to do so. Overall, 83.0% of predictions were accurate. Therefore, this iteration is marginally more accurate than the original configuration and may be considered optimal in terms of parsimony.

It was established that the model significantly predicts the likelihood of DUBR being enacted as a consequence of the way respondents interpreted the stimulus. Where

---

<sup>33</sup> Nagelkerke’s  $R^2$  approximation is reported as the measure of the model’s effect size in preference to Cox and Snell’s  $R^2$  approximation as, unlike the latter, it is capable of reaching a maximum value of 1, thereby enabling clearer interpretation (see Field, 2013, pp.765-766).

respondents believed crime to be increasing, the Exp(B) statistic indicates there was an increase in the odds of DUBR being enacted by a factor of 1.864 (95% CI = 1.353 and 2.567) for each unit increase in this predictor variable. For each unit increase in the ‘level of concern’ predictor variable, there was an increase in the odds of DUBR being enacted by a factor of 3.223 (95% CI = 2.951 and 3.520).

### 5.5.2 League Tables

In respect of the league table stimulus, regression was conducted following the same procedure as previously. The analysis produced a model that established the only significant predictors were the variable ‘Performance is poor’ from Question 1 and the ‘level of concern’ variable from Question 2. Consequently, further regression was conducted under the principles of parsimony, using these two variables. Output from this optimal model is displayed in Table 5.21.

*Table 5.21: League Tables Regression Analysis Data Table*

<b>N</b>	2,775
<b>Omnibus Chi-Square</b>	824.392 ( $p = <0.001$ , Nagelkerke's $R^2$ approximation = 0.424) (Medium effect size)
<b>Hosmer and Lemeshow statistic</b>	Chi-Square = 4.600, $p = 0.467$
<b>Predictive accuracy (DUBR likely)</b>	93.1%
<b>Predictive accuracy (DUBR unlikely)</b>	57.4%
<b>Overall predictive accuracy</b>	86.8%
<b>Significant predictor variables</b>	Q1. ‘Performance is poor’ Wald = 5.935, $p = <0.001$ , Exp(B) = 1.775 (95% CI = 1.356 – 2.235) Q2. ‘Level of concern’ Wald = 402.680, $p = <0.001$ , Exp(B) = 3.361 (95% CI = 2.990 – 3.778)

The results show that the model significantly predicts the likelihood of DUBR. Where respondents believed performance to be poor, there was an increase in the odds of DUBR being enacted by a factor of 1.775 for each unit increase in this

predictor variable. For each unit increase in the ‘level of concern’ predictor variable, there was an increase in the odds of DUBR being enacted by a factor of 3.361.

### 5.5.3 Numerical Targets

In respect of the numerical target stimulus, regression produced a model that established the only significant predictors were the ‘Performance is poor’ variable from Question 1 and the ‘level of concern’ variable from Question 2. Consequently, further regression was conducted using these two variables; output from the optimal model is displayed in Table 5.22.

*Table 5.22: Numerical Targets Regression Analysis Data Table*<sup>34</sup>

<b>N</b>	3,834
<b>Omnibus Chi-Square</b>	873.468 ( $p = <.001$ , Nagelkerke's $R^2$ approximation = 0.420) (Medium effect size)
<b>Hosmer and Lemeshow statistic</b>	Chi-Square = 8.812, $p = 0.032$
<b>Predictive accuracy (DUBR likely)</b>	96.4%
<b>Predictive accuracy (DUBR unlikely)</b>	47.5%
<b>Overall predictive accuracy</b>	91.4%
<b>Significant predictor variables</b>	Q1. ‘Performance is poor’ Wald = 6.299, $p = 0.012$ , Exp(B) = 1.457 (95% CI = 1.086 – 1.955) Q2. ‘Level of concern’ Wald = 495.826, $p = <0.001$ , Exp(B) = 3.816 (95% CI = 3.391 – 4.293)

It can be seen that the model significantly predicts the likelihood of DUBR. Where respondents believed performance to be poor, there was an increase in the odds of DUBR being enacted by a factor of 1.457 for each unit increase in this predictor variable. For each unit increase in the ‘level of concern’ predictor variable, there was an increase in the odds of DUBR being enacted by a factor of 2.990.

<sup>34</sup> It is observed that the output from the Hosmer and Lemeshow test suggests some data do not significantly fit the model. Whilst this is acknowledged, Wuensch (2014) warns this test is susceptible to producing significant results even where the fit is good, and particularly where the sample size is large. In any case, the model was highly effective at predicting DUBR.

### 5.5.4 SPC Charts

Regression was next conducted in respect of the SPC chart stimulus. It was established the only significant predictors were the variable ‘Crime is stable’ from Question 1 and the ‘level of concern’ variable from Question 2. Consequently, further regression was conducted using these two variables; output from the optimal model is displayed in Table 5.23.

*Table 5.23: SPC Charts Regression Analysis Data Table*

<b>N</b>	2,717
<b>Omnibus Chi-Square</b>	1019.772 ( $p = <0.001$ , Nagelkerke's $R^2$ approximation = 0.422) (Medium effect size)
<b>Hosmer and Lemeshow statistic</b>	Chi-Square = 3.920, $p = 0.417$
<b>Predictive accuracy (DUBR likely)</b>	50.4%
<b>Predictive accuracy (DUBR unlikely)</b>	96.5%
<b>Overall predictive accuracy</b>	77.8%
<b>Significant predictor variables</b>	Q1. ‘Crime is stable’ Wald = 7.121, $p = 0.008$ , $\text{Exp}(B) = 0.668$ (95% CI = 0.496 – 0.898) Q2. ‘Level of concern’ Wald = 457.975, $p = <0.001$ , $\text{Exp}(B) = 5.167$ (95% CI = 4.446 – 6.006)

Again, it is observed that the model significantly predicts the likelihood of DUBR. Where respondents identified crime was stable, there was a decrease in the odds of DUBR being enacted by a factor of .668 for each unit increase in this predictor.<sup>35</sup> For each unit increase in the ‘level of concern’ predictor variable, there was an increase in the odds of DUBR being enacted by a factor of 5.167. In other words, lower levels of concern equate to reduced levels of DUBR.

<sup>35</sup> A value of less than 1 indicates that as the predictor increases, the odds of the outcome occurring decrease (Field, 2013, p.786).

### 5.5.5 Contextualised Peer Comparisons

Finally, with regard to the contextualised peer comparison stimulus, regression was conducted, which established the only significant predictors were the variables ‘Performance is good’ and ‘Performance is acceptable’ from Question 1 and the ‘level of concern’ variable from Question 2. Consequently, further regression was conducted using these variables; output from the optimal model is displayed in Table 5.24.

*Table 5.24: Contextualised Peer Comparisons Regression Analysis Data Table*

<b>N</b>	2,737
<b>Omnibus Chi-Square</b>	1376.250 ( $p = <0.001$ , Nagelkerke's $R^2$ approximation = 0.529) (Large effect size)
<b>Hosmer and Lemeshow statistic</b>	Chi-Square = 28.516, $p = 0.250$
<b>Predictive accuracy (DUBR likely)</b>	64.3%
<b>Predictive accuracy (DUBR unlikely)</b>	95.4%
<b>Overall predictive accuracy</b>	81.4%
<b>Significant predictor variables</b>	Q1. ‘Performance is good’ Wald = 8.438, $p = 0.004$ , Exp(B) = 0.471 (95% CI = 0.284– 0.783) Q1. ‘Performance is acceptable’ Wald = 3.612, $p = 0.042$ , Exp(B) = 0.722 (95% CI = 0.516 – 1.010) Q2. ‘Level of concern’ Wald = 458.242, $p = <0.001$ , Exp(B) = 5.515 (95% CI = 4.717 – 6.449)

The results show that the model significantly predicts the likelihood of DUBR. Where respondents believed performance was acceptable or good, there was a decrease in the odds of DUBR being enacted by a factor of 0.471 and 0.722 respectively, for each unit increase in these predictor variables. For each unit increase in the ‘level of concern’ predictor variable, there was an increase in the odds of DUBR being enacted by a factor of 5.515.

### **5.5.6 Regression Analysis: Summary of Findings**

This regression analysis demonstrates that the patterns observed in responses to Questions 1 and 2 are significant predictors for the behavioural tendencies exhibited at Question 3. In respect of binary comparisons, league tables and numerical targets, respondents' (mis)interpretation of these stimuli and consequent levels of concern were confirmed as the origin of the subsequent high levels of disproportionate / unwarranted behavioural responses observed.

Conversely, with regard to the SPC chart and contextualised peer comparison stimuli, accurate interpretation and associated reduced levels of concern were found to be direct precursors for lower levels of DUBR. Typically, these findings were produced with high levels of significance, indicating the models possess clear predictive utility. Overall, regression confirms that performance information users' interpretation of data is a primary factor affecting subsequent behavioural responses.

### **5.6 Additional Analysis**

In addition to the tests discussed above, analysis was undertaken to establish if:

1. Respondents who selected 'Don't know' at Question 1 behaved differently to others in subsequent questions.
2. Qualifying conditions affected how respondents reacted to the stimulus. (For example, whether familiarity with SPC affects the extent of DUBR).

The results are summarised below.

#### **5.6.1 The 'Don't Know' Group**

The logical extension of the assertion that it is unwise to infer meaning from simplistic data display formats, or make judgments about performance based on variance from an isolated reference point, would be that the appropriate option at Question 1 is 'Don't know' for binary comparisons, league tables and numerical targets. Conversely, in respect of the SPC charts and contextualised peer comparisons stimuli, if respondents are unable to interpret these formats, a 'Don't know' response may affect how they react to subsequent questions.

Therefore, tests were conducted to establish if respondents who chose ‘Don’t know’ at Question 1 responded differently to others at Questions 2 and 3. It is acknowledged there are myriad reasons why respondents might select this option – for example, it could signal insight into the limitations of some formats, or simply indicate they were unwilling to commit. Similarly, lack of available information may dissuade some respondents from providing an interpretation. Consequently, findings are presented without speculating about respondents’ motives.

Mann-Whitney U tests were first conducted in respect of each of the stimuli to establish whether those who selected ‘Don’t know’ in Question 1 responded significantly differently to other respondents when asked to indicate their level of concern at Question 2. Further Mann-Whitney U tests were then carried out to establish whether there was a significant difference between the groups in respect of how they responded at Question 3. The results are presented in Table 5.25.

*Table 5.25: Mann-Whitney U-tests (‘Don’t know’ group)*

<b>Question 2</b>	<b>N</b>	<b>U</b>	<b>z</b>	<b>p</b>	<b>r</b>	<b>Outcome</b>
<b>Binary comparisons</b>	5,256	1,070,391	12.497	<0.001	0.17 (small effect size)	Reduced concern
<b>League tables</b>	5,231	5,200,250.5	34.251	<0.001	0.47 (medium to large effect size)	Reduced concern
<b>Numerical targets</b>	5,232	1,876,913	20.492	<0.001	0.28 (small to medium effect size)	Reduced concern
<b>SPC charts</b>	4,449	407,333	-11.164	<0.001	0.17 (small effect size)	Increased concern
<b>Contextualised peer comparisons</b>	4,424	745,850.5	-13.031	<0.001	0.20 (small effect size)	Increased concern
<b>Question 3</b>	<b>N</b>	<b>U</b>	<b>z</b>	<b>p</b>	<b>r</b>	<b>Outcome</b>
<b>Binary comparisons</b>	4,917	897,479.5	8.913	<0.001	0.13 (small effect size)	Reduced DUBR
<b>League tables</b>	4,927	4,929,550	25.425	<0.001	0.36 (medium effect size)	Reduced DUBR
<b>Numerical targets</b>	4,927	1,610,108.5	17.738	<0.001	0.25 (small to medium effect size)	Reduced DUBR
<b>SPC charts</b>	4,378	555,462	-1.892	0.058	N/A	No significant difference
<b>Contextualised peer comparisons</b>	4,372	902,516	-5.670	<0.001	0.09 (small effect size)	Increased DUBR

In relation to binary comparisons, league tables and numerical targets, those who belong to the ‘Don’t know’ group were significantly less likely to experience high levels of concern or enact DUBR. In respect of SPC charts and contextualised peer comparisons, respondents who selected ‘Don’t know’ recorded higher levels of concern. They also exhibited a greater propensity towards DUBR, although this tendency was non-significant in relation to SPC charts and produced a small effect size in respect of contextualised peer comparisons.

These results may arise due to respondents recognising the limitations of the first three formats and therefore choosing the appropriate option, whilst being uncertain about how to interpret the latter two; however it is not possible to draw firm conclusions. Nevertheless, it appears that there are consistently positive effects when users disengage from unreliable data display formats, along with negative effects when they do not or cannot interpret the content of more reliable formats.

### **5.6.2 Organisational Influence**

Analysis was conducted to assess whether there was a relationship between the frequency by which particular data display formats are used in respondents’ forces and the way they responded to each corresponding stimulus.<sup>36</sup> The objective of this analysis was to isolate the effect of engagement with the experimental stimuli from potential external influences generated by the frequency that the various formats were used in their forces.

Cross-tabulation was carried out between the variable at Question 1 and the variable that records how frequently each format is used within respondents’ forces. Further cross-tabulation was then conducted between the ‘frequency’ variable and the ‘level of concern’ variable at Question 2. This was followed by cross-tabulation between the ‘frequency’ variable and the ‘behavioural response’ variable at Question 3.

Next, in order to further assess whether regular use of particular formats in respondents’ forces influences levels of concern or behavioural responses, a binary variable was computed from the ‘frequency’ variable, where the descriptors ‘Never’

---

<sup>36</sup> With regard to the contextualised peer comparisons stimulus, as this format of presenting data is not widely used in UK police forces, no attempt at assessing organisational influence was made.

and ‘Occasionally’ were assigned to one condition, whilst ‘Frequently’ and ‘All of the time’ were assigned to another.

Mann-Whitney U tests were then conducted in respect of the stimuli to establish whether there was a significant difference in respect of how respondents in each group registered their level of concern at Question 2. Further Mann-Whitney U tests were also conducted to establish whether there was a significant difference between the groups in respect of how they responded at Question 3. Output from these tests can be inspected in tables 5.26 and 5.27.

*Table 5.26: Cross-tabulation (Frequency variable against output from Questions 1, 2 and 3)*

<b>Frequency / Q1</b>	<b>N</b>	<b>Chi-square</b>	<b>p</b>	<b>V</b>
<b>Binary comparisons</b>	4,278	9.465	0.663	N/A
<b>League tables</b>	4,214	20.104	0.065	N/A
<b>Numerical targets</b>	4,214	30.713	0.002	0.049 (small effect size)
<b>SPC charts</b>	4,195	64.041	<0.001	0.071 (small effect size)
<b>Frequency / Q2</b>	<b>N</b>	<b>Chi-square</b>	<b>p</b>	<b>V</b>
<b>Binary comparisons</b>	4,278	86.637	<0.001	0.07 (small effect size)
<b>League tables</b>	4,214	32.007	0.010	0.04 (small effect size)
<b>Numerical targets</b>	4,214	66.233	<0.001	0.063 (small effect size)
<b>SPC charts</b>	4,195	87.019	<0.001	0.072 (small effect size)
<b>Frequency / Q3</b>	<b>N</b>	<b>Chi-square</b>	<b>p</b>	<b>V</b>
<b>Binary comparisons</b>	4,277	148.64	<0.001	0.093 (small effect size)
<b>League tables</b>	3,808	13.345	0.647	N/A
<b>Numerical targets</b>	4,006	40.871	0.001	0.101 (small effect size)
<b>SPC charts</b>	3,504	38.743	0.001	0.105 (small effect size)

Table 5.27: Mann-Whitney U-tests (Frequency variable against output from Questions 2 and 3)

Frequency / Q2	N	U	z	p	r	Outcome
Binary comparisons	3,864	1,454,341.5	-1.72	0.085	N/A	No significant difference
League tables	3,810	1,715,773.5	-1.609	0.108	N/A	No significant difference
Numerical targets	4,008	1,021,843	-3.486	<0.001	0.06 (small effect size)	Increased concern
SPC charts	3,505	1,293,209.5	-5.011	<0.001	0.08 (small effect size)	Reduced concern
Frequency / Q3	N	U	z	p	r	Outcome
Binary comparisons	3,863	1,400,866.5	-3.275	0.001	0.05 (small effect size)	Increased DUBR
League tables	3,808	1,741,242	-0.721	0.471	N/A	No significant difference
Numerical targets	4,006	993,749.5	-4.245	<0.001	0.07 (small effect size)	Increased DUBR
SPC charts	3,504	1,281,504.5	-4.837	<0.001	0.08 (small effect size)	Reduced DUBR

In relation to the cross tabulation, there was evidence of relationships between most of the variables tested, albeit with very small effect sizes (the exception being the two pairings found to be non-significant). This alone suggests that the frequency by which these data display formats are used within respondents' forces is not a major factor in how users interpret performance information, or behave as a consequence.

With regard to the Mann-Whitney U tests, the results are mixed; no significant differences were found in respect of league table use and there was no relationship between frequent use of binary comparisons and high levels of concern. However, regular use of binary comparisons and numerical targets was found to be associated with elevated DUBR, with the latter format also being connected to increased levels of concern. Frequent use of SPC charts was associated with lower levels of concern and reduced DUBR. However, it must be noted that all effect sizes were very small.

Overall, it appears the frequency by which particular data display formats are used within respondents' forces is not a major influence on how they responded to the tests. Frequent use of SPC charts appears to exert a limited positive effect, whilst regular use of the other formats was either inconsequential or had a small negative influence. However, there is nothing to suggest respondents' interpretation of the stimulus was affected, alongside a very small impact on behavioural responses.

This indicates the ‘frequency’ qualifying condition does not significantly impact upon behaviour, suggesting the *data display format itself acts a primary influence on behavioural outcomes*, irrespective of the latent influence of common operational practice. This finding is notable, as although the thematic analysis suggested this was the case, it was not possible to isolate the influence of data display in the same way as it has been during the experimental component of the study.

### **5.6.3 Familiarity with Statistical Process Control**

During the pilot phase, familiarity with SPC was identified as a possible moderating factor on how respondents might react to the tests. Consequently, a qualifying question was inserted into the instrument to enable analysis of the presence and extent of this potential influence. The effect of the ‘SPC familiarity’ variable was first assessed by cross-tabulating it with the output of Questions 1, 2 and 3, as per the previous section.

Mann-Whitney U tests were then carried out after generating a binary variable from the ‘SPC familiarity’ variable, where the descriptors ‘Not at all familiar’ and ‘Unfamiliar’ were assigned to one condition, whilst ‘Familiar’ and ‘Very familiar’ were assigned to another.<sup>37</sup> Analysis then proceeded in the same fashion as for the previous section, with tabulated results being presented in tables 5.28 and 5.29.

---

<sup>37</sup> The mid-point (‘Somewhat familiar’) was incorporated into the former category, as unlike previous scales with a ‘Don’t know’ mid-point, this scale was anchored in such a way as to measure ascending degrees of familiarity. As the bulk of responses (64.9%) fell into the ‘Not at all familiar’ and ‘Unfamiliar’ descriptor range, it was considered appropriate to split the original scale in this way, rather than disregard mid-point responses. Furthermore, as only 21.8% of respondents indicated they were ‘Familiar’ or ‘Very familiar’, it was considered that as the purpose of the variable is to establish whether there is an effect when those most familiar with SPC engage in the tests, there was a risk that aligning the mid-point with the latter binary category could have a disproportionate and misleading impact on the analysis, thereby undermining the objective of the variable. The same rationale applies to the ‘Simon Guilfoyle familiarity’ binary variable (below), where just 7.2% of respondents fell into the ‘Familiar’ / ‘Very familiar’ category.

Table 5.28: Cross-tabulation ('SPC familiarity' variable against output from Questions 1, 2 and 3)

<b>SPC Familiarity / Q1</b>	<b>N</b>	<b>Chi-square</b>	<b>p</b>	<b>V</b>
<b>Binary comparisons</b>	4,172	51.673	<0.001	0.064 (small effect size)
<b>League tables</b>	4,172	22.890	0.086	N/A
<b>Numerical targets</b>	4,172	18.928	0.090	N/A
<b>SPC charts</b>	4,172	21.918	0.038	0.072 (small effect size)
<b>Contextualised peer comparisons</b>	4,172	22.368	0.034	0.073 (small effect size)
<b>SPC Familiarity / Q2</b>	<b>N</b>	<b>Chi-square</b>	<b>p</b>	<b>V</b>
<b>Binary comparisons</b>	4,172	38.375	0.001	0.048 (small effect size)
<b>League tables</b>	4,172	36.600	0.002	0.047 (small effect size)
<b>Numerical targets</b>	4,172	30.255	0.017	0.043 (small effect size)
<b>SPC charts</b>	4,172	66.676	<0.001	0.063 (small effect size)
<b>Contextualised peer comparisons</b>	4,172	62.906	<0.001	0.061 (small effect size)
<b>SPC Familiarity / Q3</b>	<b>N</b>	<b>Chi-square</b>	<b>p</b>	<b>V</b>
<b>Binary comparisons</b>	4,171	104.498	0.001	0.079 (small effect size)
<b>League tables</b>	4,171	69.179	0.307	N/A
<b>Numerical targets</b>	4,171	110.744	<0.001	0.081 (small effect size)
<b>SPC charts</b>	4,171	112.862	<0.001	0.082 (small effect size)
<b>Contextualised peer comparisons</b>	4,171	116.370	<0.001	0.084 (small effect size)

Table 5.29: Mann-Whitney U-tests ('SPC familiarity' variable against output from Questions 2 and 3)

<b>SPC Familiarity / Q2</b>	<b>N</b>	<b>U</b>	<b>z</b>	<b>p</b>	<b>r</b>	<b>Outcome</b>
<b>Binary comparisons</b>	4,172	1,644,833.5	2.057	0.040	0.03 (small effect size)	Reduced concern
<b>League tables</b>	4,172	1,470,665	-0.443	0.658	N/A	No significant difference
<b>Numerical targets</b>	4,172	1,580,281	3.254	0.001	0.05 (small effect size)	Reduced concern
<b>SPC charts</b>	4,172	1,425,911.5	-2.040	0.041	0.03 (small effect size)	Reduced concern
<b>Contextualised peer comparisons</b>	4,172	1,449,934	-1.161	0.245	N/A	No significant difference
<b>SPC Familiarity / Q3</b>	<b>N</b>	<b>U</b>	<b>z</b>	<b>p</b>	<b>r</b>	<b>Outcome</b>
<b>Binary comparisons</b>	4,171	1,590,666	3.353	0.001	0.05 (small effect size)	Reduced DUBR
<b>League tables</b>	4,171	1,497,029	0.416	0.677	N/A	No significant difference
<b>Numerical targets</b>	4,171	1,587,682.5	3.264	0.001	0.05 (small effect size)	Reduced DUBR
<b>SPC charts</b>	4,171	1,475,523.5	0.258	0.797	N/A	No significant difference
<b>Contextualised peer comparisons</b>	4,171	1,490,395.5	0.208	0.835	N/A	No significant difference

In respect of the cross tabulation, there was evidence of relationships between most of the variables tested, albeit with very small effect sizes. This suggests that familiarity with SPC may affect users' interactions with performance information, albeit with a limited effect.

With regard to the Mann-Whitney U tests, where significant differences were observed, familiarity with SPC was consistently associated with lower levels of concern and reduced DUBR. However, it must be noted that the condition was found to be non-significant in 50% of cases and where there was an impact, effect sizes were very small. Therefore, familiarity with SPC may be considered to be a moderating factor, but one which exerts limited influence.

Results for the SPC charts and contextualised peer comparisons stimuli are of particular interest, as there were no significant differences in behavioural responses between groups. This may indicate respondents reacted to these stimuli in a balanced manner due to the data display formats being relatively easy to interpret, rather than

because they possess prior knowledge. It suggests that although statistical literacy was previously shown to assist engagement with performance information, familiarity with SPC methodology does not seem to be an absolute prerequisite for accurate interpretation when reliable formats are utilised; i.e. even those unfamiliar with SPC charts tended to interpret them correctly, leading to reduced DUBR.

#### 5.6.4 Familiarity with the Author's Prior Work

A further consideration about potential influence on responses relates to familiarity with the author's prior work (see Guilfoyle, 2011; 2012; 2013; 2015; 2016). As a result of identifying familiarity with the author's work as a potential 'contaminating' factor during the pilot phase, a qualifying question was inserted into the instrument to enable analysis of the presence and extent of this. Tests were conducted, following the same format as in the previous section; the results are shown in tables 5.30 and 5.31.

*Table 5.30: Cross-tabulation ('SG familiarity' variable against output from Questions 1, 2 and 3)*

<b>SG Familiarity / Q1</b>	<b>N</b>	<b>Chi-square</b>	<b>p</b>	<b>V</b>
<b>Binary comparisons</b>	4,172	144.89	<0.001	0.108 (small effect size)
<b>League tables</b>	4,172	20.628	0.056	N/A
<b>Numerical targets</b>	4,172	77.931	<0.001	0.079 (small effect size)
<b>SPC charts</b>	4,172	17.706	0.125	N/A
<b>Contextualised peer comparisons</b>	4,172	23.519	0.024	0.043 (small effect size)
<b>SG Familiarity / Q2</b>	<b>N</b>	<b>Chi-square</b>	<b>p</b>	<b>V</b>
<b>Binary comparisons</b>	4,172	48.778	<0.001	0.054 (small effect size)
<b>League tables</b>	4,172	24.071	0.088	N/A
<b>Numerical targets</b>	4,172	30.517	0.016	0.043 (small effect size)
<b>SPC charts</b>	4,172	89.926	<0.001	0.070 (small effect size)
<b>Contextualised peer comparisons</b>	4,172	54.669	<0.001	0.057 (small effect size)
<b>SG Familiarity / Q3</b>	<b>N</b>	<b>Chi-square</b>	<b>p</b>	<b>V</b>
<b>Binary comparisons</b>	4,171	90.962	0.015	0.108 (small effect size)
<b>League tables</b>	4,171	78.706	0.102	N/A
<b>Numerical targets</b>	4,171	119.046	<0.001	0.084 (small effect size)
<b>SPC charts</b>	4,171	125.677	<0.001	0.087 (small effect size)
<b>Contextualised peer comparisons</b>	4,171	105.886	0.001	0.080 (small effect size)

Table 5.31: Mann-Whitney U-tests ('SG familiarity' variable against output from Questions 2 and 3)

<b>SG Familiarity / Q2</b>	<b>N</b>	<b>U</b>	<b>z</b>	<b>p</b>	<b>r</b>	<b>Outcome</b>
<b>Binary comparisons</b>	4,172	524,215.5	-2.977	0.003	0.05 (small effect size)	Reduced concern
<b>League tables</b>	4,172	542,062	-1.913	0.056	N/A	No significant difference
<b>Numerical targets</b>	4,172	540,269	-2.101	0.036	0.03 (small effect size)	Reduced concern
<b>SPC charts</b>	4,172	469,238.5	-6.151	<0.001	0.10 (small effect size)	Reduced concern
<b>Contextualised peer comparisons</b>	4,172	495,692.5	-4.520	<0.001	0.07 (small effect size)	Reduced concern
<b>SG Familiarity / Q3</b>	<b>N</b>	<b>U</b>	<b>z</b>	<b>p</b>	<b>r</b>	<b>Outcome</b>
<b>Binary comparisons</b>	4,171	531,027	-2.402	0.016	0.04 (small effect size)	Reduced DUBR
<b>League tables</b>	4,171	528,377.5	-2.536	0.011	0.04 (small effect size)	Reduced DUBR
<b>Numerical targets</b>	4,171	500,348.5	-3.948	<0.001	0.06 (small effect size)	Reduced DUBR
<b>SPC charts</b>	4,171	518,772	-5.263	<0.001	0.08 (small effect size)	Reduced DUBR
<b>Contextualised peer comparisons</b>	4,171	484,282.5	-4.740	<0.001	0.07 (small effect size)	Reduced DUBR

In respect of the cross tabulation, there was evidence of relationships between most of the variables tested, albeit with very small effect sizes. This suggests familiarity with the author's prior work may have a minor impact on respondents' interactions with performance information. However, as only 299 of the 4,172 respondents (7.2%) stated they were 'familiar' or 'very familiar' with the author's prior work, any impact on the output of the micro-experiments would be minimal in any case

With regard to the Mann-Whitney U tests, where significant differences were observed, familiarity with the author's prior work was consistently associated with lower levels of concern and reduced DUBR, albeit with very small effect sizes; the only exception was in respect of the league table stimulus and associated levels of concern, as this test produced non-significant results. Overall, familiarity with the author's prior work may therefore be considered to be a factor capable of affecting outcomes, albeit one which exerts limited influence.

### 5.6.5 Stratification and Inter-Block Analysis

Further analysis was conducted to assess whether significant differences exist in respect of tendencies observed between ranks, police forces, and thematic blocks.

The objective of these tests was to assess whether practices within different forces and / or factors at different levels of these organisations, may affect how respondents reacted to the stimuli. The analysis can be viewed at *Appendix 'G'*; this section contains a brief summary of results.

Stratification across all five data display types demonstrates that no individual rank stood out from the group as an outlier in any of the tests. Occasional pairwise differences were observed, however these fell within the overall range for the group and the majority of effect sizes were small.

Therefore, although these slight differences amount to minor points of interest, it is the *absence* of differences or patterns which is of greatest relevance to this study. The analysis strongly indicates the postulated mechanism operates consistently across the rank structure and that its operation is primarily affected by the act of interchanging data display formats, rather than organisational factors.

Furthermore, the analysis did not identify any significant differences between forces, suggesting consistency throughout the UK policing context. The analysis also found a sharp contrast in results between aggregated variables pertaining to binary comparisons, league tables and numerical targets, versus SPC charts and contextualised peer comparisons; this further indicates that the mechanism operates without being substantially affected by exogenous conditions.

### **5.7 Multiple Variants: Exploratory Analysis**

During the design phase, consideration was given to testing more than one variant of each stimulus, using configurations that depict contrasting trajectories (e.g. a binary comparison where this month's crime figure is lower than last month's). However, in light of Babbie's (1990) warning about the trade-off between response rates and number of survey questions, this prospect was ultimately discounted as it would have significantly extended the survey and potentially impaired its overall effectiveness.

Instead, a smaller-scale exploratory study was conducted to support the large scale research; the micro-experiments undertaken during the deployment of the survey instrument were replicated involving 50 police officers who had not accessed the original tests. These participants were shown the original stimuli in controlled conditions (i.e. under supervision and instructed to complete the tests individually).

In addition to the original stimuli, participants were also exposed to further variants of each of the data display types, depicting a range of positive, negative and neutral variance from reference points. The objective was to test whether respondents tend to take data patterns at face value regardless of the apparent direction of variance. Each of the stimuli used during this additional deployment are displayed in *Appendix 'H'*, along with relevant statistics.

The tests confirmed that respondents tend to opt for the most obvious interpretation of whatever the stimuli appear to depict, whether or not a particular format may be deemed reliable; this was found to apply whether a data pattern suggests variance or trajectories in any direction. These outcomes strongly support the proposition that data display can exert significant influence upon performance information users' interpretations, assumptions and actions.

Furthermore, the tests apparently resolve the puzzle of high incidences of 'Don't know' responses to the original league table stimulus. By deploying variants incorporating additional content alongside other stimuli comparable to the original design, it was found that respondents were much more willing to provide an interpretation when further information was available.

As consistent patterns of 'Don't know' responses were generated by the additional league table stimuli, this provides reassurance that it was likely a design characteristic of the original stimulus that was responsible for the anomaly. The results produced by interaction with the additional variants demonstrated that users do indeed tend to respond as anticipated, which is consistent with the behaviour observed in respect of all other stimuli, in both the large-scale and exploratory tests.

Overall, the exploratory analysis does not produce absolutely conclusive findings due to the relatively small sample sizes involved; it does however, provide useful indications about tendencies, as well as support for the prospect that data display is a primary antecedent to performance information users' cognitive and behavioural responses. Indeed, the consistency between the original micro-experiments and supplementary tests bodes well for test-retest reliability and provides further grounds for making claims about the presence of a mechanism.

## **5.8 Quantitative Analysis: Summary and Key Findings**

The primary objective of the quantitative analysis was to ascertain if statistical testing identified any strong or recurrent patterns that may indicate the presence of a mechanism responsible for producing the phenomenon of interest. Specifically, it aimed to establish if the use of different data display formats affected the operation of such a mechanism, despite other structural influences remaining constant.

The results certainly appear to confirm the presence of a mechanism operating at the interface between performance information and its users; the output also strongly complements the findings generated by the qualitative analysis. Overall, this provides valuable insight into the cognitive and behavioural responses that tend to occur when certain data display formats are used. A brief summary of the main findings relating to each stimulus is presented below.

### **5.8.1 Binary Comparisons**

With regard to the binary comparison stimulus, the analysis found:

1. An overwhelming proportion of respondents (89.6%) misinterpreted the stimulus to mean crime was increasing.
2. Those who believed crime was increasing tended to experience higher levels of concern.
3. Those who experienced higher levels of concern tended to enact disproportionate / unwarranted behavioural responses.
4. Regression analysis confirms that patterns observed in the data at Questions 1 and 2 act as a precursor for the extent of adverse behavioural tendencies at Question 3.
5. Potential moderating conditions (e.g. organisational influence / familiarity with SPC) exerted minimal or no influence on responses.

As with all the tests, these outcomes occurred with high levels of statistical significance and notable effect sizes, strongly indicating that data display exerts a substantial influence on performance information users' interpretation, assumptions and consequent behaviour.

### **5.8.2 League Tables**

In respect of the league table stimulus, there was a split between those who believed performance was poor and those who selected the ‘Don’t know’ option. In summary, the analysis found:

1. 48.9% of respondents chose ‘Don’t know’ at Question 1, whilst 45.1% concluded performance was poor; this divergence produced notably contrasting results in subsequent tests.
2. Those who believed performance was poor tended to experience higher levels of concern; conversely, those who selected ‘Don’t know’ at Question 1 did not tend to instigate DUBR.

For those who believed performance was poor, conditions 3, 4 and 5 of the original pattern were then repeated, resulting in high levels of dysfunction. This pattern was also evident in the results of the supplementary experiments; furthermore, where additional variants incorporated further content, respondents consistently interpreted those stimuli in the expected direction.

### **5.8.3 Numerical Targets**

With regard to the numerical target stimulus, the analysis found:

1. 78.4% of respondents’ interpreted variance from the target to mean performance was poor.
2. Those who believed performance was poor tended to experience higher levels of concern.

Again, conditions 3, 4 and 5 of the original pattern followed these initial steps. This suggests that where assumptions are made about performance as a result of engaging with these three data display formats, they consistently act as a catalyst for undue concern and adverse behavioural tendencies.

#### **5.8.4 SPC Charts**

Whilst the above sequence was common to binary comparisons, league tables and numerical targets, in respect of SPC charts and contextualised peer comparisons it was largely reversed. For the SPC stimulus, the analysis found:

1. An overwhelming proportion of respondents (82.3%) accurately interpreted the stimulus, correctly ascertaining crime was stable.
2. Those who identified crime was stable tended to experience lower levels of concern.
3. Those who experienced lower levels of concern tended to be less likely to initiate DUBR.
4. Regression analysis confirms that patterns observed in the data at Questions 1 and 2 act as a precursor for the extent of adverse behavioural tendencies at Question 3.
5. Potential moderating conditions exerted minimal or no influence on responses (as previously).

This suggests that use of this format promotes increased likelihood of accurate interpretation and consequent reduction in adverse behavioural tendencies. It is particularly notable that prior knowledge of SPC does not appear to be a prerequisite for these observed outcomes, as the majority of respondents were unfamiliar with it.

#### **5.8.5 Contextualised Peer Comparisons**

Similarly, in respect of contextualised peer comparisons, it was found:

1. Two thirds of respondents (66.9%) accurately interpreted the stimulus, to ascertain performance was acceptable.
2. Those who identified performance was acceptable tended to experience lower levels of concern.

Conditions 3, 4 and 5 then followed, resulting in reduced levels of adverse behavioural tendencies. Again, it is notable that although this stimulus was designed around SPC principles, these outcomes arose despite the vast majority of

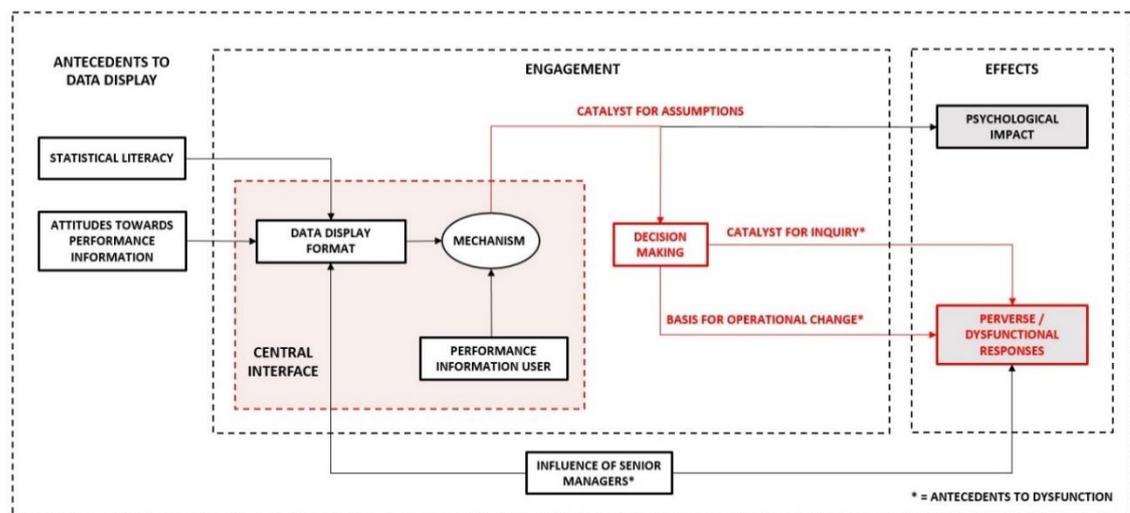
respondents being unfamiliar with them. Crucially, the sharply contrasting results produced by these latter tests are derived under the same conditions and from the same respondents who previously demonstrated a propensity to enact DUBR.

### 5.9 Quantitative Analysis: Overall Conclusions

The results of this analysis complement the findings of the qualitative analysis and strongly support the prospect of there being a mechanism that operates at the interface between performance information and its users; the output of the micro-experiments also confirms that the types of conduct identified as being antecedents to dysfunction are precisely the forms of behaviour that users tend to enact when confronted with particular data display formats.

Returning briefly to the theoretical model presented in Chapter Four, a variant is produced below, highlighting the areas tested during the micro-experiments. As the interaction of entities within the ‘Central Interface’ likely accounts for activation of the proposed mechanism, this was the initial focus of statistical testing, followed by assessments of its influence upon subsequent events.

Figure 5.32: Theoretical Process Model – Statistical Testing of Variables



The components of the model depicted in red pertain to the variables and themes subject to statistical testing; the outcomes of this analysis confirm many of the findings of the qualitative analysis and strongly corroborate the likelihood that data display exerts a direct influence on the polarity of the proposed mechanism’s operation, as well as ultimately upon the phenomenon of interest.

From a critical realist perspective, it is therefore proposed that where reference-dependent data display formats are used, their innate powers and liabilities affect the interaction between performance information and its users, causing the conceived mechanism to trigger a sequence of events, whereby users tend to misinterpret the data, formulate unfounded assumptions and exhibit adverse behavioural tendencies; these tendencies manifest themselves in behaviours that act as antecedents to dysfunction, as well as dysfunction itself.

Conversely, where statistically robust data display formats are utilised, their contrasting powers and liabilities influence the operation of this mechanism, causing it to produce significantly different events (i.e. accurate interpretation, moderated concern and reduced levels of adverse behavioural tendencies). Furthermore, as test participants were engaging with experiments rather than acting in environments whilst exposed to organisational influences, the findings suggest data display is a pivotal factor affecting the polarity of the mechanism's operation.

These proposals are further supported by the unexpected results recorded during the league tables analysis; as a large proportion of respondents did not provide an interpretation for this stimulus, the subsequent discovery that these individuals exhibited strong tendencies *not* to enact DUBR indicates it is highly likely that the mechanism is only activated if the stated entities interact. Critical realists would argue this reflects a situation where performance information is *present but not used*; its powers and liabilities remain unexercised and the mechanism is not activated.

Overall, the quantitative analysis strongly indicates the use of some data display formats is significantly more likely to trigger dysfunction than others; it also helps explain the underlying processes by which this occurs. The following chapter will explore these implications in greater depth, integrating the findings of the qualitative and quantitative analysis, and examining the operation of the theoretical model in the context of broader research and theory.

## Chapter Six

### Discussion

#### 6.1 Introduction

This chapter discusses the implications of the study's findings in the light of wider research, literature and theory.<sup>38</sup> In doing so, the theoretical process model is explored in detail, assessing relationships between components in greater depth and with regard to relevant literature. The discussion also introduces and explores a distinction between *antecedents to data display* and *core concepts* of the study; this aims to fully contextualise the operation of the model and clearly explicate the role of data display as a catalyst for dysfunction.

The discussion examines individual elements of the model and their roles. It assesses types of, and antecedents to, dysfunction, drawing parallels between the literature and themes that emerged during qualitative analysis; it also refines the typology of behavioural dysfunction by proposing the notion of *intermediate dysfunction*. The narrative considers how the study informs, and is informed by, both the performance management and data display literature, establishing firm links to prior research and integrating the qualitative and quantitative analysis.

#### 6.2 Theoretical Process Model

Central to this research is the theoretical model first presented in Chapter Four (p.170); this model depicts relationships between themes identified during thematic analysis and illustrates the operation of a mechanism responsible for triggering a sequence of events that produces the phenomenon of interest. This sequence begins with the interaction between performance information and its users, culminating in behavioural dysfunction when certain data display formats are used. The model also highlights other factors, either on the periphery of this sequence, or which otherwise bear influence on the choice of data display format or phenomenon of interest.

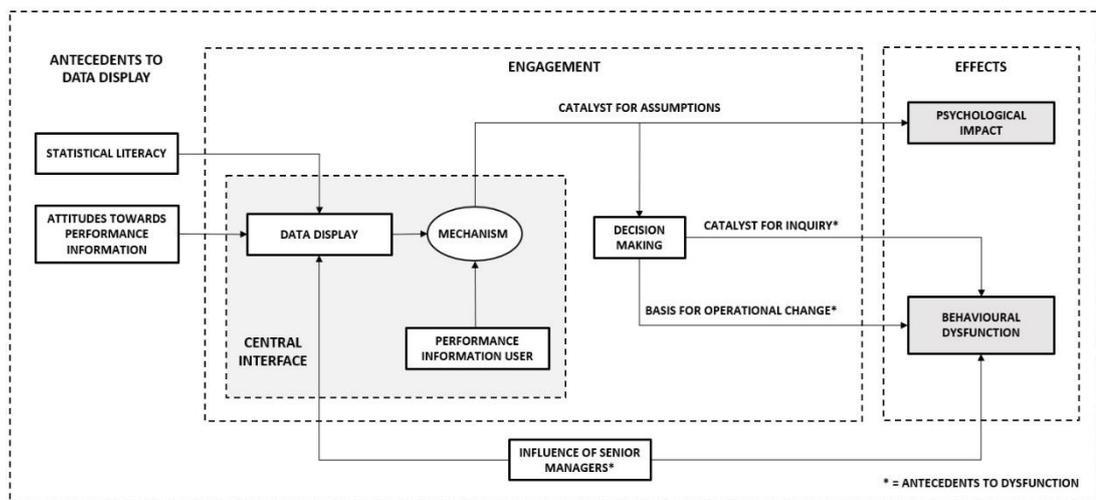
---

<sup>38</sup> In doing so, the discussion introduces new literature in places when examining constructs and relationships from within the theoretical model, in order to provide possible explanation for observed phenomena (e.g. in respect of the content exploring statistical literacy on pages 227-230). This is known as 'inference to the best explanation' (see Ketokivi and Mantere, 2010) and is part of the abductive process.

The model was subsequently revisited during Chapter Five, which highlighted significant relationships between variables tested during the micro-experiments; this confirmed strong links between particular themes and also reinforced the prospect that data display exerts considerable influence on the operation of the mechanism. Essentially, in overlaying findings of both the qualitative and quantitative analysis, this further demonstrated that a pivotal factor influencing cognitive and behavioural processes is the data display format itself.

Consequently, it is suggested the model is highly robust and possesses considerable explanatory utility. However, aside from exposing direct links between data display and behavioural dysfunction, the model also begins to open up wider considerations about how its components relate to wider theory and research. This discussion will therefore explore *how and why* elements of the model seem to operate in the way they do. The model is presented again at Figure 6.1.

Figure 6.1: Theoretical Process Model



The model comprises a range of components,<sup>39</sup> some of which are integral to the pathway from data display to the phenomenon of interest; these sit alongside others that identify factors and relationships pertinent to the overall context, but which do not necessarily sit within the linear sequence of events triggered by the activation of a mechanism.

<sup>39</sup> In this final iteration of the model, the original working labels 'Data Display Format' and 'Perverse / Dysfunctional Responses' have been simplified to 'Data Display' and 'Dysfunctional Behaviour' respectively.

Prior explication of the model's operation has drawn predominantly upon the empirical data underpinning it, to explain how the components interact and influence each other. Here, we look beyond this data to explore underlying theoretical considerations that may help explain relationships between themes in greater depth, as well as identify potential areas for further examination. Whilst some of the themes integral to the sequence of events depicted within the model are also exclusive to it (e.g. 'Catalyst for Inquiry') others overlap with broader domains and constructs (e.g. 'Statistical Literacy'); it is the latter that chiefly inform the bulk of the discussion, which aims to shed light on the following overarching considerations:

1. How and why themes affect each other, in respect of the specific directional pairings depicted within the model.
2. Whether, how and why each theme influences:
  - i) Behavioural dysfunction in performance management systems.
  - ii) Interaction with, or the use of, performance information.

The forthcoming narrative therefore explores components of the model in turn, commencing with those identified as acting as antecedents to data display.

### **6.2.1 Antecedents to Data Display**

Thematic analysis identified three themes that act as major antecedents to data display, being found to affect the way in which particular data display formats are selected and used in police performance management systems; they are 'Statistical Literacy', 'Influence of Senior Managers' and 'Attitudes to Performance Information'. Each is discussed below.

#### **6.2.1.1 Statistical Literacy**

Statistical literacy is a complex construct (Watson and Callingham, 2003). It can be defined as *the ability to interpret and critically evaluate statistical information* (Wallman, 1993; Gal, 2000). It is, however, a largely neglected skill that receives little explicit attention (Gal, 1997; 2002) and this has led for calls for greater emphasis to be placed upon it (Sharma, 2017).

Much of the research into statistical literacy is based within the education setting, where a raft of studies highlight the benefits of possessing a basic awareness of statistical concepts (Biggs and Collis, 1982; Watson and Callingham, 2003; Watson, 2003; 2006; Clarke *et al*, 2006; Franklin and Garfield, 2006; English, 2010; Callingham and Watson, 2017). Commentators on this topic assert specific statistical considerations should inform data-driven decisions, such as having a knowledge of variation, standard deviation, significance testing and inferential reasoning (Scheaffer *et al*, 1998; Watson, 2011).

Such lessons translate well into the policing context, as statistical literacy has been shown to act as a prominent antecedent to the types of data display chosen to present police performance information. Therefore, managers with sufficient statistical awareness should be well-placed to select robust formats and reject unreliable ones. Furthermore, although statistical literacy does not fit within the linear chain of events responsible for producing the phenomena of interest, analysis of the empirical data strongly suggests it directly affects engagement with performance information.

Centrally, whilst the study found reduced levels of dysfunction were associated with reliable formats, it also identified statistical literacy as being a primary factor in preventing dysfunction from occurring when users are presented with unreliable formats. This strongly indicates statistical literacy aids recognition of potentially misleading formats and encourages users to disengage, thereby halting the sequence of events that leads to dysfunction.

However, prior research exploring statistical literacy does not establish any relationship between it and behavioural dysfunction; furthermore, there is very little that specifically examines any potential links between statistical literacy and performance information use. Nevertheless, studies within this field provide insights into underlying considerations that help explain how it can affect data interpretation and ultimately influence decision-making.

Of particular relevance, research into statistical literacy finds that it can assist decision-making when individuals are presented with numeric data. For example, where performance information and risk management can be integrated (see Bourne and Mura, 2018) studies indicate statistical literacy promotes effective decisions about risks or hazards (AEC, 1991; Watson, 2011); separate research also found

perceptions of risk arising due to individuals' interpretations of numeric data can vary considerably, depending on their levels of statistical literacy (Sandman, 1993).

To counter such disparity and promote sound judgment when faced with statistical information, Watson (1997) proposes a three-tier hierarchy designed to assist in effectively assessing risk, which is paraphrased as follows:

1. Possessing knowledge of basic statistical terms;
2. Interpreting such terms in applied contexts; and;
3. The ability to question unrealistic claims made by others.

It is suggested that such an approach could well be adopted beyond the realms of risk management and applies directly to the way in which performance information can (and arguably, *should*) be interacted with. The ability to apply pre-existing statistical knowledge within relevant contexts, alongside confidence to challenge unwarranted assumptions about data, certainly appears relevant to the interpretation and use of performance information; in theory, it should act as a basis for robust assessment, which in turn would aid decision-making.

This prospect was confirmed by analysis of the study's empirical data; respondents consistently cited benefits where managers possessed satisfactory levels of statistical awareness, thereby responding to performance information in a measured and proportionate fashion. A typical observation was that the use of SPC by suitably-trained managers, "*Prevents knee jerk to normal trends*" (SPC1522); another respondent observed, "*I understand this sort of data because I have had training, but many don't. It is very useful if used correctly*" (SPC3172).

In contrast, respondents reported negative effects when managers lacked statistical literacy, citing "*...a universal lack of understanding of the normal range and variation*" (SPC3289), as a result of "*...no training in performance management*" (SPC3666). Fundamentally, one respondent asserts SPC is "*...only as good as the person reading and understanding it*" (SPC2853). Consequently, whilst the importance of statistical literacy cannot be understated in general, its principles are particularly relevant to police performance management, as they directly influence the sequence of events at the heart of the theoretical model.

The study also established that some individuals prefer simplistic formats on account of their appearance being more visually pleasing than formats such as SPC; this condition was also occasionally present within the healthcare research, where users expressed a preference for simplistic formats due to their perceived heuristic value (Hawley *et al*, 2008; Faber *et al*, 2009; Hildon *et al*, 2012).

However, this study also found where those expressing such preferences happened to be senior managers, this perpetuated the use of simplistic formats and constrained the use of comparatively visually complex (albeit more reliable) ones. Ultimately, where senior managers exert influence over their subordinates *and* performance management systems, yet lack sufficient levels of statistical literacy, this creates a fertile ground for behavioural dysfunction to thrive.

In summary, whilst there is no prior research connecting (even indirectly) statistical literacy to behavioural dysfunction, it is clear that it bears direct relevance to data display and also influences decision-making. Furthermore, although there is a paucity of literature that specifically considers the impact of statistical literacy on performance information use, the underlying concepts outlined may partially explain why it acts as a prominent antecedent to data display; there may also be an opportunity to further explore its impact on the cognitive processes activated when users engage with performance information.

#### **6.2.1.2 Influence of Senior Managers**

Another condition impacting upon data display and performance information use is the influence of senior managers. Although this condition does not form a step in the sequence from data display to the phenomenon of interest, thematic analysis confirmed it can behave as a factor capable of exacerbating this chain of events; indeed, it is explicitly identified within the model as a major antecedent to dysfunction. However, although prior research demonstrates senior managers affect behaviour and organisational conditions in a variety of ways, there is limited material directly covering its relationship with dysfunction; furthermore, there is nothing that considers how choices involving data display may be affected.

Prior studies establish that senior managers influence the systems, structures and processes of organisations, but also act on a more symbolic level (Gordon, 1991) and

this can affect both the perceptions and behaviour of lower-level managers (Vancil, 1978; Baird *et al.*, 1981). The reach of senior managers' influence has been found to impact upon a range of domains, including organisational change (Maurer, 2014), job satisfaction (Moyle, 1998), employee wellbeing (Robertson and Barling, 2014), integrity of middle managers (Van Niekerk and May, 2012), trust (Albrecht and Travaglione, 2003) the willingness of employees to act entrepreneurially (Brundin *et al.*, 2008), attitudes towards safety (Flin, 2003), organisational learning (Waddell and Pio, 2015) and absorptive capacity (Sun and Anderson, 2012).

This confirms that senior managers can exert considerable power over the operation of their organisations and behaviour of subordinates. Of primary relevance to this study are findings they can specifically influence performance management practices (Otley, 2003) and measured performance (Davies *et al.*, 2007); furthermore, effects tend to be magnified where reward schemes are present (Kessler and Purcell, 1992; Suff *et al.*, 2008; Le Meunier-FitzHugh *et al.*, 2011).

Of note, it is known that 'irrational' behaviour by managers can trigger dysfunction (Kerr, 1975; Pope and Burnes, 2013), although research also confirms they can intervene against such dysfunction (Kahn, 2012). Additionally, the literature review identified certain managerial behaviours as being associated with unintended consequences; these included managers insisting on answers for perceived deficiencies (Eterno and Silverman, 2012), as well as demanding improvements or directing unnecessary operational activities (Home Office, 2015).

This research establishes stronger links between such antecedent behaviour and dysfunction; it clearly exposes the way in which conduct by managers contributes towards dysfunction. For example, respondents spoke of domineering behaviour generating "*...pressure that breeds manipulation and fiddling of figures*" (LT2305), whilst others described an "*...authoritarian, oppressive work environment...*" (NT2532), coupled with "*...a bullying culture within the organisation*" (BC2543). Furthermore, the experimental component of the study tested the relationships between particular performance information formats and such antecedents to dysfunction, finding strong and recurrent links.

Similarly, the study uncovered cases where performance information was used to "*...name and shame...*" (LT906), along with examples of gaming arising as a direct

consequence of officers being “...*beaten up...*” (LT3182) by bosses for alleged poor performance. It also confirmed the presence of a “...*focus on narrow outcomes...*” (NT355) in some quarters, similar to that documented in the literature. Others spoke of norms where “...*quantity rather than quality is the rule of thumb...*” (LT3764).

Overall, the influence of senior managers is identified as a far-reaching condition capable of affecting behaviour, as well as broader organisational factors. Although there is a lack of research that examines its relationship with data display, by taking the findings of prior research in conjunction with the outcomes of this study, it appears that the influence of senior managers is a potent antecedent for establishing organisational norms that affect how performance information is selected and used. This, in turn, acts as a lever for behavioural change and ultimately helps to reinforce the conditions for the sequence of events depicted within the model.

### **6.2.1.3 Attitudes towards Performance Information**

The third antecedent to data display identified during thematic analysis was the theme ‘Attitudes to Performance Information’. This theme highlighted prejudices and even animosity towards police performance information; it also drew attention to difficulties associated with measuring complex policing activity and provided insights into some of the reasons why simplistic forms of data display are prevalent within UK police forces.

Respondents consistently bemoaned what they perceived as some managers’ fixation on crime figures as an arbiter of performance; commentary on this topic stacked up to provide a fascinating insight into one of the factors responsible for disquiet about the relevance of many performance measures. Certainly, for frontline officers, the use of crime data in this manner tended to generate negativity towards performance information (and even statistics as a whole), regardless of the formats used.

Furthermore, parallels were found with extant literature in respect of the suitability of using crime data as a performance measure (see Bayley, 1994; Coleman and Moynihan, 1996), along with difficulties in quantifying police performance and the measurement of outcomes (Shane, 2010). For example, respondents cited activities such as “...*spending an extra hour with a vulnerable victim of crime to offer reassurance...*” (NT2664), with many arguing such intangible contributions are

overlooked due to an emphasis on measuring simplistic outputs. Many were also emphatic about the need for effective performance measurement, whilst being highly critical of prevalent forms of data presentation.

This resonates with discussions on the definition and measurement of performance in complex systems (Pollitt, 1999; Caers *et al*, 2006; Kelman and Friedman, 2009; Moynihan *et al*, 2009), as well as the problematic issue of quantifying outcomes (Behn and Kant 1999; Mackenzie and Hamilton-Smith, 2010; Lowe, 2013); these considerations were reflected within the qualitative data, being consistently identified as major antecedents to officers' attitudes towards performance information.

Nevertheless, despite general negativity, the study found a strong sense of vocation amongst officers; although there was an occasional tendency to conceive performance information as a homogenous entity, respondents exhibited a strong desire to serve the public. This is consistent with the notion of Public Service Motivation (PSM), as discussed by Moynihan and Pandey (2010) and further explored by Kroll (2014). Although not a focal point of this study, this finding ties the research further to this literature and could be an avenue for further exploration.

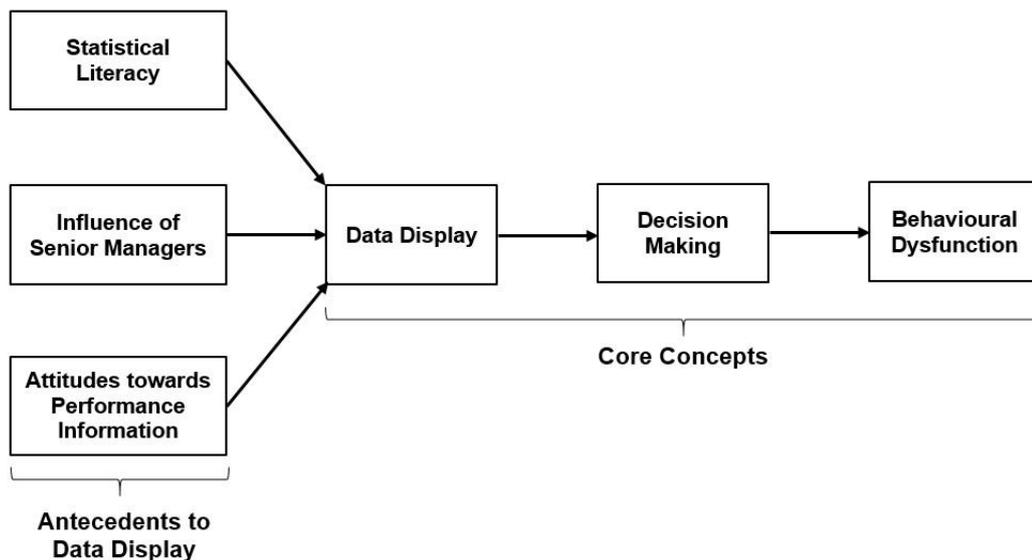
However, despite the rich narrative produced by this study, when it comes to prior research, there seems to be nothing to directly connect attitudes towards performance information with either data display *or* behavioural dysfunction. In contrast, the type of negativity towards statistics observed within the free text data is generally echoed elsewhere and it is suggested this could be partly attributed to 'statistics anxiety' (see Zeidner, 1991; Perney and Ravid, 1991; Gal and Ginsburg, 1994; Onweuegbuzie *et al*, 1997; Onweuegbuzie, 2000). This condition manifests itself in a strong dislike (or even fear) of statistics and it may therefore be instructive to explore its impact within policing, to assess how it may affect attitudes towards performance information.

Overall however, when considered alongside the themes 'Statistical Literacy' and 'Influence of Senior Mangers', it seems this theme acts as a less powerful influence on data display compared to the other two major antecedents discussed to this point; essentially, the theme 'Attitudes towards Performance Information' provides additional insights, context and nuances that help shape the organisational climate, rather than actively accelerating the sequence of events at the heart of the model.

### 6.2.2. Core Concepts

These three antecedents to data display act as levers that shape the way in which performance information is selected and used in police performance management systems. As a result of their combined influence, the character of performance management and types of data display used take on a particular disposition. Their influence also ultimately affects the sequence of events that follows once users engage with performance information. This sequence revolves around the key themes of ‘Data Display’, ‘Decision Making’ and ‘Behavioural Dysfunction’, which represent *core concepts* in this study. The links between antecedents to data display and core concepts are depicted in Figure 6.2.

Figure 6.2: Antecedents to Data Display and Core Concepts



The remainder of this discussion focuses upon these core concepts.

#### 6.2.2.1 Data Display

The ‘Data Display’ theme is central to the operation of the model. Where certain data display formats are present during the engagement phase, the resultant cognitive processes consistently trigger responses that act as antecedents to dysfunction, as well as reactions that may be considered innately dysfunctional. Therefore, this portion of the discussion examines the influence of data display in the light of relevant literature, as well as associated theories and concepts.

The data display literature is extensive (see Tufte, 1990; 1997; 2001; 2013; Card *et al*, 1999; Chen, 2006; Few, 2009; Lindquist, 2011; Ward, 2015, Isett and Hicks, 2018). It establishes the way in which data are presented can aid or impair communication of information (Vessey, 1991; Speier, 2006; Olsen, 2013a; 2015; Hirsch *et al*, 2015; Huber, 2016), affect users' comprehension (Gerteis *et al*, 2007; Kurtzman and Greene, 2016) and influence decision-making (Peters *et al*, 2007; NHS, 2018). Studies have also shown some formats promote accurate interpretation (Marshall *et al*, 2004; Sherlaw-Johnson and Bardsley, 2016; Schmidtke *et al*, 2017a), whilst others can be misleading (Anhøj and Blok Hellesøe, 2016).

Whilst the data display literature encompasses a wide range of loci, this study specifically explores the cognitive and decision-making processes that occur during *engagement with performance information*. Therefore, it most closely speaks to a comparatively recent field of data display research, primarily focused on the use of performance information in healthcare settings (see, for example: Uhrig *et al*, 2006; Peters *et al*, 2007; Damman *et al*, 2011; Hildon *et al*, 2012; Elbel *et al*, 2014; Anhøj and Blok Hellesøe, 2016; Kurtzman and Greene, 2016; Mountford and Wakefield, 2016; Neuburger *et al*, 2017; Schmidtke *et al*, 2017a; 2017b).

This study's findings confirm elements of the healthcare research; in particular, it consistently identified that performance information users are influenced by the visual appearance of data. This was demonstrated in both the qualitative and quantitative elements of the study; in particular, the psychometric micro-experiments conducted during the latter stage of the research found that users interpreted content radically differently, depending on the format used. This was despite the fact that the stimuli contained information about comparable performance dimensions (e.g. crime rates) and participants were exposed to each of the stimuli in comparable conditions.

It was also established that users' initial perceptions of performance were directly responsible for the assumptions they drew following engagement with data, being demonstrated by the ways in which they consistently reacted to perceived poor performance. This finding strengthens prior research and leads us to touch on the 'output' of the central interface, namely the theme, 'Catalyst for Assumptions' – this construct encompasses the stage immediately following activation of the mechanism and is directly prior to the decision-making step in the sequence depicted within the

model. It may be conceived as the point whereby the information absorbed during engagement is processed and users formulate their interpretation of the data.

The nature of such assumptions and the role that data display plays in shaping them were explicitly tested during the experimental phase of the study; this identified a sharp contrast between the effects of engagement with reference-dependent formats (see Festinger, 1954; Albert, 1977; Tversky and Kahneman, 1986; 1991; Olsen, 2013b; Charbonneau and Van Ryzin, 2015), versus interaction with statistically robust forms of data presentation, such as SPC charts (see Shewhart, 1939; Wheeler, 1998; Nielsen, 2015). It was found that interaction with reference-dependent formats consistently led to unwarranted assumptions; conversely, when presented with reliable formats, participants tended to draw accurate conclusions and respond significantly differently to how they reacted to the more simplistic stimuli.

Furthermore, where such adverse assumptions arose, they regularly impacted upon the theme ‘Psychological Impact’, which tended to manifest itself as damage to morale and employee wellbeing; this was often attributed to negative inferences being drawn about performance by managers. The wider literature demonstrates that this condition can arise as a consequence of managers’ attitudes and behaviour towards subordinates, resulting in psychological distress, such as anxiety, depression and emotional exhaustion (Brewer, 1996; Tepper, 2000; Kelloway and Barling, 2010; Rafferty *et al*, 2010; Skakon *et al*, 2010; Restubog *et al*, 2011).

Indeed, the study confirms the presence of adverse psychological effects previously associated with target-driven performance management. Prominent themes, such as degradation of trust and autonomy (Taylor-Gooby, 2009), loss of motivation (Deci *et al*, 1999; Moynihan, 2010; Jacobsen *et al*, 2014), feelings of disempowerment (Fitzgerald *et al*, 2002; Moynihan and Soss, 2014) and an increase in stress levels (Locke and Latham, 2009) were evident in the free text responses.

However, a particularly strong and recurrent theme amongst the comments was the overwhelming “...negative effect on morale...” (LT2346); this too reflects aspects of the literature (O’Neill, 2002; Jackson, 2005). However, the key difference between prior studies and this research is that a portion of such psychological damage observed in performance management systems can now be directly attributed to the use of data display.

Additionally, the study demonstrates how *negativity bias* (James and John, 2007; Olsen, 2013c) and *loss aversion* (Tversky and Kahneman, 1991; Quattrone and Tversky, 1988; Kahneman, 1992) manifest themselves in the form of elevated levels of concern where respondents envisaged poor performance as a result of interaction with the experimental stimuli. This may help further inform the findings of prior research that examines negativity bias in respect of performance information use by politicians (Moynihan, 2016b; Nielsen and Moynihan, 2017a) and citizens (Larsen and Olsen, 2020), not least as political-administrative systems are dominated by numerical performance measurement (Olsen, 2018).

Overall, this leads to the assertion that once users engage with a particular format, they begin to formulate assumptions about what is presented and this triggers a sequence of events directly shaped by the nature of the data display format utilised; effectively, users respond in different ways when interacting with different formats and this affects decision-making. The model demonstrates how initial interpretations are impaired when superficial formats are used and how subsequent reactions are then characterised by remarkably consistent tendencies; this confirms that data display can significantly aid or impair interpretation of performance information and thereby colour assumptions about performance.

#### **6.2.2.2. Decision Making**

Whilst the themes that act as antecedents to data display provide insights into conditions affecting performance information use, the decision-making processes that arise following engagement with performance information (i.e. the ‘thinking’ stage) represent the bridge that completes the link to observed behavioural outcomes (i.e. the ‘action’ stage).

It has been argued that one of the primary functions of performance information use is to enable effective decision-making (Ammons and Rivenbark, 2008; Moynihan, 2008; Lavertu and Moynihan, 2012); it also remains largely undisputed that performance management systems exert considerable influence upon individuals’ decisions (Dumond, 1994). This study bears out these assertions, establishing that data display acts as a pivotal lever affecting such decisions, due to the way in which it influences the operation and polarity of the proposed mechanism.

Therefore it may be conceived that activation of the mechanism produces the ‘raw materials’ required for decision-making; these take the form of users’ assumptions about the data. Where unreliable data display formats are utilised, such assumptions lead to decisions that produce counterproductive reactions; in the case of this study, such reactions manifest themselves in the themes ‘Catalyst for Inquiry’ and ‘Basis for Operational Change’, which in turn, act as antecedents to dysfunction. Typical responses usually involve managers demanding improvements in respect of perceived deficiencies and are depicted as outgoing arrows from the ‘Decision Making’ block, thereby forming the final steps prior to the phenomenon of interest.

Although such reactions may be classed as the first stage of behavioural dysfunction, an overwhelming preponderance of data from the study indicates the sequence rarely stops at this point, with users initiating disproportionate or unnecessary operational responses to potentially non-existent problems. Whilst such inappropriate responses are dysfunctional in their own right, the experiences of respondents demonstrate that they act as antecedents to more serious forms of dysfunction.

Therefore, the model not only illustrates how dysfunction can be traced to decision-making, but also how those decisions are directly influenced by the assumptions formulated as a consequence of engagement with defective forms of data display. These findings are further strengthened when one examines what occurs where reliable data display formats are used; the cognitive processes remain the same, yet decision-making is substantially altered, as users either identify there is no major deficiency to be addressed, or are able to pinpoint areas that may genuinely require attention.

One particular aspect of the decision-making process exposed by this research was the reluctance of some respondents to ‘do nothing’, even when faced with experimental stimuli that suggested a change of tactics or other intervention was not necessary. This opens up theoretical considerations about tendencies to act (or otherwise) in such circumstances. For example, Olsen (2017) conducted experiments that demonstrated action tends to be more positively evaluated than inaction when responding to perceived poor performance.

This reflects studies reporting on *action bias* (see Patt and Zeckhauser, 2000; Connolly and Reb, 2003), which highlight a general disposition towards taking

action. This can even be the norm where such action may not have been necessary and reflects other studies indicating this remains the case regardless of the eventual outcome (Zeelenberg *et al*, 2002; Bar-Eli *et al*, 2007).

This notion runs contrary to the concept of *omission bias* (see Ritov and Baron, 1990; Baron and Ritov, 1994; 2004), which asserts there is usually a preference for adverse outcomes caused by omissions rather than equivalent or lesser harm caused by positive acts. However, the more recent research finds that polarity is substantially reversed (i.e. towards action) if the choice between action or inaction is *in direct response to a perceived problem*, even “...where the action may not have any real effect or could even worsen the situation” (Olsen, 2017, p.1361).

This may help explain one facet of the decision-making process when performance information users engage with numeric data; either they are unable to interpret otherwise benign patterns of data effectively (due to inadequate levels of statistical literacy), or their assumptions lead them to believe there is a deficiency requiring intervention (due to misplaced faith in an unreliable data display format). In each case, users default to acting, even where it would be inappropriate to do so.

Overall, it is clear the ‘upstream’ activity within the model sets the direction for the subsequent decisions and that the nature of associated behavioural phenomena is likely to be affected by the choices made. Therefore, although some prior research has explored links between data display and decision making, this study is unique in the sense that it not only strengthens these links, but positions decision-making as the lynchpin that connects data display to behavioural dysfunction.

### **6.2.2.3 Behavioural Dysfunction**

This theme encapsulates the phenomenon of interest and represents the culmination of the sequence of events depicted within the model. It is clear from prior research there is no single cause of dysfunction in performance management systems and this is certainly reflected in the findings of this study. The extant literature indicates that a multitude of environmental and behavioural factors contribute towards a variety of adverse consequences; these range from relatively minor aberrations to instances of seriously unethical conduct.

The literature review explored potential antecedents to dysfunction; for instance, pressure from managers (Goodhart, 1975; Neyroud and Disley, 2007), inappropriate application of otherwise potentially useful measures (Locke and Latham, 1990; 1991; 2006; Barsky, 2008), the use of sanctions or penalties (Bevan and Hood, 2006; Rothstein 2008; Bevan and Hamblin (2009) and a focus on quantity over quality (Northcraft *et al*, 1994; Chatterton, 2008; Cumming, 2014; Murray, 2014).

Some of the factors identified in the literature review were also found during this study; indeed many are featured as first order components of the themes used to construct the theoretical model. Commentary on some others (such as the use of conflicting measures), tended to be comparatively rare and therefore insufficiently frequent to contribute towards the generation of standalone themes. Other known contributory factors, such as aggressive audit and inspection regimes, did not feature in respondents' commentary at all.

However, there is a good degree of consistency between dysfunction recounted in the literature and that identified by this study. In other words, the types of dysfunction observed in UK policing closely match those seen in performance management regimes elsewhere; furthermore, they were found to be closely connected to the use of league tables and numerical targets.

Specifically, the literature identifies a wide range of perverse behaviours, dysfunction and unintended consequences from across the public sector; these included falsification of records (Francis, 2013a; Shorrock and Licu, 2013) underreporting measures (Eterno and Silverman, 2012; Crockett, 2013; MOPAC, 2013; PASC, 2013b) and other forms of gaming designed to make performance appear better than it actually is (Bevan and Hood, 2006; Rothstein, 2008; Home Office, 2015).

These patterns of behaviour were found to be closely replicated within UK policing, as comprehensively demonstrated by the experiences of respondents. Common types of dysfunction were extensively documented within the free text responses used to construct the data structures; examples ranged from “...*unethical practices and even falsifying of figures...*” (LT647) to “...*unhealthy competition...*” (NT3562) and “...*gaming to improve scores...*” (LT2165).

However, although this study confirms potential contributory factors observed elsewhere are likely to apply in policing settings, it also identified data display as an additional condition of influence. Furthermore, whilst confirming there exists a strong association between dysfunction and the use of numerical targets and league tables, it also made the novel discovery that the use of binary comparisons similarly appears to initiate behavioural dysfunction.

As the use of binary comparisons has not previously been identified as being a potential trigger for dysfunction, this finding begins to open up the prospect that in addition to the conditions already known to adversely influence behaviour, there may be something about the visual appearance of particular performance information formats that causes users to unwittingly take the first steps along a path leading to dysfunction. This possibility is strengthened by output from the experimental phase of the study, which confirmed there were strong parallels in respect of the types of behaviour triggered by the use of certain formats.

Typically, in both the qualitative and quantitative elements of this study, although numerical targets and league tables were found to exhibit a strong association with the phenomenon of interest, it was consistently demonstrated that the binary comparison format also appeared to exert a similar influence. Respondents reported a range of reactions consistent with responses to numerical targets and league tables, such as the format being used to “...*distort the figures in favour of the author...*” (BC141), “...*focus activity in a narrow range...*” (BC3761) and cause teams to “...*compete against each other...*” (BC3667).

Furthermore, in addition to the cases of gaming or outright perversity identified both in this study and the literature, this research also highlighted relationships between these three formats and an apparent subset of dysfunctional behaviour; this may be termed *intermediate dysfunction*, as it does not necessarily arise through negligent application, nor involve deliberate dishonesty, but extends beyond the realms of unintended consequences. The primary reasons for this distinction, along with the chief characteristics of intermediate dysfunction, may be summarised as follows:

1. Intermediate dysfunction tends to manifest itself in behaviours characterised by unnecessary or disproportionate reactions to performance information, rather than directly perverse or unethical conduct.

2. Whilst unintended consequences may be unforeseen and disparate, behaviours typical of intermediate dysfunction are remarkably consistent and seem to arise with almost automatic regularity when certain performance information formats are used.
3. Intermediate dysfunction may occur even where users do not actively misuse performance information, or where environmental factors known to exacerbate dysfunction are absent.

Such cases are not widely documented within the literature, yet this study finds numerous examples of this particular category of behaviour, which seem to occur even when users engage with performance information with apparent good intentions. These include “...*short-termism and knee-jerk reactions...*” (BC3227), the inception of “...*special squads...*” (NT2071) and “...*operations to address the issue...*” (BC1302), as well as the creation of “...*plans that two days later are defunct...*” (BC2721), plus other “...*abnormal reactions, initiatives and tense meetings...*” (BC3182).

There is little to suggest such responses arise as a consequence of deliberate misuse of performance information or overbearing management styles; therefore, whilst the literature cites conditions that make the prospect of behavioural dysfunction seem reasonably predictable in some circumstances, it appears this particular type of dysfunction can occur even in the absence of such aggravating conditions. This strongly suggests that it is directly traceable to the use of simplistic data display formats.

### **6.3 Summary**

This chapter has utilised the theoretical model as a focal point for discussion, exploring its operation in the light of underlying theories, literature and research. In doing so, it explicates the relationship between data display and behavioural dysfunction more comprehensively than any previous research, by expounding the operation of the model and identifying recurrent tendencies that consistently lead to the phenomenon of interest. It also conceptualises *antecedents to data display* and *core concepts*, clearly defining their roles and influences within the model.

Centrally, the model demonstrates that data display *impels* managers to behave in certain ways when engaging with particular formats, indicating dysfunction can arise independently of misapplication, organisational conditions or other exogenous factors. Fundamentally, the study establishes clear links between engagement with certain data display formats and antecedents to dysfunction, as well as further links between such antecedents and behavioural dysfunction itself.

The discussion also highlights similarities between the study's findings and existing knowledge about behavioural dysfunction in performance management systems; this serves to verify that observed phenomena are consistent across sites, providing assurance that factors known to exacerbate dysfunction operate similarly within policing. Furthermore, the 13,135 free text comments submitted by respondents represent an unparalleled source of commentary on performance information use and dysfunction in police performance management systems.

However, building on this foundation, the study diverges from much previous research, as it draws upon data display and reference dependence theories to uncover underlying factors responsible for triggering dysfunction. Whilst the extant literature cites a range of causes, the prospect that data display may act as a standalone catalyst for behavioural dysfunction in performance management systems appears to have been overlooked until now.

Overall, this study enriches the body of research that indicates data display can influence interpretation and decision-making, by demonstrating it also exerts a substantial impact on behaviour. This finding sits at the nexus of the performance management and data display literatures and the contribution it makes will be fully explored in the final chapter.

## Chapter Seven

### Conclusions

#### 7.1 Introduction

This chapter presents the final conclusions of this study. It first provides a summary of the integrated findings of the research, before considering its theoretical and practical implications. Next, it presents the original theoretical contribution that sits at the nub of this thesis, before concluding with a brief overview of the study's limitations and potential avenues for future research.

#### 7.2 Summary of Findings

It is not disputed that a variety of organisational conditions can lead to behavioural dysfunction in performance management systems (see Moynihan, 2008; Moynihan and Lavertu, 2011; Moynihan, 2016a). However, this study finds that a fundamental condition responsible for either exacerbating or moderating such dysfunction lies in the choice of *data display* format (see Tufte, 1990; 2001; 2013; Olsen, 2013a; 2015; Isett and Hicks, 2018) used to present numeric performance information.

This prospect was first identified through qualitative analysis of empirical data, initially via sentiment analysis (see Turney, 2002; Wilson *et al*, 2005), then thematic analysis (see Corley and Gioia, 2004; Gioia *et al*, 2012). The sentiment analysis identified recurrent patterns and tendencies for certain polarities to be aligned with particular formats; binary comparisons, league tables and numerical targets were all strongly associated with negative effects, whilst SPC charts were less likely to align with negative sentiments and more likely to produce positive effects.

Next, thematic analysis established there is a traceable path from data display to dysfunction; specifically, engagement with simplistic reference-dependent formats consistently produces adverse effects, whilst interaction with statistically robust formats moderates such effects. The data structures and theoretical model comprehensively elucidate the relationships between themes and further strengthen the assertion that a generative mechanism is activated when users engage with performance information.

Furthermore, thematic analysis identified recurrent antecedents to dysfunction, such as managers demanding improvement for perceived deficiencies; a direct link was also established between certain data display formats and such conduct, as well as further links between these behaviours and subsequent dysfunction. Whilst it has long been known that certain behaviours may trigger dysfunction (see Smith, 1990; Bevan and Hood, 2006; Hood, 2006; Rothstein, 2008; Shorrocks and Licu, 2013) this research therefore addresses the question of what *causes* managers to respond in a particular fashion when using unreliable formats.

The analysis also uncovered instances of *intermediate dysfunction*, where adverse outcomes arose following interaction with simplistic formats. These reactions often took the shape of unnecessary operational responses and were more commonplace than extreme forms of dysfunction. Such occurrences may be termed ‘intermediate’, as they extend beyond the realms of unintended consequences, yet do not involve gaming, dishonesty or negligent application.

The findings of the qualitative analysis therefore strongly indicate the presence of a mechanism operating at the interface between performance information and its users; this mechanism systematically produces certain types of events when simplistic formats are used, whilst generating strikingly different events when statistically robust formats are utilised. Although the mere *existence* of unreliable formats may not cause adverse effects, it appears the act of engagement with them is a catalyst for consistent reactions by users, which ultimately lead to dysfunction.

These prospects are further supported by the experimental component of the study; when presented with hypothetical police performance information, respondents consistently misinterpreted stimuli depicting binary comparisons, league tables and numerical targets, recording elevated levels of concern and adverse behavioural tendencies. Conversely, these patterns were substantially moderated when presented with comparable data using SPC charts or contextualised peer comparison diagrams.

Furthermore, statistical analysis of this experimental output identified significant relationships between variables, whilst regression confirmed the variables pertaining to interpretation of the stimuli acted as a predictor to dysfunction. These outcomes arose despite the absence of aggravating conditions (such as pressure from managers). Additionally, other exogenous factors, such as the latent influence of

organisational culture, did not exert a major influence on observed patterns of behaviour. Again, this strongly suggests that a fundamental cause of dysfunction is likely to be an underlying agent, situated at the nexus where the powers and liabilities of the stated entities operate and interact.

The experimental portion of the study was instructive in confirming links between simplistic data display formats and common antecedents to dysfunction, along with subsequent relationships between such antecedents and dysfunction itself. Furthermore, the micro-experiments explicitly tested key relationships within the theoretical model, with the findings strongly reinforcing the sequence of events first identified during thematic analysis.

The research also found the most effective way of preventing dysfunction is to simply not engage with unreliable formats; *statistical literacy* (see Wallman, 1993; Gal, 1997; 2000) was found to be a primary factor in prompting disengagement, yet the research found considerable lack of awareness and training in this respect. Consequently, whilst superficial formats may remain inert when performance information users resist interaction, there is a strong tendency for most of them to engage, causing activation of the mechanism; this catalytic interaction consistently affects interpretation, assumptions and decision-making, leading to dysfunction.

From a methodological perspective, the substantial alignment between the quantitative and qualitative components of the research is notable, as triangulation of the outcomes at each stage identifies significant consistency throughout. This greatly enhances the findings, as it indicates the effects of using particular formats are comparable whether tested in an experimental setting, or examined through the operational experiences of practitioners. This therefore provides a powerful basis from where to launch further research, as well as inform policy and practice.

Finally, it is evident that recurrent forms of dysfunction in UK policing reflect those documented within the literature; furthermore, that observed dysfunction is closely associated with the same types of performance information formats. Taken together, this indicates consistency between the various elements of the study and strongly suggests data display is a major antecedent to the phenomenon of interest.

### 7.3 Theoretical and Practical Implications

There are several theoretical and practical implications of this research; primarily, as data display appears to be capable of amplifying or moderating dysfunction, this introduces an entirely new consideration into the study of this phenomenon. It also opens up the possibility of conflict with longstanding assumptions about conditions responsible for triggering such dysfunction.

Consequently, this study generates implications for dominant thinking about the causes of dysfunction in performance management systems. For instance, what does the interplay between factors identified within this study and other influencing conditions look like? Would further research find contradictions between the claims made in this study and pre-existing knowledge about causes of dysfunction, or would they complement each other? Could the outcomes of this study potentially enhance some prior findings and undermine others?

The study also carries implications relevant to the specific theories cited within this thesis. For example, with regard to reference dependence theories (see Festinger, 1954; Albert, 1977; Tversky and Kahneman, 1986; 1991; Olsen, 2013b) the outcomes of this research indicate the use of isolated reference points in making temporal, social or aspirational comparisons is a highly counterproductive practice. Therefore, it may be beneficial to further explore the implications of reference dependence for performance information use, in light of the study's findings.

Similarly, in respect of Goal Setting Theory (see Locke, 1968; Locke *et al*, 1981, 1989; Locke and Latham, 1990; 2006), the findings could impact upon established thinking regarding the design of organisational goals, as it is clear that where targets are presented in a simplistic binary fashion, this is likely to trigger dysfunction, whereas different visual presentation of the same goals may not. Therefore, this research has implications for the theoretical concept of 'goals as reference points' (Heath *et al*, 1999) by establishing that the format used to present a particular target may influence behaviour as much as the existence of the target itself.

Furthermore, the thesis has cited studies that examine the way in which data display can affect the interpretation of performance information (see Brewer *et al*, 2012; Kurtzman and Greene, 2016; Schmidtke *et al*, 2017a), along with its impact on

decision-making (Elbel *et al*, 2014; NHS, 2018); this study enriches this field of research by demonstrating how the influence of data display reaches beyond cognitive processes, to have a direct bearing on behaviour. This has theoretical implications for how we understand engagement between performance information and its users, as well as how data display can exert influences on the cognitive and behavioural processes that lead to dysfunction.

Additionally, using data display to explain behavioural phenomena associated with performance information use is rare; consequently, theoretical implications could extend far into uncharted reaches of this field. In particular, the postulation of a mechanism responsible for behavioural tendencies opens up the prospect of a ‘blind spot’ in this strand of the literature (see Moynihan, 2008; Micheli and Neely, 2010; Moynihan *et al*, 2012; Nielsen, 2013; Kroll, 2013; 2014; 2015; Kroll and Moynihan, 2017). Fundamentally, the theory generated as a result of this study compels scholars to consider the possibility that data display exerts a direct influence on behaviour.

With regard to practical implications, the study’s chief impact is in respect of how it informs the design and use of police performance information. As the research establishes the effects of data display are far-reaching and can trigger dysfunction even in the absence of misapplication, this places deep responsibility on those charged with designing and using performance information within the police service to ensure formats are statistically robust. Such implications may also reach beyond policing and it could be instructive to further explore this through further research.

Finally, the findings suggest there is a firm case for ensuring performance information users are suitably trained and equipped to recognise and interpret reliable forms of data display, in order to avoid the pitfalls of simplistic formats. It is however recognised that such a fundamental shift in what has become embedded thinking and practice is likely to involve significant challenges, both culturally and organisationally. It is therefore suggested such a shift is not simply a case of supplanting reference-dependent formats with statistically robust ones, as a fresh mindset is required; police managers need to envisage performance data as a source of information to aid decision-making, rather than a tool to ‘name and shame’.

## 7.4 Contribution

In reviewing an early iteration of this thesis, Professor Donald Moynihan (Director of the La Follette School of Public Affairs, University of Wisconsin-Madison and President of the Public Management Research Association) remarked that conventional wisdom tends to ascribe perversity in performance management systems to well-documented factors, such as organisational or contextual conditions; meanwhile, he observed there is scant research into how *the presentation of performance information itself* might also be partly responsible. He then went on to comment upon the main contribution of the thesis, as he saw it:

*“...you both contribute to work on perversity, but suggest that at least some of the underpinnings come from cognitive biases that arise from the information we are presented with. That is to say, not all perversity is the function of incentives or leadership or culture, but the information we are presented with predictably leads to faulty heuristics kicking in, and sets us on the path to engaging in perverse use of data. I think that’s a real contribution to make” (Moynihan, 2016a).*

Therefore, it is proposed this thesis makes a substantial and original theoretical contribution, high in methodological rigour and practical relevance. Specifically, it identifies data display as being a significant antecedent to the phenomenon of interest, establishing the likely presence of an underlying generative mechanism and providing an explanatory theoretical account. Ultimately, the study *extends* the boundaries of where it is believed the influence of data display may reach, by establishing a verifiable path from data display, through decision making, to the phenomenon of interest; this prospect has not been explored elsewhere.

The study contributes to management research by identifying underlying factors in the design of data display responsible for triggering dysfunction. It introduces the notion that certain formats consistently impair engagement with data, producing a pattern of cognitive and behavioural responses that culminate in dysfunction; the study proposes the notions of *antecedents to data display* and *core concepts* to explicate this sequence of events. The research also identifies pivotal relationships between certain formats and prominent antecedents to dysfunction, demonstrating how such antecedents produce adverse outcomes; these range from unintended

consequences and *intermediate dysfunction* at one end of the scale, to outright perversity and unethical behaviour at the other.

Fundamentally, the study establishes that data display formats founded on comparisons to isolated reference points consistently trigger unwarranted assumptions about apparent changes or differences and encourage a narrative describing apparent increases or decreases with a plausible air of authority that belies the efficacy of such formats. This then frames and shapes conversations about performance, leading to a foreseeable sequence of events responsible for producing behavioural dysfunction.

It is therefore asserted that behavioural phenomena arising from engagement with performance information are *directly shaped by the chosen data display format*. This is demonstrated from multiple perspectives in this study; where all factors remain constant apart from the data display format, when engaging with performance information, the same users operating within the same context exhibit radically different assumptions and responses, depending solely on how data are presented.

From a critical realist perspective, the study utilises Easton's (2010) retroductive four step model as a framework for addressing the research question, by identifying the entities and mechanisms responsible for producing events. Consequently, by envisioning the interaction of entities and activation of a generative mechanism, this enables the production of an explicative theoretical model, alongside a robust explanatory account. Such an account supplements the visual representation proffered by the model and may be outlined as follows:

“*Entities* (performance information users) having *structures* (e.g. knowledge, personality traits), operating within other structures (UK police service) and necessarily possessing *causal powers* (to act / influence others) and *liabilities* (e.g. extent of statistical literacy) will, under different specific conditions (using different data display formats - i.e. further entities with causal powers and liabilities), produce *events* (particular decisions and actions), which contribute towards the phenomenon of interest (behavioural dysfunction)”.

This account is founded upon the tenets of critical realist causal explanation; it also acts as a straightforward expository narrative, yet one which possesses the

ontological and epistemological rigour necessary to make a theoretical claim. Additionally, it not only distils key factors that produce the phenomenon of interest in a manner consistent with critical realism, but retains significant relevance to the practical application of performance information within operational settings.

This doctoral research therefore provides a solid foundation from which to argue for a significant rethink about the standalone behavioural power of data display, as well as a sea change in how police performance information is presented and used. It is therefore critical that data display formats must ‘reveal the truth’ as asserted by Tufte (2013); this is because their visual appearance exerts a heavy influence on the nature of what follows once users engage with the data.

Consequently, returning to the research question...

*“Does data display influence the likelihood, nature or extent of behavioural dysfunction in police performance management systems, and if so, why?”*

...this study finds data display is indeed singularly capable of triggering behavioural dysfunction in police performance management systems and this can be attributed to the way in which it shapes initial perceptions and subsequent decisions. Most significantly however, the research comprehensively establishes data display directly influences the likelihood, nature *and* extent of such dysfunction.

## **7.5 Limitations and Further Research**

There are limitations to this study that need to be acknowledged. For example, it is accepted it would be impossible for engagement with the micro-experiments to be absolutely sterile. Nevertheless, respondents were, as much as practicable, able to interact with the stimuli away from influences or pressures that may exist in their operational environments.

Similarly, it is acknowledged that researcher bias (Pannucci and Wilkins, 2010) cannot be completely eliminated, although again, steps were taken to mitigate this through the research design. Consequently, although this is the first work to examine all of the constructs depicted within the model together, it is not claimed that the findings are infallible. Therefore, it is recommended further research be undertaken

to test the claims made, as well as explore associated areas that lie beyond the scope of this study.

Firstly, elements of the theoretical model could be further assessed; it may be possible to test the strength of relationships between particular components, or via the introduction of new variables, to ascertain if they exert an influence on the postulated mechanism or observed outcomes. Questions could include: Would the introduction of additional factors substantially impact upon the patterns and tendencies produced by the experiments? If so, what are they and how do they exert an influence?

Similarly, it may be worthwhile evaluating whether there are any reinforcing loops, reciprocal effects, or currently unidentified associations between components of the model. Further experiments might also locate new mechanisms capable of influencing themes (or exerting moderating effects along one or more of the arrows). Also, where a theme does not currently have an 'incoming' arrow, what factors influence it?

Next, there are opportunities for extending existing aspects of this study. For instance, this could involve testing additional variants of the stimuli depicting a range of proximities from reference points; in particular, where negative variance exists, it may be possible to assess whether negativity bias could help explain extreme tendencies towards dysfunction. If additional stimuli depicted performance at greater or lesser variances, would assumptions about performance being 'good', 'poor', or 'acceptable' vary, depending on relative distances, and would this affect behavioural outcomes?

Similarly, labelling stimuli to denote crime types of a more or less serious nature (e.g. 'rape' vs 'shop theft') could be worth exploring to see if this affects behaviour, especially if the actual interpretation of each such stimulus is comparable; the same applies to incorporating a numeric scale alongside some stimuli, to test whether respondents react differently to when being assessed against patterns of data with no scale. A more complex set of stimuli could therefore be deployed to test whether different labels or scales affect participants' interpretations or behavioural responses; however, at the very least, it seems worthwhile to replicate the original exploratory tests on a larger scale to produce results that are more representative.

A further consideration not addressed by this research is how the characteristics of performance information users might affect operation of the model. The study found users tended to respond to the experimental stimuli consistently despite pre-existing characteristics; nevertheless, an interesting line of inquiry could be to assess the impact of heuristics or biases capable of affecting how data are interpreted or used (see Tversky and Kahneman, 1974; Jacobsen *et al*, 2014; Baekgaard and Serritzlew, 2016; Nielsen and Moynihan, 2017b; Christensen *et al*, 2018).

Finally, it may be beneficial to utilise the methodology employed by this research to conduct parallel studies within the fields of healthcare and education (and potentially other public or even private sector domains) to investigate whether data display is also a contributory factor towards behavioural dysfunction there. Overall, it appears that synthesis of the fields of performance information use and data display could pay dividends by unlocking new perspectives on the relationship between data display and behavioural dysfunction.

## References

- Adab, P., Rouse, A. M., Mohammed, M. A. and Marshall, T. (2002) *Performance League Tables: The NHS Deserves Better*. [Online]  
<https://www.bmj.com/content/324/7329/95> [Accessed 15th June 2018]
- Adams, R. (2013) [Online] *GCSE Results Could Be Distorted*.  
<http://www.theguardian.com/education/2013/aug/01/gcse-exam-results-could-be-distorted> [Accessed 17th August 2013]
- AEC (Australian Education Council) (1991). *A National Statement on Mathematics for Australian Schools*. Carlton, Victoria: Curriculum Corporation
- Albrecht, S. and Travaglione, A. (2003) 'Trust in Public-Sector Senior Management'. *International Journal of Human Resource Management*. 14 (1): 76-92
- Altman, L. K. (1990) *Heart-Surgery Death Rates Decline in New York*. [Online]  
<http://www.nytimes.com/1990/12/05/nyregion/heart-surgery-death-rates-decline-in-new-york.html?pagewanted=all&src=pm> The New York Times. 5<sup>th</sup> December 1990. [Accessed 22<sup>nd</sup> July 2013]
- Alvesson, M., and Kärreman, D. (2007). 'Constructing Mystery: Empirical Matters in Theory Development'. *Academy of Management Review*. 32: 1265-1281
- Ammons, D. N. (1997) 'Raising the Performance Bar Locally'. *Public Management Magazine*. 79: 10–16
- Ammons, D. N. and Rivenbark, W. C. (2008) 'Factors Influencing the Use of Performance Data to Improve Municipal Services: Evidence from the North Carolina Benchmarking Project'. *Public Administration Review*. 68: 304–18
- Andersen, S. C. and Moynihan, D. P. (2016a) 'How Leaders Respond to Diversity: The Moderating Role of Organizational Culture on Performance Information Use'. *Journal of Public Administration Research and Theory*. 26 (3): 448-460
- Andersen, S. and Moynihan, D. P. (2016b). 'Bureaucratic Investments in Expertise: Evidence from a Randomized Controlled Field Trial'. [In press]. *Journal of Politics*
- Anderson, D. R. (2009) "Peirce and Cartesian Rationalism". In Shook, J. R. and Margolis, J. (Eds). *A Companion to Pragmatism: Blackwell Companions to Philosophy*. Oxford: Wiley-Blackwell. pp.154-165
- Anderson, N., Herriot, P. and Hodgkinson, G. P. (2001) 'The Practitioner–Researcher Divide in Industrial, Work and Organizational (IWO) Psychology: Where Are We Now, and Where Do We Go From Here?' *Journal of Occupational and Organizational Psychology*. 74: 391–411
- Anderson, P. J. J., Blatt, R., Christianson, M. K., Grant, A. M., Marquis, C., Neuman, E. J., Sonenshein, S. and Sutcliffe, K. M. (2006) 'Understanding Mechanisms in Organizational Research: Reflections from a Collective Journey'. *Journal of Management Inquiry*. 15: 102–113

- Anhøj, J. and Blok Hellesøe, A. (2016) 'The Problem with Red, Amber, Green: The Need to Avoid Distraction by Random Variation in Organisational Performance Measures'. *BMJ Quality & Safety*. 26: 81-84
- Antaki, C., Billig, M., Edwards, D. and Potter, J. (2003) 'Discourse Analysis Means Doing Analysis: A Critique of Six Analytic Shortcomings'. *Discourse Analysis Online*. 1(1):1-24
- Appleton, J. V. and King, L. (2002) 'Journeying From the Philosophical Contemplations of Constructivism to the Methodological Pragmatics of Health Services Research'. *Journal of Advanced Nursing*. 40(6): 641–648
- Argyris, C. (1952) *The Impact of Budgets on People*. New York: The Controllershship Foundation
- Armstrong, J. (1997) *Stewardship and Public Service*. Ottawa: Public Service Commission of Canada
- Ary, D., Jacobs, L., and Razavieh, A. (1996) *Introduction to Research in Education*. Fort Worth, TX: Holt, Rinehart, and Winston
- Ashforth, B. E. and Anand, V. (2003) 'The Normalization of Corruption in Organizations'. *Research in Organizational Behavior* 25: 1–52
- Audit Commission (2000) *Local Authority Performance Indicators 1998-99*. London: Audit Commission
- Audit Commission (2007) *Response Times: Great Western Ambulance Service NHS Organisation in Respect of the Former Wiltshire Ambulance Service NHS Organisation*. London: Audit Commission
- Babbie, E. R. (1990) *Survey research methods*. Belmont, CA: Wadsworth
- Bacharach, S. B. (1989) 'Organizational Theories: Some Criteria for Evaluation'. *The Academy of Management Review*. 14 (4): 496-515
- Bacon, M. (2012) *Pragmatism: An Introduction*. Cambridge: Polity Press
- Baekgaard, M. and Serritzlew, S. (2016) 'Interpreting Performance Information: Motivated Reasoning or Unbiased Comprehension?!' *Public Administration Review*. 76 (1): 73-82
- Baird, D., Baker, D., Gordon, G., Smoker, R. and Whitney, R. (1981) *Managing for performance*. Unpublished manuscript, Philadelphia: The Hay Group
- Bandura, A. (1999) 'Moral Disengagement in the Perpetration of Inhumanities'. *Personality and Social Psychology Review*. 3: 193 - 209
- Bandura, A., and Cervone, D. (1983) 'Self-Evaluative and Self-Efficacy Mechanisms Governing the Motivational Effects of Goal Systems'. *Journal of Personality and Social Psychology*. 45: 1017-1028

- Bar-Eli, M., Azar, O. H., Ritov, I., Keidar-Levin, Y. and Schein, G. (2007) 'Action Bias among Elite Soccer Goalkeepers: The Case of Penalty Kicks'. *Journal of Economic Psychology*. 28 (5): 606–621
- Baron, J. and Ritov, I. (1994) 'Reference Points and Omission Bias'. *Organizational Behavior and Human Decision Processes*. 59: 475–498
- Baron, J., and Ritov, I. (2004) 'Omission Bias, Individual Differences, and Normality'. *Organizational Behavior and Human Decision Processes* 94. (2): 74–85
- Baron, R. M. and Kenny, D. A. (1986) 'The Moderator – Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic and Statistical Considerations'. *Journal of Personality and Social Psychology*. 51 (6): 1173-1182
- Barratt, D. (2007) *Targets 'Force Police to Make Ludicrous Arrests'*. The Independent. [Online] <http://www.independent.co.uk/news/uk/crime/targets-force-police-to-make-ludicrous-arrests-448921.html> [Accessed 30th July 2013]
- Barsky, A. (2008) Understanding the Ethical Cost of Organizational Goal-Setting: A Review and Theory Development. *Journal of Business Ethics*. 81: 63–81
- Barton, H. and Beynon, N. (2011) 'Targeted Criteria Performance Improvement – An Investigation of a 'Most Similar' UK Police Force'. *International Journal of Public Sector Management*. 24 (4): 356-367
- Baruch, Y. (1999) 'Response Rates in Academic Studies: A Comparative Analysis'. *Human Relations*. 52: 451-468
- Bass, B., and Riggio, E. G. (2005) *Transformational Leadership*. (2nd ed.) Mahwah, NJ: Lawrence Erlbaum Associates
- Bayley, D. (1994) *Policing for the Future*. New York, Oxford University Press
- Baumeister, R. F., Bratslavsky, E., Finkenauer, C., and Vohs, K. D. (2001) 'Bad is Stronger Than Good'. *Review of General Psychology*. 5 (4): 323-370
- Bedfordshire Police (2008) *Performance Report to the Police Authority*. [Online] [http://www.bedfordshirepoliceauthority.co.uk/documents/archive/Audit-and-Business-Assurance-Committee/18-March-2008\\_07-08-Performance-Q3.pdf](http://www.bedfordshirepoliceauthority.co.uk/documents/archive/Audit-and-Business-Assurance-Committee/18-March-2008_07-08-Performance-Q3.pdf) [Accessed 13th July 2013]
- Behn, R. D. (2001) *Rethinking Democratic Accountability*. Washington, DC: Brookings Institution
- Behn, R. D. (2002) 'The Psychological Barriers to Performance Management: Or Why Isn't Everyone Jumping on the Performance Management Bandwagon?' *Public Performance and Management Review*. 26 (1): 5-25
- Behn, R. D. (2003) 'Why Measure Performance? Different Purposes Require Different Measures'. *Public Administration Review*. 63: 586 – 606
- Behn, R. D. (2006) 'The Varieties of CitiStat'. *Public Administration Review*. 66 (3): 332-340

Behn, R. D. (2008a) *The Seven Big Errors of PerformanceStat*. Rappaport Institute for Greater Boston and the Taubman Center for State and Local Government (Policy Brief, February), Cambridge, Mass.

Behn, R. D. (2008b) *PerformanceStat as a Leadership Strategy: It Don't Mean a Thing if it Ain't Got that Follow-Up*. A Paper Prepared for the Twelfth Annual Conference of The International Research Society for Public Management Panel on Public Management in Practice March 26-28, 2008, Brisbane, Australia

Behn, R. D. (2014) *The PerformanceStat Potential: A Leadership Strategy for Producing Results*. Washington, DC: Brookings Institution Press/Ash Center

Behn, R. D. and Kant, P. A. (1999) 'Strategies for Avoiding the Pitfalls of Performance Contracting'. *Public Productivity and Management Review*. 22 (4): 470-489

Benaine, S. L. and Kroll, A. (2019) 'Explaining Effort Substitution in Performance Systems: The Role of Task Demands and Mission Orientation'. *Public Management Review*. 1-23. DOI: 10.1080/14719037.2019.1604794

Bergin, M., Wells, J. S. G. and Owen, S. (2008) 'Critical Realism: A Philosophical Framework for the Study of Gender and Mental Health'. *Nursing Philosophy*. 9: 169-179

Berliner, J. S. (1956) 'A Problem in Soviet Business Management'. *Administrative Science Quarterly*. 1: 86–101

Berman, E. and Wang, X. (2000) 'Performance Measurement in U.S. Counties: Capacity for Reform'. *Public Administration Review*. 60(5): 409–420

Berman, G. and Dar, A. (2013) *Police Service Strength*. London: House of Commons Library, Standard Note: SN00634

Bernstein, D. and Isackson, N. (2014) The Truth About Chicago's Crime Rates. Chicago Magazine. [Online] <http://www.chicagomag.com/Chicago-Magazine/May-2014/Chicago-crime-rates/> [Accessed 19<sup>th</sup> April 2014]

Besharov, M. L. (2014) 'The Relational Ecology of Identification: How Organizational Identification Emerges When Individuals Hold Divergent Values'. *Academy of Management Journal*. 57(5): 1185-1512

Bethlehem, J. G. (2007) *Reducing the Bias of Web survey Based Estimates*. Discussion paper 07001. Voorburg / Heerlen: Statistics Netherlands

Bethlehem, J. (2008) *How Accurate are Self-Selection Web Surveys?* [Online] <http://www.cbs.nl/NR/rdonlyres/EEC0E15B-76B0-4698-9B26-8FA04D2B3270/0/200814x10pub.pdf> The Hague: Statistics Netherlands [Accessed 23<sup>rd</sup> January 2014]

Bevan, G. and Hood, C. (2006) 'What's Measured is What Matters: Targets and Gaming in the English Public Healthcare System'. *Public Administration* 84 (3): 517-538

- Bevan, G. and Hamblin, R. (2009). 'Hitting and Missing Targets by Ambulance Services for Emergency Calls: Effects of Different Systems of Performance Measurement within the UK'. *Journal of the Royal Statistical Society (A)*. 172 (1): 161 – 190
- Beverley, C. and Haynes, J. (2005) *Franchised Trusts. Health Management Specialist Library: Management Briefing*. NeLH Health Management Specialist Library
- Bhaskar, R. (1975) *A Realist Theory of Science*. London: Verso
- Bhaskar, R. (1979) *The Possibility of Naturalism: A Philosophical Critique of the Contemporary Human Science*. Atlantic Highlands, NJ: Humanities Press
- Bhaskar, R. (1986) *Scientific Realism and Human Emancipation*. London: Verso
- Bhaskar, R. (1989) *Reclaiming Reality*. London: Verso
- Bhaskar, R. (1997) 'On the Ontological Status of Ideas'. *Journal for the Theory of Social Behaviour*. 27 (2/3): 139-147
- Bhaskar, R. (1998a) "Philosophy and Scientific Realism". In Archer, M., Bhaskar, R., Collier, A., Lawson, T. and Norrie, A. (Eds.) *Critical Realism: Essential Readings*. (pp. 16–47). London: Routledge
- Bhaskar, R. (1998b) *The Possibility of Naturalism: A Philosophical Critique of the Contemporary Human Sciences*. (3rd edn). London: Routledge
- Bhaskar, R. (1998c) "General Introduction". In Archer, M. S., Bhaskar, R., Collier, A., Lawson, T. and Norrie, A. (eds.) *Critical Realism: Essential Readings*. London: Routledge. pp. ix-xxiv
- Bhaskar, R. (1998d) "Societies". In Archer, M. et al (Eds) *Critical Realism: Essential Readings*. London: Routledge
- Bhaskar, R. (2002) *From Science to Emancipation: Alienation and the Actuality of Enlightenment*. Delhi: Sage Publications India Pvt Ltd.
- Biggs, J. and Collis, K. (1982) *Evaluating the Quality of Learning: The SOLO Taxonomy*. New York, NY: Academic Press
- Birnberg, J. G., Luft, J. and Shields, M. D. (2009) "Psychology Theory in Management Accounting Research". In Chapman, C. S., Hopwood, A. G. and Shields, M. D. (Eds.) *Handbook of Management Accounting Research*. (Vol 1: pp. 113-135) Oxford: Elsevier
- Bird, S. M., Cox, D., Farewell, V. T., Goldstein, H., Holt, T. and Smith, P. C. (2005) 'Performance Indicators: Good, Bad and Ugly'. *Journal of the Royal Statistical Society (A)*. 168 (1): 1-27
- Bititci, U. S., Bourne, M., Cross, J., Nudurupati, S. and Sang, K. (2018) 'Towards a Theoretical Foundation for Performance Measurement and Management'. *International Journal of Management Reviews*. 20 (3): 653-660

- Bititci, U. S., Carrie, A. S. and McDevitt, L. (1997) 'Integrated Performance Measurement Systems: a Development Guide'. *International Journal of Operations & Production Management*. 17 (5-6): 522-534
- Blau, P. M. (1955) *The Dynamics of Bureaucracy*. Chicago: University of Chicago Press
- Bloor, M., Frankland, J., Thomas, M. and Robson, K. (2001) *Focus Groups in Social Research*. London: Sage
- Bommer, W. H., Johnson, J. L., Rich, G. A., Podsakoff, P. M. and McKenzie, S. B. (1995) 'On the Interchangeability of Objective and Subjective Measures of Employee Performance: a Meta-Analysis.' *Personnel Psychology* 48 (3): 587-605
- Borg, W. R., and Gall, M. D. (1983). *Educational Research: An Introduction*. New York: Longman
- Botterill, C. B. (1977) 'Goal Setting and Performance on an Endurance Task'. (Doctoral Thesis). Alberta: University of Alberta
- Bourne, M. (2008) 'Performance Measurement: Learning from the Past and Projecting the Future'. *Measuring Business Excellence*. 12 (4): 67-72
- Bourne, M. and Bourne, P. A. (2011) *The Handbook of Corporate Performance Management*. London: Wiley
- Bourne, M. and Franco-Santos, M. (2010) *Target Setting*. [Online] <http://www.som.cranfield.ac.uk/som/p14513/Think-Cranfield/2010/March-2010/Setting-targets> [Accessed 9th August 2013]
- Bourne, M., Franco-Santos, M., Micheli, P. and Pavlov, A. (2018) 'Performance Measurement and Management: A System of Systems Perspective'. *International Journal of Production Research*. 56 (8): 2788-2799
- Bourne, M., Melnyk, S. and Bititci, U. S. (2018) 'Performance Measurement and Management: Theory and Practice'. *International Journal of Operations & Production Management*. 38 (11): 2010-2021
- Bourne, M. and Mura, M. (2018) 'Performance and Risk Management'. *Production Planning & Control*. 29 (15): 1221-1224
- Bourne, M., Micheli, P. and Franco, M. (2009) *Business Performance and Target Setting*. [Online] <http://www.som.cranfield.ac.uk/som/dinamic-content/think/documents/setting-targets.pdf> [Accessed 10th August 2015]
- Bourdeaux, C., and Chikoto, G. (2008) 'Legislative Influences on Performance Management Reform'. *Public Administration Review*. 68 (2): 253-265
- Boyatzis, R. (1998) *Transforming Qualitative Information: Thematic Analysis and Code Development*. Thousand Oaks, CA: Sage

- Boyne, G. A. and Chen, A. A. (2006) 'Performance Targets and Public Service Improvements'. *Journal of Public Administration Research and Theory*. 17: 455 – 477
- Boyne, G., Martin, S. and Walker, R. (2004) 'Explicit Reforms, Implicit Theories and Public Service Improvement'. *Public Management Review*. 6 (2): 211 – 228
- Brace, N., Kemp, R. and Snelgar, R. (2009) *SPSS for Psychologists*. (4<sup>th</sup> Edn). Basingstoke: Palgrave Macmillan
- Bratton, W. J. (1998) "Crime is Down in New York City: Blame the Police". In Dennis, N. (ed.). *Zero tolerance: Policing a Free Society*, 2nd ed. (pp.29-42). London: IEA
- Bratton, W. J. And Malinowski, S. W. (2008) 'Police Performance Management in Practice: Taking COMPSTAT to the Next Level'. *Policing*. (2) 3: 259–265
- Braun, V. and Clarke, V. (2006) 'Using Thematic Analysis in Psychology'. *Qualitative Research in Psychology*. 3(2): 77-101
- Brewer, A. M. (1996) 'Developing Commitment between Managers and Employees'. *Journal of Managerial Psychology*. 11 (4): 24
- Brewer, N. T., Gilkey, M. B., Lillie, S. E., Hesse, B. W. and Sheridan, S. L. (2012) 'Tables or Bar Graphs? Presenting Test Results in Electronic Medical Records'. *Medical Decision Making*. 32: 545–553
- Brickman, P., and Bulman, R. J. (1977) "Pleasure and Pain in Social Comparison". In Suls, J. M. and Miller, R. L. (Eds.) *Social Comparison Processes: Theoretical and Empirical Perspectives*. Washington, DC: Hemisphere
- British Medical Association (2007) *Emergency Medicine: A Report of a National Survey of Emergency Medicine*. Health Policy and Economic Research Unit. London: BMA
- British Transport Police (2012) *London Underground / DLR Policing Plan 2012-2013*. [Online] <http://www.btp.police.uk/pdf/ludlr-policing-plans-web-2012-13.pdf> [Accessed 2nd August 2013]
- British Transport Police Authority (2015) *PRC: Performance 2014-2015 Q4*. [Online] [http://btpa.police.uk/livesite/wp-content/uploads/2014/06/Item-3.-Q4-Performance-Report-and-Appendices-combined-for-website\\_NPM.pdf](http://btpa.police.uk/livesite/wp-content/uploads/2014/06/Item-3.-Q4-Performance-Report-and-Appendices-combined-for-website_NPM.pdf) [Accessed 11th June 2015]
- Britten, N. (1999) "Qualitative Interviews in Healthcare". In Pope, C. and Mays, N. (eds) *Qualitative Research in Health Care*. 2nd ed. Pp. 11–19. London: BMJ Books
- Broadhurst, K., Wastell, D., White, S., Hall, C, Peckover, S., Thompson, K., Pithouse, A. and Davey, D. (2010) 'Performing 'Initial Assessment': Identifying the Latent Conditions for Error at the Front-Door of Local Authority Children's Services'. *British Journal of Social Work*. 40 (2): 352-370
- Brown, M. (1996) *Keeping Score: Using the Right Metrics to Drive World Class Performance*. New York: Quality Resources

- Brown, D. (2013) *Comment: The Met's Stephen Lawrence Outrage is Just the Tip of the Iceberg*. [Online] <http://www.politics.co.uk/comment-analysis/2013/06/26/comment-the-met-s-stephen-lawrence-outrage> [Accessed 28th April 2014]
- Brundin, E., Patzelt, H. and Shepherd, D.A. (2008). 'Managers' Emotional Displays and Employees' Willingness to Act Entrepreneurially'. *Journal of Business Venturing*. 23 (2): 221-243
- Bryman, A. (2001) *Social Research Methods*. Oxford: Oxford University Press
- Bunge, M. (2004) "How Does it Work? The Search for Explanatory Mechanisms". *Philosophy of the Social Sciences*. 34 (2): 182-210
- Burrell, G. and Morgan, G. (1979) *Sociological Paradigms and Organisational Analysis*. London: Heinemann Educational Books
- Bygstad, B. and Munkvold, B. E. (2011) *In Search of Mechanisms: Conducting a Critical Realist Data Analysis*. Research paper: Thirty Second International Conference on Information Systems. Shanghai, 2011.
- Bylander, T., Allemang, D., Tanner, M. C. and Josephson, J. R. (1991) 'The Computational Complexity of Abduction'. *Artificial Intelligence* 49: 25-60
- Caers, R., Du Bois, C., Jegers, M., De Gieter, S., Schepers, C. and Pepermans, R. (2006) 'Principal-Agent Relationships on the Stewardship-Agency Axis'. *Nonprofit Management and Leadership*. 17 (1): 25-47
- Caldwell, B. J. (1984) 'Some Problems with Falsificationism in Economics'. *Philosophy of the Social Sciences*. 14: 489-495
- Callingham, R. and Watson, J. M. (2017) 'The Development of Statistical Literacy at School'. *Statistics Education Research Journal*. 16 (1): 181-201
- Capon, N., Farley, J. and Hubert, J. (1987) *Corporate Strategic Planning*. New York: Columbia University Press
- Card, S. K., Mackinlay, J. D. and Schneiderman, B. (eds.) (1999) *Readings in Information Visualization: Using Vision to Think*. San Francisco: Morgan Kaufmann
- Carifio, J. and Perla, P. (2007) 'Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes'. *Journal of Social Sciences*. 3 (3): 106-116
- Carifio, J. and Perla, P. (2008) 'Resolving the 50-Year Debate Around Using and Misusing Likert Scales'. *Medical Education*. 42: 1150-1152
- Carnoy, M., Jacobsen, R., Mishel, L. and Rothstein, R. (2005) *The Charter School Dust-Up. Examining the Evidence on Enrollment and Achievement*. New York: Teachers College Press
- Carter B (2000) *Realism and Racism: Concepts of Race in Sociological Research. Critical Realism Interventions Series*. London: Routledge

Carvel, J. (2003) *Hospitals Faked Wait Times Test*. The Guardian. May 13, 2003 p. 1

Cassell, C. and Johnson, P. (2006) 'Action Research: Explaining the Diversity'. *Human Relations*. 59 (6): 783-814

Castellano, J. F., Young, S. and Roehm, H. A. (2004) 'The Seven Fatal Flaws of Performance Management'. *CPA Journal*. 74 (6): 32-35

Centre for Crime and Justice Studies (2007) *Ten Years of Criminal Justice Under Labour – An Independent Audit*. London: Centre for Crime and Justice Studies

Charbonneau, E. and Van Ryzin, G. G. (2015) 'Benchmarks and Citizen Judgments of Local Government Performance: Findings from a Survey Experiment'. *Public Management Review*. 17 (2): 288–304

Chatterton M., (2008) *Losing the Detectives; Views from the Frontline*. Surbiton: Police Federation of England and Wales

Chatterton M. and Bingham, E. (2006) *24/7 Response Policing in the Modern Police Organisation – Views from the Frontline*. Police Federation of England and Wales: December 2006

Chen, C. (1999) *Information Visualisation and Virtual Environments*. Springer: Berlin

Chen, C. (2006) *Information Visualization: Beyond the Horizon*. (2nd Ed). London: Springer-Verlag

Cherryholmes, C. H. (1992) 'Notes on Pragmatism and Scientific Realism'. *Educational Researcher*. 21 (6): 13-17

Cherryholmes, C. H. (1994) 'More Notes on Pragmatism'. *Educational Researcher*. 23 (1): 16-18

Chief Secretary to the Treasury. (1998a) *Modern Public Services for Britain: Investing in Reform. Comprehensive Spending Review: New Public Spending Plans 1999–2002*. London: HMSO

Chief Secretary to the Treasury. (1998b) *Public Services for the Future: Modernisation, Reform, Accountability. Comprehensive Spending Review: Public Service Agreements 1999–2002*. London: HMSO

Choi, I. and Moynihan, D. (2019) 'How to Foster Collaborative Performance Management? Key Factors in the US Federal Agencies'. *Public Management Review*. 21 (10): 1538-1559

Christensen, J., Dahlmann, C. M., Mathiasen, A. H., Moynihan, D. P. and Petersen, N. B. G. (2018) 'How do Elected Officials Evaluate Performance? Goal Preferences, Governance Preferences, and the Process of Goal Reprioritization'. *Journal of Public Administration Research and Theory*. 28 (2): 197-211

- Chrystal, K. A. and Mizen, P. D. (2001) *Goodhart's Law: Its Origins, Meanings and Implications for Monetary Policy*. [Online]  
[http://www.cyberlibris.typepad.com/blog/files/Goodharts\\_Law.pdf](http://www.cyberlibris.typepad.com/blog/files/Goodharts_Law.pdf) [Accessed 11th May, 2015]
- CIPD (2003) *People and Public Services. Why Central Targets Miss the Mark. The Change Agenda*. London: Chartered Institute of Personnel and Development
- CIPD (2016a) *Rapid Evidence Assessment on the Research Literature on the Effect of Goal Setting on Workplace Performance*. London: CIPD
- CIPD (2016b) *Could do Better: Assessing What Works in Performance Management*. London: CIPD
- Claramunt, C., Jiang, B. and Bargiela, A. (2006) 'A New Framework for the Integration, Analysis and Visualisation of Urban Traffic Data within Geographic Information Systems'. *Transportation Research Part C*. 8: 167-184
- Clarke, B., Clarke, D., and Cheesman, J. (2006) 'The Mathematical Knowledge and Understanding Young Children Bring to School'. *Mathematics Education Research Journal*. 18 (1): 78–103
- Cleveland, W. S. (1985) *The Elements of Graphing Data*. Monterey: Wadsworth
- Cleveland, W. S. and McGill, R. (1984) 'Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods'. *Journal of the American Statistical Association*. 79 (387): pp. 531-554
- Cockcroft, T. and Beattie, I. (2009) 'Shifting Cultures: Managerialism and the Rise of 'Performance''. *Policing: An International Journal of Police Strategies and Management*. 32 (3): 526 – 540
- Cohen, J. (1960) 'A Coefficient of Agreement for Nominal Scales'. *Educational and Psychological Measurement*. 20 (1): 37–46
- Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press
- Colbert, A. E., Kristof-Brown, A. I., Bradley, B. H., and Barrick, M. R. (2008) 'CEO Transformational Leadership: The Role of Goal Importance Congruence in Top Management Teams'. *Academy of Management Journal*: 51: 81–96
- Colman, A. M., Norris, C. E., and Preston, C. C. (1997) 'Comparing Rating Scales of Different Lengths: Equivalence of Scores From 5-point and 7-point Scales'. *Psychological Reports*. 80: 355-362
- Coleman, C. and Moynihan, J. (1996) *Understanding Crime Data: Haunted by the Dark Figure*. London: Open University Press
- College of Policing (2013) *Performance Management / Service Delivery Commission*. 20<sup>th</sup> May 2013. Unpublished

College of Policing (2014) *Effective Performance Measurement Principles*. (Draft: version 6.1. Guilfoyle / Wiggett. December 2014). Unpublished

College of Policing (2015) *College Response to Police Targets Report*. [Online] [http://www.college.police.uk/News/College-news/Pages/police\\_targets\\_report.aspx](http://www.college.police.uk/News/College-news/Pages/police_targets_report.aspx) [Accessed 19th October 2018]

College of Policing (2020) *Authorised Professional Practice*. [Online] <https://www.app.college.police.uk/> [Accessed 6<sup>th</sup> May 2019]

Collier, A. (1994) *Critical Realism: An Introduction to Roy Bhaskar's Philosophy*. London: Verso

Collier, P.M. (2006) 'In Search of Purpose and Priorities: Police Performance Indicators in England and Wales'. *Public Money & Management*. 26 (3): 165-172

Commission for Health Improvement (2003) *What CHI has Found in: Ambulance Organisations*. London: The Stationery Office

Connolly, T. and Reb, J. (2003) 'Omission Bias in Vaccination Decisions: Where's the "Omission"? Where's the "Bias"?' *Organizational Behavior and Human Decision Processes*. 91: 186–202

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

Copperfield, D. (2012) *Wasting More Police Time*. Croydon: Monday Books

Corcoran, K., Crusius, J. and Mussweiler, T. (2011) "Social Comparisons: Motives, Standards and Mechanisms". In Chadee, D. (Ed.) *Theories in Social Psychology*. (pp.119-139) Oxford: Wiley-Blackwell

Corley, K. G., and Gioia, D. A. (2004) 'Identity Ambiguity and Change in the Wake of a Corporate Spin-off'. *Administrative Science Quarterly*. 49: 173-208

Cotter, P., Cohen, J. and Coulter, P. (1982) 'Race-of-Interviewer Effects in Telephone Interviews'. *Public Opinion Quarterly*. 46: 278-84

Cottrell, S. (2005) *Critical Thinking Skills: Developing Effective Analysis and Argument*. Basingstoke: Palgrave Macmillan

Coulter, J. (1984) *State of Siege: Miner's Strike 1984. Politics Policing the Coal Field*. London: Canary Press

Courty, P., Heinrich, C., and Marschke, G. (2007) 'Setting the Standard in Performance Measurement Systems'. *International Public Management Journal*. 8 (3): 321-347

Crace, J. (2007) *Rose-Tinted Memoirs*. [Online] <http://www.guardian.co.uk/education/2007/jun/12/schools.education> [Accessed 2nd June 2013]

Creswell, J. (2008) *Research Design: Qualitative and Quantitative Approaches*. (3<sup>rd</sup> Edition). London: Sage

Creswell, J. (2009) *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Thousand Oaks, CA: Sage

Creswell, J. (2013) *What is Mixed Methods Research?* [Online]  
<http://www.youtube.com/watch?v=1OaNiTIpyX8> [Accessed 4<sup>th</sup> December 2013]

Crockett, N. (2013) *Gwent Police Boss Ian Johnston's Concern Over Crime Stats*. [Online]  
[http://www.southwalesargus.co.uk/news/10428840.Gwent\\_Police\\_boss\\_\\_concern\\_over\\_crime\\_stats/](http://www.southwalesargus.co.uk/news/10428840.Gwent_Police_boss__concern_over_crime_stats/) [Accessed 15h October 2013]

Cuddeback, G., Wilson, E., Orme, J. G., and Combs-Orme, T. (2004) 'Detecting and Statistically Correcting Sample Selection Bias'. *Journal of Social Service Research*. 30 (3): 19-33

Cumming, M. (2014) *To What Extent do Performance Targets Influence Policing?* MSc Dissertation. University of Portsmouth. Unpublished

Curtis, I. (2013) *Police Targets – The Debate that Will Not Go Away*. [Online]  
[http://www.policeoracle.com/news/Comment/2013/Sep/23/Police-targets-The-debate-that-will-not-go-away\\_71120.html](http://www.policeoracle.com/news/Comment/2013/Sep/23/Police-targets-The-debate-that-will-not-go-away_71120.html) [Accessed 16th October 2013]

Cyert, R. M. and March, J. G. (1963) *Behavioral Theory of the Firm*. Englewood Cliffs, NJ: Prentice-Hall

Daily Mail (2006) *UK's Worst Police Forces Named. UK's Worst Police Forces Named*. [Online] <http://www.dailymail.co.uk/news/article-412255/UKs-worst-police-forces-named.html> [Accessed 23rd June 2013]

Daly, J., Kellehear, A. and Gliksman, M. (1997) *The Public Health Researcher: A Methodological Approach*. Melbourne: Oxford University Press

Damman, O. C., Hendriks, M., Rademakers, J., Spreeuwenberg, P., Delnoij, D. M. and Groenewegen, P. P. (2011) 'Consumers' Interpretation and use of Comparative Information on the Quality of Health Care: The Effect of Presentation Approaches'. *Health Expect*. 15: 197–211

Danermark, B. (2001) *Interdisciplinary Research and Critical Realism: The Example of Disability Research*. [Online]  
[http://www.criticalrealism.com/archive/iacr\\_conference\\_2001/bdanermark\\_ircr.pdf](http://www.criticalrealism.com/archive/iacr_conference_2001/bdanermark_ircr.pdf) [Accessed 21st September 2015]

Danermark, B., Ekström, M., Jokabsen, L. and Karlsson, J. C. (2002) *Explaining Society: Critical Realism in the Social Sciences*. Abingdon: Routledge

Dayley, D. J. and Mayers, D. (1999) "A Statistical Model for Dynamic Ride-Matching in the World Wide Web". *In: Proceedings of the ITSC'99 Conference*. The IEEE Computer Society, Tokyo, pp. 154-165

- Davis, R. C. (2012a) *Selected International Best Practices in Police Performance Measurement*. Santa Monica: The RAND Centre on Quality Policing
- Davies, H. T. O., Mannion, R., Jacobs, R., Powell, A. E. and Marshall, M. N. (2007) 'Exploring the Relationship between Senior Management Team Culture and Hospital Performance'. *Medical Care Research and Review*. 64 (1): 46-65
- Dawson, D., Gravelle, H., O'Mahony, M., Street, A., Weale, M., Castelli, A., Jacobs, R., Kind, P., Loveridge, P., Martin, S., Stevens, P. and Stokes, L. (2005) *Developing New Approaches to Measuring NHS Outputs and Productivity. Final Report York*: Centre for Health Economics, University of York
- De Bruijn, H. (2002) *Performance Measurement in the Public Sector*. London: Routledge
- De Bruijn, H. (2007) *Managing Performance in the Public Sector*. London: Routledge
- De Fanti, T. A., Brown, M. D. and McCormick, B. H. (1989) 'Visualization: Expanding Scientific and Engineering Research Opportunities'. *Computer*. 22: 12-16
- De Lancer Julnes, P. and Holzer, M. (2001) 'Promoting the Utilization of Performance Measures in Public Organizations: An Empirical Study of Factors Affecting Adoption and Implementation'. *Public Administration Review*. 61 (6): 693-708
- De Maillard, J. and Savage, S. (2012) 'Comparing Performance: The Development of Police Performance Management in France and Britain'. *Policing and Society: An International Journal of Research and Policy*. 22 (4): 363-383
- Deci, Edward L., Richard Koestner and Ryan, R. M. (1999) 'A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation'. *Psychological Bulletin*. 125: 627-68
- Deming, W. E. (1986) *Out of the Crisis*. Cambridge: MIT Press
- Deming, W. E. (1994) *The New Economics for Industry, Government, Education*. (2<sup>nd</sup> Ed) Cambridge: MIT Press
- Denrell, J. (2003) 'Vicarious Learning, Undersampling of Failure, and the Myths of Management'. *Organization Science*. 14: 227-243
- Department for Business Innovation & Skills (2011) *New Challenges, New Chances*. [Online] <http://www.bis.gov.uk/assets/biscore/further-education-skills/docs/f/11-1380-further-education-skills-system-reform-plan> [Accessed 25th June, 2013]
- Department for Education (2012) *Guidance for Local Authorities and Schools on Setting Education Performance Targets for 2012*. [Online] <http://media.education.gov.uk/assets/files/pdf/2/2012%20target%20setting%20guidance.pdf> [Accessed 9<sup>th</sup> August 2013]
- Department of Health (2001) *Department of Health, NHS performance Ratings Acute Trusts 2000/01*. London: Department of Health

- DeSanctis, G. (1984) 'Computer Graphics as Decision Aids: Directions for Research'. *Decision Sciences*. 15 (4): 463-487
- DETR (1999) *Best Value and Audit Commission Performance Indicators for 2000/2001*. Vol. 1. London: HMSO
- De Vaus, D. (2002) *Surveys in Social Research*. (5<sup>th</sup> edn.) London: Routledge
- Dewey, J. (1931) "The Development of American Pragmatism". In H. S. Thayer (Ed.) (1989) *Pragmatism: The Classic Writings* (pp. 23-40). Indianapolis, IN: Hackett
- Di Biase, D.A., MacEachren, M., Krygier, J. B. and Reeves, C. (1992) 'Animation and the Role of Map Design in Scientific Visualisation'. *Cartography and Geographic Information Systems*. 19 (4): 201-214
- Donahue, J. D. (1989) *The Privatization Decision: Public Ends, Private Means*. New York: Basic Books
- Dull, M. (2009) 'Results-Model Reform Leadership: Questions of Credible Commitment'. *Journal of Public Administration Research & Theory*. 19 (2): 255–284
- Dumond, E.J. (1994) Making Best Use of Performance Measures and Information. *International Journal of Operations & Production Management*. 14 (9): 16-31
- Dunleavy, P. and Hood, C. (1994) 'From Old Public Administration to New Public Management'. *Public Money & Management*. 14 (3): 9 - 16
- Durham Police and Crime Commissioner (2018) *Quarter 4 Performance Report (January 2018 to March 2018)* [Online] <https://www.durham-pcc.gov.uk/Document-Library/Performance/2017-18/Q4-Performance-Report.pdf> [Accessed 11th October 2018]
- Earley, P. C., Connolly, T., and Ekegren, G. (1989) 'Goals, Strategy Development, and Task Performance: Some Limits on the Efficacy of Goal Setting'. *Journal of Applied Psychology*. 74: 24-33
- Easterby-Smith, M., Thorpe, R. and Lowe, A. (2002) *Management Research: An Introduction*. London: Sage
- Easton, G. (2010) 'Critical Realism in Case Study Research'. *Industrial Marketing Management*. 39: 118-128
- Eijkenaar, F. (2011) 'Key Issues in the Design of Pay for Performance Programs'. *European Journal of Health and Economics*. 14: 117 - 131
- Eisenhardt, K. M. (1989) 'Agency Theory: An Assessment and Review'. *The Academy of Management Review*. 14 (1): 57-74
- Elbel, B., Gillespie, C. and Raven, M. C. (2014) 'Presenting Quality Data to Vulnerable Groups: Charts, Summaries or Behavioral Economic Nudges'. *Journal. Health Serv. Res. Policy*. 19: 161–168

- Elting, L. S. (1999) 'Influence of Data Display Formats on Physician Investigators' Decisions to Stop Clinical Trials: Prospective Trial with Repeated Measures'. *British Medical Journal*. doi: <http://dx.doi.org/10.1136/bmj.318.7197.1527>
- English, L. D. (2010) 'Young Children's Early Modelling with Data'. *Mathematics Education Research Journal*. 22 (2): 24–47
- Essex Police (2013) *Operation Brightshadow Reduces Burglaries by a Third*. [Online] [http://www.essex.police.uk/news\\_features/features\\_archive/2011/october/operation\\_brightshadow\\_res.aspx](http://www.essex.police.uk/news_features/features_archive/2011/october/operation_brightshadow_res.aspx) [Accessed 3rd July, 2013]
- Eterno, J. A, and Silverman, E. B. (2012) *The Crime Numbers Game: Management by Manipulation*. Boca Raton, FL: CRC Press
- Faber, M., Bosch, M., Wollersheim, H., Leatherman, S. and Grol, R. (2009) 'Public Reporting in Health Care: How do Consumers use Quality-of-Care Information? A Systematic Review'. *Med. Care*. 47: 1–8
- Fagerland, M. W, and Sandvik, L. (2009) 'Performance of Fie Two-Sample Location Tests for Skewed Distributions with Unequal Variances'. *Contemporary Clinical Trials*. 30: 490-496
- Fagerland, M. W. (2012) 't-tests, Non-Parametric Tests, and Large Studies – A Paradox of Statistical Practice?' *BMC Medical Research Methodology*. 12 (78): 1-7
- Fama, E. and Jensen, M. (1983) 'Separation of Ownership and Control'. *Journal of Law and Economics*. 26: 301-325
- Faul, F., Erdfelder, E., Lang, A. G., and Buchner, A. (2007) 'G\*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences'. *Behavior Research Methods*. 39: 175-191
- Fereday, J. and Muir-Cochrane, E. (2006) 'Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development.' *International Journal of Qualitative Methods*. 5 (1): 80-92
- Festinger, L. A. (1954) 'A Theory of Social Comparison Process'. *Human Relations*. 7: 117-140
- Few, S. (2009) *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Berkeley: Analytics Press
- Few, S. (2012) *Show Me the Numbers: Designing Tables and Graphs to Enlighten*. (2<sup>nd</sup> Edition) Burlingame, CA: Analytics Press
- Field, A. (2013) *Discovering Statistics using IBM SPSS Statistics*. London: Sage
- Figlio, D. (2005) *Testing, Crime and Punishment*. NBER Working Paper W11194: March 2005

- Fildes, R. (1985) 'Quantitative Forecasting - The State of the Art: Econometric Models'. *Journal of the Operational Research Society*. 36: 549–580
- Fillichio, C. (2005) 'Getting Ahead of the Curve: Baltimore Citistat'. *Public Manager*. 34: 51 – 53
- Fink, A. (1995) *How to Analyze Survey Data*. London: Sage
- Fish, S., Munro, E. and Bairstow, S. (2008) *Learning Together to Safeguard Children: Developing a Multi-Agency Systems Approach for Case Reviews*. Report 9. London: SCIE
- Fitzgerald, M., Hough, M. Joseph, I. and Qureshi, T. (2002), "Policing for London: Key Findings". Cullompton: Willan (p.xxii)
- Flanagan, R. (2008) *The Review of Policing – Final Report*. London: HMSO
- Fleetwood, S. (2013) *What is (and what isn't) Critical Realism?* [CESR Seminar presentation slides]. Room 2D73 (EDC) 14:30-16:30, 20<sup>th</sup> September 2013. University of the West of England, Bristol
- Fleiss, J. L. (1971) 'Measuring Nominal Scale Agreement among Many Raters'. *Psychological Bulletin*. 76: 378-382
- Fleiss, J. L. and Cohen, J. (1973) 'The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability'. *Educational and Psychological Measurement*. 33: 613–619
- Flin, R. (2003) "'Danger - Men at Work": Management Influence on Safety'. *Human Factors and Ergonomics in Manufacturing & Service Industries*. 13 (4): 261-268
- Forsyth, J. (2020) *Boris's New Target – Cut Violent Crime by 20%*. The Spectator. [Online] <https://www.spectator.co.uk/article/boris-s-new-target-cut-violent-crime-by-20-per-cent> [Accessed 14th March 2020]
- Fortuin, L. (1988) 'Performance Indicators: Why, Where and How?' *European Journal of Operational Research*. 34 (1): 1–9
- Foss, N. J. (2010) 'Micro-Foundations for Management Research: What, Why and Whither?' *Cuadernos de Economía y Dirección de la Empresa*. 42: 11-34. ISSN: 1138-5758
- Foss, N. J. (2014) *Theory-Building Workshop (Day Two)*. [Lecture to PhD students: Presentation slides]. Warwick Business School, University of Warwick. 11<sup>th</sup> November 2014
- Foss, N. J. and Stea, D. (2014) 'Putting a Realistic Theory of Mind into Agency Theory: Implications for Reward Design and Management in Principal-Agent Relations'. *European Management Review*. 11: 101–116

Frakes, N. (2018) *Forces are Setting Targets in Order to Investigate Fewer Crimes*. [Online] <https://www.express.co.uk/news/uk/1028226/police-news-police-set-targets-investigate-fewer-crimes> [Accessed 13th May 2019]

Francis, R. (2013a) *Report of the Mid Staffordshire NHS Foundation Trust Public Inquiry: Executive Summary*. London: The Stationary Office

Francis, R. (2013b) *Report of the Mid Staffordshire NHS Foundation Trust Public Inquiry: Volume 1 – Analysis of Evidence and Lessons Learned*. London: The Stationary Office

Franco-Santos, M. and Bourne, M. (2011) 'The Impact of Performance Targets on Behaviour: A Close Look at Sales Force Contexts'. *Research Executive Summaries Series*. Vol 5: Issue 5. Cranfield. [Online] [http://www.cimaglobal.com/Documents/ImportedDocuments/cid\\_ressum\\_impact\\_performance\\_targets\\_behaviour\\_sales\\_force\\_oct\\_2009.pdf](http://www.cimaglobal.com/Documents/ImportedDocuments/cid_ressum_impact_performance_targets_behaviour_sales_force_oct_2009.pdf) [Accessed 14th August 2013]

Franco-Santos, M., Lucianetti, L. and Bourne, M. (2012) 'Contemporary Performance Measurement Systems: A Review of their Consequences and a Research Agenda'. *Management Accounting Research*. 23 (2): 79-119

Franklin, A. L. (2000) 'An Examination of Bureaucratic Reactions to Institutional Controls'. *Public Performance and Management Review*. 24: 8-21

Franklin, C. A. and Garfield, J. (2006) The GAISE project: developing statistics education guidelines for grades pre-K-12 and college courses. In G. Burrill and P. Elliott (Eds.), *Thinking and Reasoning with Data and Chance, 68th Yearbook*. (pp. 345–376). Reston: National Council of Teachers of Mathematics

Garcia, S. M., Tor, A and Gonzalez, R. 2006. 'Ranks and Rivals: A Theory of Competition'. *Personality and Social Psychology Bulletin*. 32 (7): 970-982

Garrison, J. (1994) 'Deweyan Pragmatism, and Educational Research'. *Educational Researcher*. 23 (1): 5-14

Gal, I. (1997). 'Numeracy: Reflections on Imperatives of a Forgotten Goal'. In L. A. Steen (Ed.). *Quantitative Literacy*. Washington, DC: College Board. 36–43

Gal, I. (2000) "Statistical Literacy: Conceptual and Instructional issues". In Coben, D., O'Donoghue, J. and Fitzsimons G. E. (eds). *Perspectives on Adults Learning Mathematics. Mathematics Education Library, vol 21*. Springer: Dordrecht

Gal, I. (2000) Statistical Literacy: Conceptual and Instructional issues. In Coben, D., O'Donoghue, J. and Fitzsimons G. E. (eds). *Perspectives on Adults Learning Mathematics. Mathematics Education Library, Vol 21*. Springer: Dordrecht

Gal, I. and Ginsburg, L. (1994) 'The Role of Beliefs and Attitudes in Learning Statistics: Towards an Assessment Framework'. *Journal of Statistics Education*. 2 (2): 1-16

- Gay, B. and Weaver, S. (2011) 'Theory Building and Paradigms: A Primer on the Nuances of Theory Construction'. *American International Journal of Contemporary Research*. 1 (2): 24-32
- Gelso, C. J. (2006) "Applying Theories to Research: The Interplay of Theory and Research in Science". In Leong, F. T. and Austin J. T. (Eds.) *The Psychology Research Handbook*. Thousand Oaks, CA: Sage
- Gerteis, M., Gerteis, J. S., Newman, D. and Koepke, C. (2007) 'Testing Consumers' Comprehension of Quality Measures using Alternative Reporting Formats'. *Health Care Finance Review*. 28: 31-45
- Ghobadian, A., Viney, H. and Redwood, J. (2009) 'Explaining the Unintended Consequences of Public Sector Reform'. *Management Decision*. 47 (10): 1514-1535
- Gibbons, R. (1998) 'Incentives in Organizations'. *Journal of Economic Perspectives*. 12: 115 – 132
- Gilbert, D. T., Giesler, R. B., and Morris, K. A. (1995) 'When Comparisons Arise'. *Journal of Personality and Social Psychology*. 69: 227-236
- Gioia, D. A. and Chittipeddi, K. (1991) 'Sensemaking and Sensegiving in Strategic Change Initiation.' *Strategic Management Journal*. 12 (6): 433-448
- Gioia, D. A., Thomas, J. B., Clark, S. M. and Chittipeddi, K. (1994) 'Symbolism and Strategic Change in Academia: The Dynamics of Sensemaking and Influence'. *Organization Science*. 5: 363-383
- Gioia, D. A., Corley, K. G. and Hamilton, A. L. (2012) 'Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology'. *Organizational Research Methods*. (16 (1): 15-31
- Goethals, G. R. (1986) 'Social Comparison Theory: Psychology from the Lost and Found'. *Personality and Social Psychology Bulletin*. 12: 261-278
- Goethals, G. R., and Darley, J. (1977). "Social Comparison Theory: An Attributional Approach". In Suls, J. and Miller, R. L. (Eds.) *Social Comparison Process: Theoretical and Empirical Perspectives*. Washington, DC: Hemisphere
- Golding, B. and Savage, S., (2008) "Leadership and Performance Management". In: Newburn, T. (Ed.) *Handbook of Policing*. Cullompton: Willan. pp. 725 - 759
- Goldstein, H. and Spiegelhalter, D. (1996) 'League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance.' *Journal of the Royal Statistical Society* 159 (3): 385-443
- Goles, T. and Hirschheim, R. (2000) 'The Paradigm is Dead, the Paradigm is Dead...Long Live the Paradigm: The Legacy of Burrell and Morgan'. *Omega* 28 (3): 249-268

- Goodhart, C.A.E. (1975) 'Problems of Monetary Management: The UK Experience'. *Papers in Monetary Economics*. (Volume I) Reserve Bank of Australia
- Gordon, G. G. (1991) 'Industry Determinants of Organizational Culture'. *Academy of Management Review*. 16 (2): 396-415
- Gorski, P. S. (2013) 'What is Critical Realism? And Why Should You Care?' *Contemporary Sociology: A Journal of Reviews*. 42: 658-670
- Grainger, S., Mao, F. and Buytaert, W. (2016) 'Environmental Data Visualisation for Non-scientific Contexts: Literature Review and Design Framework'. *Environmental Modelling & Software*. 85: 299-318
- Granick, D. (1954) *Management of the Industrial Firm in the USSR*. New York: Columbia University Press
- Greater London Authority (2014) *London Data Store. Dashboard Theme: Crime*. [Online] <http://data.london.gov.uk/dashboard-summary/crime/> [Accessed 6<sup>th</sup> December 2014]
- Green, J. and Wintfeld, N. (1995) 'Report Cards on Cardiac Surgeons: Assessing New York State's Approach'. *New England Journal of Medicine*. 332: 1229-1232
- Greenhalgh, S. (2014) *Principles for Policing in Austerity - ACPO Conference speech*, 18<sup>th</sup> June 2014 [Online] <https://www.london.gov.uk/sites/default/files/ACPO%20speech%20SG%20FINAL.pdf> [Accessed 18<sup>th</sup> June 2014]
- Greenwood, C. (2010) 'Theresa May Axes Police Performance Targets'. (The Independent). [Online] <http://www.independent.co.uk/news/uk/home-news/theresa-may-axes-police-performance-targets-2013288.html> [Accessed 8th June 2013]
- Gretton, A. (2013) *Ambulance Watch: Chaos at the N & N as Seventeen Ambulances Stuck in Queue*. [Online] [http://www.edp24.co.uk/news/health/ambulance\\_watch\\_chaos\\_at\\_the\\_n\\_n\\_as\\_17\\_ambulances\\_stuck\\_in\\_queue\\_1\\_1970818](http://www.edp24.co.uk/news/health/ambulance_watch_chaos_at_the_n_n_as_17_ambulances_stuck_in_queue_1_1970818) [Accessed 15th August 2013]
- Greve, H. R. (1998) 'Performance, Aspirations and Risky Organizational Change'. *Administrative Science Quarterly*. 43 (1): 58-86
- Groff R (2004) *Critical Realism: Post-Positivism and the Possibility of Knowledge. Routledge Studies in Critical Realism*. London: Routledge
- Groves, R. (2006) 'Nonresponse Rates and Nonresponse Bias in Household Surveys'. *Public Opinion Quarterly*. 70 (5): 646-67
- Groves, R., and Kahn, R. (1979) *Surveys by Telephone: A National Comparison with Personal Interviews*. New York: Academic Press
- Groves, R. and Fultz, N. (1985) 'Gender Effects among Telephone Interviewers in a Survey of Economic Attitudes'. *Sociological Methods & Research*. 14: 31-52

- Guilfoyle, S. J. (2011) *InspGuilfoyle – Police Inspector and Systems Thinker*. [Online] <https://inspguilfoyle.wordpress.com/> [Accessed 12<sup>th</sup> August 2014]
- Guilfoyle, S. J. (2012) 'On Target? Public Sector Performance Management: Recurrent Themes, Consequences and Questions'. *Policing: A Journal of Policy and Practice*. 6 (3): 250 - 260
- Guilfoyle, S. J. (2013) *Intelligent Policing: How Systems Thinking Methods Eclipse Conventional Management Practice*. Axminster: Triarchy Press
- Guilfoyle, S. J. (2015) 'Binary Comparisons and Police Performance Measurement: Good or Bad?' *Policing: A Journal of Policy and Practice*. 9 (2): 195-209
- Guilfoyle, S. J. (2016) 'Getting Police Performance Measurement Under Control' *Policing: A Journal of Policy and Practice*. 10 (1): 71-87
- Hale, C., Uglow, S., and Heaton, R. (2005) 'Uniform Styles II: Police Families and Policing Styles'. *Policing and Society*. 15 (1): 3 – 20
- Hales, G. (2018) 5<sup>th</sup> June. Available at <https://twitter.com/gmhales/status/1004024675385651201> [Accessed 6th June 2018]
- Hallgren, K. A. (2012) 'Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial'. *Tutor Quant Methods Psychol*. 8 (1): 23-34
- Hamel, G. (1998) 'Strategy Innovation and the Quest for Value'. *Sloan Management Review*. 39 (2): 7–15
- Hamel, G., and Prahalad, C. K. (1993) 'Strategy as Stretch and Leverage'. *Harvard Business Review*. 71 (2): 75–84
- Hammer, M. (2007) 'The 7 Deadly Sins of Performance Measurement (And How to Avoid Them)'. *MIT Sloan Management Review*. Spring 2007: 19 – 28
- Hansen , K. M., Olsen, A. L. and Bech, M. (2015). 'Cross-National Yardstick Comparisons: A Choice Experiment on a Forgotten Voter Heuristic '. *Political Behavior*. 37 (4): 767-789
- Hardin, G. (1968) 'The Tragedy of the Commons'. *Science*. (162): 1243-1248
- Harper, T. (2020) *Police Fear Return of Targets as Price of 20,000 Recruits*. The Times. [Online] <https://www.thetimes.co.uk/article/police-fear-return-of-targets-as-price-of-20-000-recruits-cqvmzfcqp> [Accessed 14th March 2020]
- Harré, R. and Secord, P. F. (1972) *The Explanation of Social Behavior*. Oxford: Blackwell
- Harrison, J., MacGibbon, L., and Morton, M. (2001) 'Regimes of Trustworthiness in Qualitative Research: The Rigors of Reciprocity'. *Qualitative Inquiry*. 7(3): 323-345
- Hatry, H. (1999a) *Performance Measurement: Getting Results*. Washington, DC: Urban Institute Press

Hatry, H. (1999b) 'Mini-Symposium on Intergovernmental Comparative Performance Data'. *Public Administration Review*. 59: 101–104

Hatzivassiloglou, V., and McKeown, K. R. (1997) 'Predicting the Semantic Orientation of Adjectives. *Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL*. (pp. 174-181) New Brunswick, NJ: ACL

Hawley, S.T., Zikmund-Fisher, B., Ubel, B., Jancovic, A., Lucas, T and Fagerlin, A. (2008) 'The Impact of the Format of Graphical Presentation on Health-Related Knowledge and Treatment Choices'. *Patient Educ. Couns.* 73: 448–455

Healthcare Commission (2009) *Investigation into Mid Staffordshire NHS Foundation Trust*. London: HMSO

Heath, C., Larrick, R. P. and Wu, G. (1999) 'Goals as Reference Points'. *Cognitive Psychology*. 38: 79-109

Heckman J. J., Heinrich C. J., and Smith. J. A. (2002) 'The Performance of Performance Standards'. *The Journal of Human Resources*. 37 (4): 778 - 811

Heinen, J. (1985) 'A Primer on Psychological Theory'. *The Journal of Psychology*. 119: 413-421

Heinrich, C. J. (2008) 'Advancing Public Sector Performance Analysis.' *Applied Stochastic Models in Business and Industry*. 24: 373-389

Heinrich, C. J. and Marschke, G. (2010) 'Incentives and Their Dynamics in Public Sector Performance Management Systems'. *Journal of Policy Analysis and Management*. 29 (1): 183 – 208

Henderson, S. G., and Mason, A. J. (2005) "Ambulance Service Planning: Simulation and Data Visualisation". In: Brandeau, M. L., Sainfort, F. and Pierskalla, W. P. (eds) *Operations Research and Health Care. International Series in Operations Research & Management Science, vol 70*. Springer: Boston, MA

Hibbard, J. H., Peters, E., Slovic, P. and Finucane, M. L. (2001) 'Making Health Care Quality Reports Easier to Use'. *Joint Commission Journal on Quality Improvement*. 27: 591–604

Hibbard, J. H., Slovic, P., Peters, E., Finucane, M. L. (2002) 'Strategies for Reporting Health Plan Performance Information to Consumers: Evidence from Controlled Studies'. *Health Serv Res*. 37: 291–313

Hildon, Z., Allwood, D. and Black, N. (2012) 'Making Data More Meaningful: Patients' Views of the Format and Content of Quality Indicators Comparing Health Care Providers'. *Patient Education and Counseling*. 88(2): 298–304

Hill, W. Y. and Milner, M. (2003) 'Guidelines for Graphical Displays in Financial Reporting'. *Accounting Education*. 12 (2): 135-157

- Hirsch, B., Seubert, A and Sohn, M. (2015) 'Visualisation of Data in Management Accounting Reports'. *Journal of Applied Accounting Research*. 16 (2): 221-239
- HMIC (1999a) 1999/2000 Inspection: Suffolk Constabulary. London: HMSO
- HMIC (1999b) *Police Integrity: Securing and Maintaining Public Confidence*. London: HMSO
- HMIC (2008) *Leading From The Frontline*. London: HMSO
- HMIC (2009) *Responsive Policing: Delivering the Policing Pledge*. London: Central Office of Information
- HMIC (2012) *A Step in the Right Direction: The Policing of Anti-Social Behaviour*. London: HMSO
- HMIC (2013a) *Crime and Policing Comparator*. [Online]  
<http://www.hmic.gov.uk/crime-and-policing-comparator/about-the-data/#allcrimes>  
 [Accessed 21<sup>st</sup> July 2013]
- HMIC (2013b) *Technical Notes: How Most Similar Groups are Formed*. [Online]  
<http://www.hmic.gov.uk/media/most-similar-groups-technical-note.pdf> [Accessed 21st July 2013]
- HMIC (2013c) *Crime and Policing Comparator: How to Interpret Charts that use Most Similar Group Comparisons*. [Online]  
<http://www.justiceinspectors.gov.uk/hmic/crime-and-policing-comparator/about-the-data/#peerforces> [Accessed 7<sup>th</sup> June 2015]
- HMIC (2014a) *PEEL 2014 Methodology*. [Online]  
<https://www.justiceinspectors.gov.uk/hmic/wp-content/uploads/2014-peel-methodology.pdf> [Accessed 7th June 2015]
- Hodgkinson, G. P. and Healey, M. P. (2008) 'Toward a (Pragmatic) Science of Strategic Intervention: Design Propositions for Scenario Planning'. *Organization Studies*. 29 (3): 435–457
- Hodgkinson, G. P. and Rousseau, D. M. (2009) 'Bridging the Rigour–Relevance Gap in Management Research: It's Already Happening!' *Journal of Management Studies*. 46 (3): 534-546
- Hodgkinson, G. P. and Starkey, K. (2011) 'Not Simply Returning to the Same Answer Over and Over Again: Reframing Relevance'. *British Journal of Management*. 22: 355–369
- Hoffman, P. J., Festinger, L., and Lawrence, D. H. (1954) 'Tendencies Toward Group Comparability in Competitive Bargaining'. *Human Relations*. 7: 141-159
- Holdaway, S. (1983) *Inside the British Police: A Force In Action*. Oxford: Blackwell
- Holloway, D., Horton, S. and Farnham, D. (1999) "Education". In Horton, S. and Farnham, D. (Eds). *Public Management in Britain*. Basingstoke: Macmillan. pp. 194-212

Holmström, B. (1979) 'Moral Hazard and Observability'. *Bell Journal of Economics*. 10: 74-91

Holmström, B. (1999) 'The Firm as a Mini-Economy'. *Journal of Law, Economics and Organization*. 15: 74-102

Holmström, B. And Milgrom, P. (1991) 'Multitask Principal – Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design'. *Journal of Law, Economics and Organisation*. 7 (1) 24 – 52

Home Office (1983) *Circular 114/83: Manpower, Effectiveness and Efficiency in the Police Service*. London: Home Office

Home Office (1984) *Police and Criminal Evidence Act (PACE): Code of Practice 'A' - Statutory Powers of Stop and Search. (Revised 27<sup>th</sup> October 2013)* [Online] [https://www.gov.uk/government/uploads/system/uploads/attachment\\_data/file/384108/2013PACEcodeA.pdf](https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/384108/2013PACEcodeA.pdf) [Accessed 27th January 2016]

Home Office (1990) *Trends in Crime and their Interpretation*. London: HMSO

Home Office (1993a) *Police Reform*. White Paper Cm2281. London: HMSO

Home Office (1993b) *Inquiry into Police Responsibilities and Rewards*. Cm2280. London: HMSO

Home Office (1993c) *Police Reform: A Police Service for the Twenty-First Century*. London: HMSO

Home Office (1994) *Police and Magistrates Court Act*. London: HMSO

Home Office (1997) *No More Excuses: A New Approach to Tackling Youth Crime in England and Wales*. London: HMSO

Home Office (1999) *Strategic Plan*. London: HMSO

Home Office (2000a) *Best Value Performance Indicators*. London: HMSO

Home Office (2000b) *Crime Targets Published Today: Driving Up Performance – Driving Down Crime*. London: HMSO

Home Office (2001) *Policing a New Century: A Blueprint for Reform*. London: HMSO

Home Office (2002) *Police Reform Act*. London: HMSO

Home Office (2005) *Policing Performance Assessment Framework*. London: HMSO

Home Office (2007) *Assessments of Policing and Community Safety: Strategic Consultation*. London: Home Office

Home Office (2008a) *From the Neighbourhood to the National: Policing Our Communities Together*. London: HMSO

Home Office (2008b) *Improving Performance - A Practical Guide to Police Performance Management*. London: HMSO

Home Office (2008c) *Policing in the 21<sup>st</sup> Century – Volume 1*. London: HMSO

Home Office (2010a) *Policing in the 21st Century: Reconnecting Police and the Public*. London: Home Office

Home Office (2010b) *Performance and Measurement – iQuanta*. [Online]  
<http://tna.europarchive.org/20100419081706/http://www.police.homeoffice.gov.uk/performance-and-measurement/iquanta/index358e.html?version=6> [Accessed 21<sup>st</sup> April 2013]

Home Office (2011) *Police Reform and Social Responsibility Act*. London: HMSO

Home Office (2012b) *Letter to Community Safety Managers*. (Dated 17<sup>th</sup> September 2012). [Online]  
<http://eservices.solihull.gov.uk/mgInternet/Data/Safer%20Solihull%20Board/201210171400/Agenda/Letter%20from%20Home%20Office%20%28126K%20bytes%29%20-%20att37050.pdf> [Accessed 3<sup>rd</sup> May 2015]

Home Office (2015) *The Use of Targets in Policing*. [Online]  
[https://www.gov.uk/government/uploads/system/uploads/attachment\\_data/file/466058/Review\\_Targets\\_2015.pdf](https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/466058/Review_Targets_2015.pdf) [Accessed 3<sup>rd</sup> January 2016]

Hood, C. (1991) 'A Public Management for All Seasons?' *Public Administration*. 69 (1): 3-19

Hood, C. (1995) 'The 'New Public Management' in the 1980s: Variations on a Theme'. *Accounting, Organizations and Society*. 20 (2/3): 93 - 109

Hood, C. (1996) "Exploring Variations in Public Management Reform of the 1980s". In Bekke, H.A., Perry, J.L. and Toonen, T.A. (eds.) *Civil Service Systems in Comparative Perspective*. Bloomington: Indiana University Press

Hood, C. (1998) *The Art of the State: Culture, Rhetoric and Public Management*. Oxford: Oxford University Press

Hood, C. (2006) 'Gaming in Targetworld: The Targets Approach to Managing British Public Services'. *Public Administration Review*. 66 (4): 515-521

Hood, C. (2007) 'Public Service Management by Numbers: Why Does it Vary? Where Has it Come From? What are the Gaps and the Puzzles?' *Public Money & Management*. 27 (2): 95-102

Hood, C., Dixon, R. and Wilson, D. (2009) *'Managing by Numbers': The Way to Make Public Services Better?* Oxford: ESRC Public Services Programme

Hood, C. and Dixon, R. (2010) 'The Political Payoff from Performance Target Systems: No-Brainer or No-Gainer?' *Journal of Public Administration and Theory*. 20 (2): 281-298

- Hoogenboezem, J. and Hoogenboezem, D. (2005) 'Coping With Targets: Performance Measurement in The Netherlands Police'. *International Journal of Productivity and Performance Management*. 54 (7): 568-578
- Hookway, C. (2009) *Pragmatism*. *The Stanford Encyclopedia of Philosophy* [Online] <http://plato.stanford.edu/entries/pragmatism/> [Accessed 18<sup>th</sup> December 2013]
- Hoque, Z. (2014) '20 Years of Studies on the Balanced Scorecard: Trends, Accomplishments, Gaps and Opportunities for Future Research'. *The British Accounting Review*. 46: 33-59
- Hoque, Z, Arends, S and Alexander, R (2004) 'Policing the Police Service - A Case Study of the Rise of "New Public Management" Within an Australian Police Service'. *Accounting, Auditing & Accountability Journal*. 17 (1): 59-84
- Horton, C. and Smith, D. (1988) *Evaluating Police Work*. London: Policy Studies Institute
- Hoskin, T. (2014) *Parametric and Non-Parametric: Demystifying the Terms*. [Online] <http://www.mayo.edu/mayo-edu-docs/center-for-translational-science-activities-documents/berd-5-6.pdf> [Accessed 18th July 2014]
- House, E. R. (1992) 'Response to "Notes on Pragmatism and Scientific Realism"'. *Educational Researcher*. 21 (6): 18-19
- Huber, A. (2016) 'Is Seeing Intriguing? Practitioner Perceptions of Research Documents'. *Journal of Interior Design*. 41 (1): 13-32
- Huber, V., & Neale, M. (1987) 'Effects of Self and Competitor Goals on Performance in an Interdependent Bargaining Task'. *Journal of Applied Psychology*, 72: 197 - 203
- Hughes, O.E. (2003) *Public Management and Administration*, 3<sup>rd</sup> ed. Basingstoke: Palgrave Macmillan
- Hume, D. (1967) *Enquiries Concerning Human Understanding and the Principles of Morals*. Oxford: Clarendon Press
- Hume, L. (1981) *Bentham and Bureaucracy*. Cambridge: Cambridge University Press
- Hunton, P. J., Jones, A. and Baker, P. (2009) 'New Development: Performance Management in a UK Police Force'. *Public Money & Management*. 29 (3): 195 – 200
- Hutcheson, G. and Sofroniou, N. (1999) *The Multivariate Social Scientist*. London: Sage
- Hvidman, U. and Andersen, S. C. (2013) 'Impact of Performance Management in Public and Private Organizations'. *Journal of Public Administration Research and Theory*. 24: 35-58

IBM Corp. (2013). *IBM SPSS Statistics for Windows, Version 22*. Armonk, NY: IBM Corp

IPCC (Independent Police Complaints Commission) (2013) *IPCC Finds Failings in the Working Practices of Southwark Sapphire Unit Between July 2008 and September 2009*. [Online] <http://www.ipcc.gov.uk/news/ipcc-finds-failings-working-practices-southwark-sapphire-unit-between-july-2008-and-september> [Accessed 27th October 2013]

Ittner, C. D. and Larkner, D. F. (2003) 'Coming up Short on Nonfinancial Performance Measures'. *Harvard Business Review*. 81 (11): 88-95

Jackson, A. (2005) 'Falling From a Great Height: Principles of Good Practice in Performance Measurement and the Perils of Top Down Determination of Performance Indicators'. *Local Government Studies*. 31 (1): 21 – 38

Jackson, P. M. (2011) 'Governance by Numbers: What Have We Learned Over The Past 30 Years?' *Public Money and Management*. 31 (1): 13-26

Jacob B. A. (2005) 'Accountability, Incentives and Behavior: Evidence from School Reform in Chicago'. *Journal of Public Economics*. 89 (5–6): 761 - 796

Jacob B. A. and Levitt S. D. (2003) 'Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating'. *The Quarterly Journal of Economics*. CXVIII (3): 843 – 877

Jacobs, R. and Goddard, M. (2007) 'How Do Performance Indicators Add Up? An Examination of Composite Indicators in Public Services.' *Public Money & Management*. 27 (2): 103-110

Jacobsen, C. B., Hvitved, J. and Andersen, L. B. (2014) 'Command and Motivation: How the Perception of External Interventions Relates to Intrinsic Motivation and Public Service Motivation'. *Public Administration*. 92: 790–806

Jacobsen, R., Snyder, J. W. and Saultz, A. (2014) 'Informing or Shaping Public Opinion? The Influence of School Accountability Data Format on Public Perceptions of School Quality'. *American Journal of Education*. 121 (1): 1–27

James, O. (2004) 'The UK Core Executive's Use of Public Service Agreements as a Tool of Governance'. *Public Administration*. 82 (2): 397-419

James, O. (2011) 'Performance Measures and Democracy: Information Effects on Citizens in Field and Laboratory Experiments'. *Journal of Public Administration Research and Theory*. 21 (3): 399–418

James, O. and John, P. (2007) 'Public Management at the Ballot Box: Performance Information and Electoral Support for Incumbent English Local Governments'. *Journal of Public Administration Research and Theory*. 17 (4): 567–580

James, O. and Moseley, A. (2014) 'Does Performance Information about Public Services Affect Citizens' Perceptions, Satisfaction, and Voice Behaviour? Field Experiments with Absolute and Relative Performance Information'. *Public Administration*. 92 (2): 493-511

Jamieson, S. (2004) 'Likert Scales: How to (Ab)use Them'. *Medical Education*. 38: 1217-1218

Jansen, E. P. (2008) 'New Public Management: Perspectives on Performance and the Use of Performance Information'. *Financial Accountability and Management*. 24 (2): 169-191

Jennings, E. T., and Haist, M. P. (2004) "Putting Performance Measurement in Context". In *The Art of Governance: Analyzing Management and Administration*. Ingraham, P. W. and Lynn, L. E. (eds.) Washington, DC: Georgetown University Press

Jick, T. D. (1979) 'Mixing Qualitative and Quantitative Methods: Triangulation in Action'. *Administrative Science Quarterly*. 24: 602-611

Johansson, T., and Siverbo, S. (2009) 'Explaining the Utilization of Relative Performance Evaluation in Local Government: A Multi-Theoretical Study Using Data from Sweden'. *Financial Accountability & Management*. 25 (2): 197-224

Johl, S., K. and Renganathan, S. (2010) 'Strategies for Gaining Access in Doing Fieldwork: Reflection of two Researchers'. *The Electronic Journal of Business Research Methods*. 8 (1): 42-50

Johnson, W. (2013) *Met Sex Crimes Squad 'Pressured Victims to Drop Rape Claims'*. [Online] <http://www.telegraph.co.uk/news/uknews/law-and-order/9894859/Met-sex-crimes-squad-pressured-victims-to-drop-rape-claims.html> [Accessed 27th October 2013]

Joiner, B. (1994) *Fourth Generation Management*. New York: McGraw-Hill

Jones, T. (2003) "The Governance and Accountability of Policing". In: Newburn, T. (Ed) *Handbook of Policing*. Cullompton: Willan. pp. 603 – 627

Jowett, P. and Rothwell, M. (1988) *Performance Indicators in the Public Sector*. London: Macmillan

Jüni, P., and Egger, M. (2005) 'Empirical Evidence of Attrition Bias in Clinical Trials'. *International Journal of Epidemiology*. 34 (1): 87-88

Kahn, W. A. (2012) 'The Functions of Dysfunction: Implications for Organizational Diagnosis and Change'. *Consulting Psychology Journal: Practice and Research*. 64 (3): 225-241

Kahneman, D. (1992) 'Reference Points, Anchors, Norms and Mixed Feelings'. *Organizational and Human Decision Processes*. 51: 296-312

Kahneman, D. (2011) *Thinking Fast and Slow*. London: Penguin

Kahneman, D., and Tversky, A. (1979) 'Prospect Theory: An Analysis of Decision Under Risk'. *Econometrica: Journal of the Econometric Society*. 47 (2): 263-291

- Kalpana, C. (2011) *How Many People do I Need to Take My Survey?* [Online] <https://www.surveymonkey.com/blog/en/blog/2011/09/15/how-many-people-do-i-need-to-take-my-survey/> [Accessed 23<sup>rd</sup> February 2014]
- Kanfer, R., and Ackerman, P. L. (1989). 'Motivation and Cognitive Abilities: An Integrative Aptitude Treatment Interaction Approach to Skill Acquisition'. *Journal of Applied Psychology*. 74: 657 – 690
- Kantrowitz, B. and Springen, K. (1997) *Why Johnny Stayed Home*. Newsweek, 6th October 1997. p. 6
- Kaplan, A. (1964) *The Conduct of Inquiry*. New York: Harper and Row
- Kaplan, R. and Norton, D. (1996) *The Balanced Scorecard*. Boston: Harvard Business School Press
- Kelley, K., Clark, B., Brown, V. and Sitzia, J. (2003) 'Good Practice in the Conduct and Reporting of Survey Research'. *International Journal for Quality in Healthcare*. 15 (3): 261-266
- Kelling, G. L., and Bratton, W. J. (1998) 'Declining Crime Rates: Insiders' Views on the New York City Story'. *The Journal of Criminal Law and Criminology*. 88: 1217 - 1231
- Kelloway, E. K., and Barling, J. (2010) 'Leadership Development as an Intervention in Occupational Health Psychology'. *Work & Stress*. 24: 260-279
- Kelman, S. and Friedman, J. N. (2009) 'Performance Improvement and Performance Dysfunction: An Empirical Examination of Distortionary Impacts of the Emergency Room Wait-Time Target in the English National Health Service'. *Journal of Public Administration and Theory*. 19: 917 – 946
- Kent Police (2013a) *Review of Performance Culture: Officer Interviews*. Unpublished
- Kent Police (2013b) *Crime Detection and Performance Culture Review*. Unpublished
- Kerlinger, F. N., and Lee, H. B. (2000) *Foundations of Behavioral Research*. Fort Worth, TX: Harcourt
- Kerr, S. (1975) 'On the Folly of Rewarding A, While Hoping for B'. *Academy of Management Journal*. 18 (4): 769-783
- Kessler, I. and Purcell, J. (1992) 'Performance Related Pay: Objectives and Application'. *Human Resource Management Journal*. 2 (3):16-33
- Ketokivi, M. and Mantere, S. (2010) 'Two Strategies for Inductive Reasoning in Organizational Research'. *Academy of Management Review*. 35 (2): 315-333
- Kitzinger, J. (1994) 'The Methodology of Focus Groups: The Importance of Interaction Between Research Participants'. *Sociology of Health and Illness*. 16: 103–121

- Kline, P. (1994) *An Easy Guide to Factor Analysis*. London: Routledge
- Kline, P. (1999) *The Handbook of Psychological Testing*. London: Routledge
- Knapp, H. (2014) *Introductory Statistics Using SPSS*. Thousand Oaks, CA: Sage
- Knapp, T. R. (1990) 'Treating Ordinal Scales as Interval Scales: an Attempt to Resolve the Controversy'. *Nursing Research*. 39: 121–123
- Koch, T. (1994) 'Establishing Rigour in Qualitative Research: The Decision Trail'. *Journal of Advanced Nursing*. 19: 976-986
- Koretz, D. (2002) 'Limitations in the Use of Achievement Tests as Measures of Educators' Productivity'. *The Journal of Human Resources*. 37 (4): 752 - 777
- Kosslyn, S. M. (1994) *Elements of Graph Design*. New York: W. H. Freeman & Co.
- Kramer, R.M. (2009) *Organizational Trust: A Reader*. New York: Oxford University Press
- Kroll, A. (2013) 'Explaining the Use of Performance Information by Public Managers: A Planned-Behavior Approach'. *American Review of Public Administration*. 20 (10): 1-15
- Kroll, A. and Vogel, D. (2013) 'The PSM-Leadership Fit: A Model of Performance Information Use'. *Public Administration*. 92 (4): 974-991
- Kroll, A. (2014) *Drivers of Performance Information Use: Systematic Literature Review and Directions for Future Research*. [Online]  
<http://www.lafollette.wisc.edu/images/publications/PIP/Kroll-2014-PPMR.pdf>  
 [Accessed 8th March 2015]
- Kroll, A. (2015) *Performance Information Use, Outcome Effects, and Social Equity: Probing the Links*. Conference Paper: Submission #15609. Accepted by the 2015 Academy of Management Annual Meeting
- Kroll, A. (2017) 'Can Performance Management Foster Social Equity? Stakeholder Power, Protective Institutions, and Minority Representation'. *Public Administration*. 95 (1): 22-38
- Kroll, A. and Moynihan, D. P. (2017) 'The Design and Practice of Integrating Evidence: Connecting Performance Management with Program Evaluation'. *Public Administration Review*. 78 (2): 183-194
- Kroll, A., Neshkova, M. I. and Pandey, S. K. (2019) 'Spillover Effects from Customer to Citizen Orientation: How Performance Management Reforms Can Foster Public Participation'. *Administration & Society*. 51(8): 1227-1253
- Kurtzman, E. T. and Greene, J. (2016) 'Effective Presentation of Health Care Performance Information for Consumer Decision-making: A Systematic Review'. *Patient Education and Counseling*. 99: 36–43

- Kuzon, W. M. Jr., Urbanchek, M. G. and McCabe, S. (1996) 'The Seven Deadly Sins of Statistical Analysis'. *Annals of Plastic Surgery*. 37: 265-72
- Landis, J., Sullivan, D. and Sheley, J. (1973) 'Feminist Attitudes as Related to Sex of the Interviewer'. *Pacific Sociological Review*. 16: 305-314
- Landis, J. R. and Koch, G. G. (1977) 'The Measurement of Observer Agreement for Categorical Data'. *Biometrics*. 33: 159-174
- Larsen, M. V. and Olsen, A. L. (2020) 'Reducing Bias in Citizens' Perception of Crime Rates: Evidence from a Field Experiment on Burglary Prevalence'. *The Journal of Politics*. 82 (2). Published online January 16, 2020. <https://doi.org/10.1086/706595>
- Latham, G. P. (2004) 'The Motivational Benefits of Goal-Setting'. *Academy of Management Executive*. 18 (4): 126-129
- Lau, R. R. (1985) 'Two Explanations for Negativity Effects in Political Behavior'. *American Journal of Political Science*. 29 (1): 119-138.
- Laubschagne, A. (2003) 'Qualitative Research - Airy Fairy or Fundamental?' *The Qualitative Report*. 8 (1): 100-103
- Lavertu, S. and Moynihan, D. P. (2012) 'The Empirical Implications of Theoretical Models: A Description of the Method and an Application to the Study of Performance Management Implementation'. *Journal of Public Administration Research and Theory*. 23: 333-260
- Laville, S. (2012) *Kent Police Officers Arrested Over Crime Statistics 'Irregularities'*. [Online] <http://www.theguardian.co.uk/2012/nov-15/kent-police-arrested-statistics-irregularities> [Accessed 29th September 2013]
- Lavrakas, P. (2008) *Encyclopaedia of Survey Research Methods*. Thousand Oaks, CA
- Lawson, T. (1997) *Economics and Reality*. London: Routledge
- Lawson, T. (1998) "Economic Science without Experimentation". In Archer, M. et al (Eds) *Critical Realism: Essential Readings*. London: Routledge
- Layder, D. (2003) *New Strategies in Social Research*. Cambridge: Polity Press
- Lazard, A. and Atkinson, L. (2015) 'Putting Environmental Infographics Center Stage: The Role of Visuals at the Elaboration Likelihood Model's Critical Point of Persuasion'. *Science Communication*. 37 (1): 6-33
- Le Grand, J. (2010) 'Knights and Knaves Return: Public Service Motivation and the Delivery of Public Services'. *International Public Management Journal*. 13 (1): 56-71

- Le Meunier-FitzHugh, K., Massey, G. R. and Piercy, N. F. (2011) 'The Impact of Aligned Rewards and Senior Manager Attitudes on Conflict and Collaboration between Sales and Marketing'. *Industrial Marketing Management*, 40 (7): 1161-1171
- LeRoux, K. and Wright, N. S. (2010) 'Does Performance Measurement Improve Strategic Decision Making? Findings from a National Survey of Nonprofit Social Service Agencies'. *Nonprofit and Voluntary Sector Quarterly*. 39 (4): 571-587
- Leapman, B. (2007) *Leaks Reveal Police Target Soft Touches to Meet Detection Rates*. The Telegraph. [Online] <http://www.telegraph.co.uk/news/uknews/1550045/Leaks-reveal-police-target-soft-touches-to-hit-detection-rates.html> [Accessed 29<sup>th</sup> July 2013]
- Leckie, G. and Goldstein, H. (2009) 'The Limitations of Using School League Tables to Inform School Choice'. *Journal of the Royal Statistical Society, (A)*. 172 (4): 835–851
- Legard, R., Keegan, J. and Ward, K. (2003) "In-depth Interviews". In Ritchie, J. and Lewis, J. (eds) *Qualitative Research Practice: A Guide for Social Science Students and Researchers*. pp. 139–169. London: Sage Publications
- Lehmann, E. L. (1998) *Nonparametrics: Statistical Methods Based on Ranks*. NJ: Prentice-Hall
- Leicestershire Police Authority (2010) *Three Year Policing Plan 2010-2013*. [Online] <http://www.leics.police.uk/media/uploads/library/file/your-policeV9.pdf> [Accessed 14th October 2013]
- Leicestershire Police (2014) *Police Officer Establishment*. Leicestershire Police: Unpublished. Data retrieved by Chief Inspector Sally Chivers, 9<sup>th</sup> December 2014
- Leishman, F, Cope, S. and Starie, P. (1995) 'Reforming the Police in Britain. New Public Management, Policy Networks and a Tough "Old Bill."' *International Journal of Public Sector Management*. 8 (4): 26-37
- Levitt, B., and March, J. G. (1988) 'Organizational Learning'. *Annual Review of Sociology*. 14: 319–340
- Lewis, P. A. (2002) 'Agency, Structure and Causality in Political Science: A Comment on Sibeon'. *Politics*. 22 (1): 17-23
- Lewis, P. (2010) *Blair Peach Killed by Police at 1979 Protest, Met Report Finds*. [Online] <http://www.theguardian.com/uk/2010/apr/27/blair-peach-killed-police-met-report> [Accessed 28th April 2014]
- Likert, R. (1932) *A Technique for the Measurement of Attitudes*. *Archives of Psychology*. 140: 1–55
- Likierman A. (1993) 'Performance Indicators: 20 Early Lessons from Managerial Use'. *Public Money & Management*. 13 (4): 15 – 22

- Linder, S. and Foss, N. J. (2013) *Agency Theory*. SMG Working Paper no. 7/2013, April 2013. Copenhagen: Department of Strategic Management and Globalization, Copenhagen Business School. ISBN: 978-87-91815-89-8
- Lindquist, E. (2011) *Surveying the World of Visualization*. Canberra: Australian National University
- Lipscomb, M. (2011) *Critical Realism and Realist Pragmatism in Mixed Methods: Problematics of Event Identity and Abductive Inference*. Conference Paper: Professors of Education Research Symposium: Evolving Paradigms in Mixed Methods Research. American Education Research Association, New Orleans, Louisiana. 8<sup>th</sup>-12<sup>th</sup> April 2011
- Litwin, M. S. (1995) *How to Measure Survey Reliability and Validity*. London: Sage
- Liu, W. B., Cheng, Z. L., Mingers, J., Qi, L. and Meng, W. (2010) 'The 3E Methodology for Developing Performance Indicators for Public Sector Organizations'. *Public Money & Management*. 30 (5): 305–312
- Locke, E. A. (1968) 'Toward a Theory of Task Motivation and Incentives'. *Organizational Behavior and Human Performance*. 3: 157 – 189
- Locke, E. A. (2001) "Motivation by Goal Setting". In Golembiewski, R.T. (Ed.) *Handbook of Organizational Behavior*. (pp. 43-56) New York: Marcel Dekker, Inc
- Locke, E. A., Shaw, K. N., Saari, L. M., and Latham, G. P. (1981) 'Goal-Setting and Task Performance: 1969–1980'. *Psychological Bulletin* 90: 125–152
- Locke, E. A., Chah, D., Harrison, S., and Lustgarten, N. (1989) 'Separating the Effects of Goal Specificity from Goal Level'. *Organizational Behavior and Human Performance*. 43: 270-287
- Locke, E. A. and Latham, G. P. (1984) *Goal Setting: A Motivational Technique that Works*. Prentice Hall: Englewood Cliffs, NJ
- Locke, E. A. and Latham, G. P. (1990) *A Theory of Goal Setting and Task Performance*. Prentice Hall: Englewood Cliffs, NJ
- Locke, E. A., and Latham, G. P. (1991) 'Self-Regulation through Goal Setting'. *Organizational Behavior and Human Decision Processes*. 50: 212–247
- Locke, E. A. and Latham, G. P. (2002) 'Building a Practically Useful Theory of Goal Setting and Task Motivation'. *American Psychologist* 57: 705 – 717
- Locke, E. A. and Latham, G. P. (2006) 'New Directions in Goal-Setting Theory'. *Association for Psychological Science*. 15 (5): 265-268
- Locke, E. A. and Latham, G. P. (2009) 'Has Goal Setting Gone Wild, or Have Its Attackers Abandoned Good Scholarship?' *Academy of Management Perspectives*. 23 (1): 17 - 23
- Locke, K. (2001) *Grounded Theory in Management Research*. Thousand Oaks, CA: Sage

Loftus, B. (2010) *Police Culture in a Changing World*. Oxford: Oxford University Press

Lohman, C., Fortuin, L. and Wouters, M. (2004) 'Designing a Performance Measurement System: A Case Study'. *European Journal of Operational Research*. 156: 267-286

Longman, H. (2013) *A&E Has a Mountain to Climb – But Which One?* [Online] <http://www.patient-access.org.uk/wordpress/wp-content/uploads/2013/06/SIPR-Charts2.pdf> [Accessed 14th August 2013]

Longshore-Smith, M. (2006) 'Overcoming Theory-Practice Inconsistencies: Critical Realism and Information Systems Research'. *Information and Organization*. 16: 191–211

Losby, J. and Wetmore, A. (2012) Using Likert Scales in Evaluation Survey Work. [Online] [http://www.cdc.gov/dhdsp/pubs/docs/CB\\_February\\_14\\_2012.pdf](http://www.cdc.gov/dhdsp/pubs/docs/CB_February_14_2012.pdf) [Accessed 11th March 2013]

Loveday, B. (1998) 'Improving the Status of Police Patrol'. *International Journal of Sociology of Law*. 26: 161 – 198

Loveday B. (1999) *The Impact of Performance Culture on Criminal Justice Agencies: The Case of the Police and the Crown Prosecution Service*. Occasional Paper No 9. January 1999. Institute of Criminal Justice Studies. University of Portsmouth

Loveday, B. (2000) 'Managing Crime: Police Use of Crime Data as an Indicator of Effectiveness.' *International Journal of the Sociology of Law*, 28: 215-237

Loveday, B. (2005) 'The Challenge of Police Reform in England and Wales'. *Public Money and Management*. 25(5): 275–281

Loveday, B. (2006) 'Policing Performance: The Impact of Performance Measures and Targets on Police Forces in England and Wales'. *International Journal of Police Science and Management*. 8 (4): 282–293

Loveday, B. (2008) 'Performance Management and the Decline of Leadership within Public Services in the United Kingdom'. *Policing*. 2 (1): 120–130

Loveday, B., Button, M., Fletcher, R., and Blackburn, D. (2004). *Isle of Wight, Safer Communities Partnership Board, Crime and Drugs Audit 2004*. Portsmouth: University of Portsmouth

Lowe, T. (2013) 'New Development: The Paradox of Outcomes - The More We Measure, The Less We Understand'. *Public Money & Management*. 33 (3): 213-216

Ltifi, H., Ben Ayed, M., Alimi, A. M. and Lepreux, S. (2009) "Survey of information visualization techniques for exploitation in KDD". *IEEE/ACS International Conference on Computer Systems and Applications*. Rabat, 2009, pp. 218-225. doi: 10.1109/AICCSA.2009.5069328

- Lusk, E. J. and Kersnick, M. (1979) 'Effects of Cognitive Style and Report Format on Task Performance: The MIS Design Consequences'. *Management Science*. 25: 787-798
- Mace, C. A. (1935) 'Incentives: Some Experimental Studies'. *Industrial Health Research Board Report No 72. Medical Research Council*. London: HMSO
- Machamer, P., Darden, L. and Craver, C. F. (2000) 'Thinking about Mechanisms'. *Philosophy of Science*. 67: 1-25
- Mackenzie, S. and Hamilton-Smith, N. (2010) 'Measuring Police Impact on Organised Crime: Performance management and Harm Reduction'. *Policing: An International Journal of Police Strategies & Management*. 34 (1): 7-30
- Manning, P. K. (1977) *Police Work: the Social Organization of Policing*. Cambridge, MA: MIT Press
- Mannion R. and Braithwaite, J. (2012) 'Unintended Consequences of Performance Measurement in Healthcare: 20 Salutory Lessons from the English National Health Service'. *Internal Medical Journal*. 5: 569–574
- March, J. G. (1988) 'Variable Risk Preferences and Adaptive Aspirations'. *Journal of Economic Behavior and Organizations* 9: 5-24
- March, J. G. (1991) 'Exploration and Exploitation in Organizational Learning'. *Organization Science*. 2: 71–87
- Marks, M. (2004) 'Researching Police Transformation: The Ethnographic Imperative'. *The British Journal of Criminology*. 44 (6): 866-888
- Markus, M. L. (1997) "The Qualitative Difference in Information Systems Research and Practice". In: Lee, A., Liebenau, J, and De Gross, J. (eds.) *Information Systems and Qualitative Research*. London: Chapman & Hall, 1997. (pp.11-27)
- Marlowe, H. A. Jnr. (2000) *Identifying and Controlling for Sources of Bias & Error in Focus Group Assessment Research: Working paper*. [Online] <http://analyticaconsulting.co/wp-content/uploads/2012/02/Identifying-and-Controlling-for-Sources-of-Bias-in-focus-group-research.pdf> [Accessed 12th March 2016]
- Marshall, G. (1998) *Interviewer Bias*. [Online] <http://www.encyclopedia.com/doc/1O88-interviewerbias.html> [Accessed 12th March 2016]
- Marshall, M., Shekelle, P., Brook, R. and Leatherman, S. (2000) *Dying to Know: Public Release of Information about Quality of Care*. London: The Nuffield Trust
- Marshall, T., Mohammed, M. A. and Rouse, A. A. (2004) 'A Randomized Controlled Trial of League Tables and Control Charts as Aids to Health Service Decision-Making'. *International Journal for Quality in Health Care*. 16(4): 309–315
- Maurer, R. (2014) 'The Influence of Senior Leaders in Successful Change'. *The Journal for Quality and Participation*. 37 (2): 435-451

- Mastrofski, S. D. (1999) 'Policing for People'. [Online] <http://www.policefoundation.org/pdf/Mastrofski.pdf> *Ideas in American Policing*. Police Foundation, March 1999. [Accessed 16<sup>th</sup> November 2013]
- May, T. (2015) *Speech given by Home Secretary Theresa May at Police Federation Annual Conference 2015*. [Online] <https://www.gov.uk/government/speeches/home-secretarys-police-federation-2015-speech> [Accessed 26th May 2015]
- McDermott, J. (2013) *Home Sec Blasts Target Culture Comeback*. [Online] [http://www.policeoracle.com/news/HR,+Personnel+and+Staff+Development/2013/September/10/Home-Sec-blasts-target-culture-comeback\\_70575.html](http://www.policeoracle.com/news/HR,+Personnel+and+Staff+Development/2013/September/10/Home-Sec-blasts-target-culture-comeback_70575.html) [Accessed 21st September 2013]
- McLean, I., Haubrich, D. and Gutierrez-Romero, R. (2007) 'The Perils and Pitfalls of Performance Measurement: The CPA Regime for Local Authorities in England'. *Public Money & Management*. 27 (2): 111 – 118
- Melkers, J. and Willoughby, K. (2005). 'Models of Performance Measurement Use in Local Governments: Understanding Budgeting, Communication, and Lasting Effects'. *Public Administration Review*. 65: 180-90
- Menand, L. (ed.) (1998) *Pragmatism*. New York: Random House
- Merton, R. K. (1967) *On Theoretical Sociology*. New York: Free Press
- Metropolitan Police (2005) *Policing and Performance Plan 2005/06*. [Online] [http://www.met.police.uk/foi/pdfs/priorities\\_and\\_how\\_we\\_are\\_doing/corporate/mps\\_policing\\_plan\\_2005-06.pdf](http://www.met.police.uk/foi/pdfs/priorities_and_how_we_are_doing/corporate/mps_policing_plan_2005-06.pdf) [Accessed 2nd August 2013]
- Metropolitan Police (2013a) *CrimeFighters Performance Document*. Unpublished: Metropolitan Police
- Metropolitan Police (2013b) *Stop and Search – Freedom of Information Request QEP8-001 STOP IT Monthly Data – August 2013*. [Online] [http://www.met.police.uk/foi/pdfs/priorities\\_and\\_how\\_we\\_are\\_doing/corporate/stop-search\\_kpi\\_august2013.pdf](http://www.met.police.uk/foi/pdfs/priorities_and_how_we_are_doing/corporate/stop-search_kpi_august2013.pdf) Metropolitan Police, TP Patrol OCU, Stop and Search Team: 08/09/2013 [Accessed 26<sup>th</sup> January 2014]
- Metropolitan Police Federation (2014) [Online] *The Consequences of a Target-Driven Culture in Policing: From the Voices of Metropolitan Police Officers*. <https://www.metfed.org.uk/support/uploads/1396550032target%20report%20final.pdf> [Accessed 18<sup>th</sup> April 2015]
- Metropolitan Police (2018) *Crime Data Dashboard*. [Online] <https://www.met.police.uk/sd/stats-and-data/met/crime-data-dashboard/> [Accessed 11<sup>th</sup> October 2018]
- Micheli, P. (2012a) *Strategy into Action*. [Online] <http://www.i-b-e.co.uk/media/files/04-strategy-into-action.pdf> Cranfield: Cranfield Business School [Accessed 27th August 2014]

- Micheli, P. (2012b) *The Seven Myths of Performance Management*. [Online] <http://www.wbs.ac.uk/news/the-seven-myths-of-performance-management/> Warwick: Warwick Business School [Accessed 26<sup>th</sup> April 2014]
- Micheli, P. and Manzoni, J. F. (2010) 'Strategic Performance Measurement: Benefits, Limitations and Paradoxes'. *Strategic Performance Measurement, Long Range Planning*. 43 (4): 465-476
- Micheli, P. and Mari, L. (2014) 'The Theory and Practice of Performance Measurement'. *Management Accounting Research*. 25 (2): 147-156
- Micheli, P. and Neely, A. (2010) 'Performance Measurement in the Public Sector in England: Searching for the Golden Thread'. *Public Administration Review*. July / August 2010: 592-600
- Micheli, P. and Pavlov, A. (2017) 'The Interplay of Purposeful and Passive Uses of Performance Information within Organisations'. *Draft paper for submission to Public Administration*
- Miller, D. T. (1983, October). *Presentation given at the Ontario Symposium on Relative Deprivation and Social Comparison*. London: Ontario
- Miller, K. D. and Tsang, E. W. K. (2010) 'Testing Management Theories: Critical Realist Philosophy and Research Methods'. *Strategic Management Journal*. 32: 139-158
- Miller, K. L., and Smith, L. E. (1983) 'Handling Nonresponse Issues'. *Journal of Extension*. 21 (5): 45-50
- Mingers, J. (2001) 'Combining IS Research Methods: Towards a Pluralist Methodology'. *Information Systems Research* 12 (3): 240-259
- Mingers, J. (2002) 'Real-izing Information Systems: Critical Realism as an Underpinning Philosophy for Information Systems'. *ICIS 2002 Proceedings*. Paper 27. <http://aisel.aisnet.org/icis2002/27>
- Mingers, J. (2003) "Future Directions in Management Science Modeling: Critical Realism and Multimethodology". In Fleetwood, S. and Ackroyd, S. (eds.) *Critical Realism in Action in Organizations and Management Studies*. London: Routledge
- Mingers, J. (2006) 'A Critique of Statistical Modelling in Management Science from a Critical Realist Perspective: Its Role Within Multimethodology'. *Journal Operational Research Society*. 55 (2): 202-219
- Mingers, J., Mutch, A. and Willcocks, L. (2013) 'Critical Realism in Information Systems Research'. *MIS Quarterly*. 37 (3): 795-802
- Monro, A. (2008) *Police Focus on Minor Crimes to Meet Targets* [Online] <http://www.timesonline.co.uk/tol/news/uk/article4033441.ece> [Accessed 18th June 2013]

Moore, M. and Braga, A. (2003) 'Measuring and Improving Police Performance: The Lessons of Compstat and its Progeny'. *Policing: An International Journal of Police Strategies and Management*. 26 (3): 439-453

Moore, D. A. and Klein, W. M. P. (2008) 'Use of Absolute and Comparative Performance Feedback in Absolute and Comparative Judgments and Decisions'. *Organizational Behavior and Human Decision Processes*. 107 (1): 60-74

MOPAC (2012) *MOPAC Performance Challenge*. [Online]  
<http://www.london.gov.uk/sites/default/files/MOPAC%20Challenge%20presentation%20-%202-10-12%20%5BCompatibility%20Mode%5D.pdf> [Accessed 9<sup>th</sup> August 2013]

MOPAC (2013) *MOPAC Stop and Search Working Group. Minutes – 11<sup>th</sup> July 2013*. [Online] <http://www.london.gov.uk/moderngov/documents/b9012/Minutes%20-%20Transcript%20-%20Appendix%201%20Thursday%2011-Jul-2013%2014.00%20Stop%20and%20Search%20Working%20Group.pdf?T=9> [Accessed 17<sup>th</sup> October 2013]

Morgan, D. L. (1998) *The Focus Group Guide Book*. London: Sage

Morse, S. and Gergen, K. J. (1970) 'Social Comparison, Self-Consistency and the Concept of Self'. *Journal of Personality and Social Psychology*. 16: 148-156

Moss, Q. Z., Alho, J. and Alexander, K. (2007) 'Performance Measurement Action Research'. *Journal of Facilities Management*. 5 (4): 290-300

Mountford, J. and Wakefield, D. (2016) 'From Stoplight Reports to Time Series: Equipping Boards and Leadership Teams to Drive Better Decisions'. *BMJ Quality & Safety*. 26: 9-11

Moyle, P. (1998) 'Longitudinal Influences of Managerial Support on Employee Well-being'. *Work & Stress*. 12 (1): 29-49

Moynihan, D. P. (2005) 'Goal-Based Learning and the Future of Performance Management'. *Public Administration Review*. 65 (2): 203–216

Moynihan, D. P. (2008) *The Dynamics of Performance Management*. Washington DC: Georgetown University Press

Moynihan, D. P. (2009) 'Through a Glass, Darkly: Understanding the Effects of Performance Regimes'. *Public Performance & Management Review*. 32(4): 592–603

Moynihan, D. P. (2010) 'A Workforce of Cynics? The Effects of Contemporary Reform on Public Service Motivation'. *International Public Management Journal*. 13: 24–34

Moynihan, D. P. (2015) 'Uncovering the Circumstances of Performance Information Use'. *Public Performance and Management Review*. 39 (1): 33–57

Moynihan, D. P. (2016a) Email to Simon Guilfoyle, dated 6<sup>th</sup> June 2016.

Moynihan, D. P. (2016b) 'Political Use of Performance Data'. *Public Money & Management*. 36 (7): 479-481

Moynihan, D. P., Baekgaard, M. and Jakobsen, M. L. (2019) 'Tackling the Performance Regime Paradox: A Problem-Solving Approach Engages Professional Goal-Based Learning'. *Public Administration Review*.  
<https://doi.org/10.1111/puar.13142>

Moynihan, D. P. and Hawes, D. (2012) 'Responsiveness to Reform Values: The Influence of the Environment on Performance Information Use'. *Public Administration Review*. 72 (s1): 95–105

Moynihan, D. P. and Ingraham, P. W. (2004) 'Integrative Leadership in the Public Sector: A Model of Performance Information Use'. *Administration and Society*. 36 (4): 427-453

Moynihan, D. P. and Landuyt, N. (2009) 'How Do Public Organizations Learn? Bridging Cultural and Structural Perspectives'. *Public Administration Review*. 69: 1097-1105

Moynihan, D. P. and Lavertu, S. (2011) *Does Involvement in Performance Management Routines Encourage Performance Information Use? Evaluating GPRA and PART*. La Follette School Working Paper No 2011-017. Wisconsin: University of Wisconsin-Madison

Moynihan, D. P. and Kroll, A. (2016) 'Performance Management Routines that Work? An Early Assessment of the GPRA Modernization Act'. *Public Administration Review*. 76 (2): 314-323

Moynihan, D. P., Kroll, A. and Neilsen, P. A. (2016) Managerial Use of Performance Data by Bureaucrats and Politicians. (Draft chapter sent to Simon Guilfoyle by email on 22<sup>nd</sup> May 2016. [In press]

Moynihan, D. P., Nielsen, P. A. and Kroll, A. (2017) "Managerial Use of Performance Data by Bureaucrats and Politicians". *In Experiments in Public Management Research* (pp. 244-269). Cambridge: Cambridge University Press

Moynihan, D. P. and Pandey, S. K. (2005) 'Testing How Management Matters in an Era of Government by Performance Management'. *Journal of Public Administration Research and Theory*. 15: 421–39

Moynihan, D. P. and Pandey, S. K. (2010) 'The Big Question for Performance Management: Why do Managers Use Performance Information?' *Journal of Public Administration Research and Theory*. 20 (4): 849-66

Moynihan, D. P., Pandey, S. K. and Wright, B. E. (2009) 'Leadership and Reform: Mapping the Causal Pathways of Performance Information Use'. Paper presented at the 10th Public Management Research Conference, Columbus, Ohio, October 1-3, 2009.

- Moynihan, D. P. and Soss, J. (2014) 'Policy Feedback and the Politics of Administration'. *Public Administration Review*. 74: 320–332
- Mulley, S. (2012) *Stop Playing the Net Migration Numbers Game*. [Online] <http://www.theguardian.com/commentisfree/2012/aug/30/stop-playing-net-migration-numbers-game> [Accessed 18th August 2013]
- Munro, E. (2005) 'A Systems Approach to Investigating Child Abuse Deaths.' *British Journal of Social Work*. 25: 531–46
- Murfitt, N. (2016) *Ambulance Services are Accused of 'Fiddling' 999 Figures to Make it Look Like they are Quicker at getting to Emergencies*. [Online] <http://www.dailymail.co.uk/news/article-3446094/Ambulance-services-accused-fiddling-999-figures-avoid-fines.html> [Accessed 14th February 2016]
- Murray, K. (2014) *The Proactive Turn: Stop and Search in Scotland. (A Study in Elite Power)*. Doctoral Thesis. Edinburgh: University of Edinburgh.
- Nasukawa, T. and Yi, J. (2003) Sentiment Analysis: Capturing Favorability Using Natural Language Processing. *Proceedings of the 2nd International Conference on Knowledge Capture* (Sanibel Island, Florida, October 2003, pp. 70-77). ACM
- Neave, H. R. (1990) *The Deming Dimension*. Knoxville: SPC Press
- Neely, A. (1998) *Measuring Business Performance: Why, What and How*. London: The Economist and Profile Books Ltd.
- Neely, A. (2007) 'The Search for Meaningful Measures'. *Management Services*. 51 (2): 14-17
- Neely, A., Adams, C. and Kennerley, M. (2002) *The Performance Prism: The Scorecard for Measuring and Managing Business Success*. London: Prentice Hall
- Neely, A., Richards, A. H., Mills, J. F., Platts, K. W. and Bourne, M. C. S. (1997) 'Designing Performance Measures: A Structured Approach.' *International Journal of Operations and Production Management*. 17 (11): 1131-1153
- Neely, A., Adams, C. and Crowe, P. (2001) 'The Performance Prism in Practice'. *Measuring Business Excellence*. 5 (2): 6-12
- Neely, A. and Powell, S. (2004) The Challenges of Performance Measurement: Andy Neely in Conversation with Sarah Powell'. *Management Decision*. 42 (7/8): 1017-1023
- Ness, R. G., and Patton, R. W. (1979) 'The Effect of Beliefs on Maximum Weight-lifting Performance'. *Cognitive Therapy and Research*. 3: 205-211
- Neuburger J, Walker K, Sherlaw-Johnson C, van der Meulen, J. and Cromwell, D. A. (2017) 'Comparison of Control Charts for Monitoring Clinical Performance Using Binary Data'. *BMJ Quality and Safety*. 26: 919-928

Neyroud, P. and Disley, E. (2007) "The Management, Supervision and Oversight of Criminal Investigation". In Newburn, T. (Ed.) *Handbook of Policing*. Cullompton: Willan Publishing

NHS (2018) *Making Data Count*. [Online]  
[https://improvement.nhs.uk/documents/2748/NHS\\_MAKING\\_DATA\\_COUNT\\_FINAL.pdf](https://improvement.nhs.uk/documents/2748/NHS_MAKING_DATA_COUNT_FINAL.pdf)  
[Accessed 15<sup>th</sup> June 2018]

Nielsen, M. (2015) *Statistical Process Control: What is SPC?* [Online]  
<http://www.statisticalprocesscontrol.info/whatisSPC.html> [Accessed 30th May 2015]

Nielsen, P. A. (2013) 'Performance Management, Managerial Authority, and Public Service Performance'. *Journal of Public Administration Research and Theory*. 24: 431-458

Nielsen, P. A. and Baekgaard, M. (2015) 'Performance Information, Blame Avoidance, and Politicians' Attitudes to Spending and Reform: Evidence from an Experiment'. *Journal of Public Administration Research and Theory*. 25 (2): 545–69

Nielsen, P. A. and Moynihan, D. P. (2017a) 'How do Politicians Attribute Bureaucratic Responsibility for Performance? Negativity Bias and Interest Group Advocacy'. *Journal of Public Administration Research and Theory*. 27(2): 269-283

Nielsen, P. A. and Moynihan, D. P. (2017b) 'Romanticizing Bureaucratic Leadership? The Politics of how Elected Officials Attribute Responsibility for Performance'. *Governance*. 30 (4): 541-559

Norris, D. and Hatcher, J. (1994) *The Impact of Interviewer Characteristics on Response in a National Survey of Violence Against Women*. Proceedings of the Survey Research Methods Section. American Statistics Association

Northcraft, G., Neale, M., & Earley, P. (1994) 'Joint Effects of Assigned Goals and Training on Negotiator Performance'. *Human Performance*. 7: 257 – 272

North Yorkshire OPCC (2013) *Annual Report: 2012 – 2013*. [Online]  
<http://www.northyorkshire-pcc.gov.uk/CHttpHandler.ashx?id=10947&p=0> [Accessed 14<sup>th</sup> July 2013]

North Yorkshire Police (2013) *Almost 4,000 Fewer Victims of Crime in North Yorkshire*. [Online] <http://www.northyorkshire.police.uk/11111> [Accessed 3rd July 2013]

Nunnally, J. C. (1967) *Psychometric Theory*. New York: McGraw-Hill

Observer (2011) *Police Forces Comparison*. [Online]  
<http://observer.guardian.co.uk/secondterm/table/0,8173,609883,00.html> [Accessed 21<sup>st</sup> July 2013]

Olsen, A. L. (2012) *Naming Bad Performance: Can Performance Disclosure Drive Improvements?* Paper Presented at the Midwest Political Science Association (MPSA). Session: The Psychology and Perception of Performance Measures. Chicago, 14th of April, 2012.

- Olsen, A. L. (2013a) 'Leftmost-Digit-Bias in an Enumerated Public Sector? An Experiment on Citizens Judgment of Performance Information'. *Judgment and Decision-making*. 8 (3): 365–371
- Olsen, A. L. (2013b) '*Compared to What? Reference Points in Performance Evaluation*'. Paper presented at the 11<sup>th</sup> Public Management Research Conference. Madison, Wisconsin, June 30-23 2013
- Olsen, A. L. (2013c) '*Framing Performance Information: An Experimental Study of the Negativity Bias*'. Paper presented at the 11th Public Management Research Conference Madison, Wisconsin, June 20-23 2013
- Olsen, A. R. (2015) 'Citizen (Dis)satisfaction: An Equivalence Framing Study'. *Public Administration Review*. 75 (3): 469–478
- Olsen, A. L. (2017) 'Human Interest or Hard Numbers? Experiments on Citizens' Selection, Exposure, and Recall of Performance Information'. *Public Administration Review*. 77 (3): 408-420
- Olsen, A. L. (2018) 'Precise Performance: Do Citizens Rely on Numerical Precision as a Cue of Confidence?' *Journal of Behavioral Public Administration*. 1 (1): 1-10
- Olsen, A. L., Hjorth, F., Harmon, N. and Barfort, S. (2019) 'Behavioral Dishonesty in the Public Sector'. *Journal of Public Administration Research and Theory*. 29 (4): 572-590
- Olsen, W. (1999) *Developing Open-Systems Interpretations of Path Analysis: Fragility Analysis using Farm Data from India Critical Realism: Implications for Practice*. Orebro University Sweden: Centre for Critical Realism
- O'Neill, O. (2002) *A Question of Trust*. Cambridge: Cambridge University Press
- ONS (Office for National Statistics) (2013) *Crime in England and Wales, Year Ending December 2012*. London: HMSO
- Onwuegbuzie, A. J. (2000) 'Statistics Anxiety and the Role of Self-Perceptions'. *The Journal of Educational Research*. 93 (5): 323-330
- Onwuegbuzie, A. J., Da Ros, D. and Ryan, J. M. (1997) 'The Components of Statistics Anxiety: A Phenomenological Study'. *Focus on Learning Problems in Mathematics*. 19 (4): 11-35
- Opdenakker, R. (2006) 'Advantages and Disadvantages of Four Interview Techniques in Qualitative Research'. *Forum: Qualitative Social Research*. 7 (4): Art 11
- Otley, D. (2003) 'Management Control and Performance Management: Whence and Whither?' *The British Accounting Review*. 35 (4): 309-326

- Oxley, J. E., Rivkin, J. W. and Ryall, M. D. (2010) 'The Strategy Research Initiative: Recognizing and Encouraging High-Quality Research in Strategy'. *Strategic Organization*. 8 (4): 377–386
- Pace, C. R. (1939) 'Factors Influencing Questionnaire Returns from Former University Students'. *Journal of Applied Psychology*. 23 (3): 388-397
- Pang, B. and Lee, L. (2004) A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (.July 2004, p. 271) Association for Computational Linguistics
- Pannucci, C. J. and Wilkins, E. G. (2010) 'Identifying and Avoiding Bias in Research'. *Plastic Reconstructive Surgery*. 126 (2): 619-625
- PASC (House of Commons Public Administration Select Committee) (2003) *Fifth Report. On Target? Government by Measurement*. (HC 62-I). London: The Stationery Office
- PASC (House of Commons Public Administration Select Committee) (2013a) *Public Administration Select Committee Takes Evidence on Crime Statistics*. [Online] <http://www.parliament.uk/business/committees/committees-a-z/commons-select/public-administration-select-committee/news/crime-statistics/> [Accessed 21<sup>st</sup> November 2013]
- PASC (House of Commons Public Administration Select Committee) (2013b) *Commons Select Committee: Crime Statistics*. [Online] <http://www.parliament.uk/business/committees/committees-a-z/commons-select/public-administration-select-committee/inquiries/parliament-2010/crime-statistics/?type=Written#pnlPublicationFilter> [Accessed 21<sup>st</sup> November 2013]
- PASC (House of Commons Public Administration Select Committee) (2014) *Caught Red-Handed: Why We Can't Count on Police Recorded Crime Statistics*. [Online] <http://www.publications.parliament.uk/pa/cm201314/cmselect/cmpublicadm/760/760.pdf> [Accessed 9th April 2014]
- Pasha, O., Kroll, A. and Ash, M. (2018) 'Assessing Police Performance Systems: The Impact of CompStat on Crime'. *Academy of Management Proceedings*. 1: 151-169
- Paton, G. (2013a) *Ofqual: Schools Playing the System to Boost GCSE Results*. [Online] <http://www.telegraph.co.uk/education/educationnews/10091658/Ofqual-schools-playing-the-system-to-boost-GCSE-grades.html> [Accessed 17th August 2013]
- Paton, G. (2013b) *Examiners Criticise 'Misguided' Schools over GCSE Ploy*. [Online] <http://www.telegraph.co.uk/education/educationnews/10240576/Examiners-criticise-misguided-schools-over-GCSE-ploy.html> [Accessed 17th August 2013]
- Patt, A. and Zeckhauser, R. (2000) 'Action Bias and Environmental Decisions'. *Journal of Risk and Uncertainty*. 21: 45–72

Patrick, R. (2009) *Performance Management, Gaming and Police Practice*. (PhD Thesis). Birmingham: University of Birmingham

Patrick, R. (2013) House of Commons Public Administration Select Committee: Written Evidence from Dr Rodger Patrick. November 2013 [Online]  
<http://data.parliament.uk/writtenevidence/writtenevidence.svc/evidencehtml/3383>  
November 2013 [Accessed 15th November 2013]

Patton, M. (2002) *Qualitative Research & Evaluation Methods* (3rd ed.). Thousand Oaks, CA: Sage

Pavlov, A. and Bourne, M. (2011) 'Explaining the Effects of Performance Measurement on Performance: An Organizational Routines Perspective'. *International Journal of Operations and Production Management*. 31 (1): 101-122

Pavlov, A., Mura, M., Franco-Santos, M. and Bourne, M. (2017) 'Modelling the Impact of Performance Management Practices on Firm Performance: Interaction with Human Resource Management Practices'. *Production Planning & Control*. 28 (5): 431-443

Pawson, R. and Tilley, N. (1997) *Realistic Evaluation*. London: Sage

Peirce, C. S. (1877-1878) 'Illustrations of the Logic of Science'. *Popular Science Monthly*. Vols. 12-13

Peirce, C. S. (1878) 'How to make Our Ideas Clear'. *Popular Science Monthly*. 12: 286-302

Peirce, C. S. (1905) "Review of Nichols' A Treatise on Cosmology". In H. S. Thayer (Ed.) (1984) *Meaning and Action: A Critical History of Pragmatism*. (pp. 493-495). Indianapolis: IN: Hackett

Peirce, C. S. (1931) *Harvard Lectures on Pragmatism: Collected Papers* v. 5. Para 188 [Online] <http://www.textlog.de/7664-2.html> [Accessed 30th January 2014]

Perney, J., and Ravid, R. (1991) *The Relationship Between Attitudes Towards Statistics, Math Self-Concept, Test Anxiety and Graduate Students' Achievement in an Introductory Statistics Course*. Unpublished manuscript, National College of Education: Evanston, IL

Peters, E., Dieckmann, N., Dixon, A., Hibbard, J. H. and Mertz, C. K. (2007) 'Less is More in Presenting Quality Information to Consumers'. *Med. Care Res. Review*. 64: 169-190

Pettigrew, A. M. (1995) 'The Double Hurdles for Management Research'. Distinguished Scholar Address to the Organization and Management Theory Division of the US Academy of Management. Vancouver, Canada, August 1995

Pettigrew, A. M. (1997) 'The Double Hurdles for Management Research'. In T. Clarke (ed.), *Advancement in Organizational Behaviour: Essays in Honour of Derek S. Pugh*. (pp. 277-296). London: Dartmouth Press

- Pettigrew, T. F. (1967) 'Social Evaluation Theory: Convergences and Applications'. *Nebraska Symposium on Motivation*. Vol 15: 241-311
- Phillips-Carson, P., Carson, K. D., and Heady, R. B. (1994) 'Cecil Alec Mace: The Man Who Discovered Goal-Setting'. *International Journal of Public Administration*. 17 (9): 1679 – 1708
- Pidd, M. (2005) 'Perversity in Public Service Performance Measurement'. *International Journal of Productivity and Performance Management*. 54 (5/6): 482-493
- Pollanen, R., Abdel-Maksoud, A., Elbanna, S. and Mahama, H. (2017) 'Relationships Between Strategic Performance Measures, Strategic Decision-making, and Organizational Performance: Empirical Evidence from Canadian Public Organizations'. *Public Management Review*. 19 (5):725-746
- Pollitt, C. (1993) *Managerialism and the Public Services*. Oxford: Blackwell
- Pollitt, C. (1999) *Integrating Financial Management and Performance Management*. Paris: OECD/PUMA
- Pollitt, C. (2006) 'Performance Management in Practice: A Comparative Study of Executive Agencies'. *Journal of Public Administration Research and Theory*. 16 (1): 25–44
- Pollitt, C. and Bouckaert, G. (2000) *Public Management Reform: A Comparative Analysis*, 2<sup>nd</sup> ed. Oxford University Press: Oxford
- Pollitt, C. (2002) "The New Public Management in International Perspective: An Analysis of Impacts and Effects". In: McLaughlin, K., Osborne, P. and Ferlie, E. (eds.) *New Public Management: Future Trends and Current Prospects*. London: Routledge. pp. 274 - 292
- Pope, R. and Burnes, B. (2013) 'A Model of Organisational Dysfunction in the NHS'. *Journal of Health, Organisation and Management*. 27 (6): 676-697
- Porpora, D. (1998) 'Do Realists Run Regressions?' *Paper presented at the Second Annual Conference on Critical Realism*. University of Essex, Wivenhoe Park, Colchester, Essex. September 1998
- Pratschke, J. (2003) 'Realistic Models? Critical Realism and Statistical Models in the Social Sciences'. *Philosophica*. 71: 13-38
- Pratt, M. G. (2009) 'For the Lack of a Boilerplate: Tips on Writing Up (And Reviewing) Qualitative research'. *Academy of Management Journal*. 52: 856-862
- Pratt, M., Rockmann, K. and Kaufmann, J. (2006) 'Constructing Professional Identity: The Role of Work and Identity Learning Cycles in the Customization of Identity among Medical Residents.' *Academy of Management Journal*. 49 (2): 235-262

Prenzler, T. (1997) *'Is there a Police Culture?'* Australian Journal of Public Administration. 56 (4): 47-56

Proctor, G. and Winter, C. (1998) "Information Flocking: Data Visualisation in Virtual Worlds Using Emergent Behaviours". *In*: Heudin, J. C. (eds) *Virtual Worlds. Lecture Notes in Computer Science, vol 1434*. Springer: Berlin, Heidelberg

Propper, C., Sutton, M. Whitnall, C., and Windmeijer, F. (2010) 'Incentives and Targets in Hospital Care: Evidence from a Natural Experiment'. *Journal of Public Economics*. 94 (3-4): 318 – 335

Pursula, M. (1998) "Simulation of Traffic Systems: An Overview". *In*: Bargiela, A. and Kerckhofs, E. (Eds.) *Proceedings of the 10th European Simulation Symposium*, pp. 20-24

Qualtrics (2014) *Qualtrics (software version 56837)*. Provo, UT, USA

Quantz, R. A. (1992) "On Critical Ethnography (With Some Postmodern Considerations)". *In* LeCompte, M. D., Millroy, W. L. and Preissle, J. (Eds). *The Handbook of Qualitative Research in Education*. New York: Academic Press

Quattrone, G. A. and Tversky, A. (1988) 'Contrasting Rational and Psychological Analyses of Political Choice'. *The American Political Science Review*. 82 (3): 719-736

Radin B. (2000) 'The Government Performance and Results Act and the Tradition of Federal Management Reform: Square Pegs in Round Holes?' *Journal of Public Administration Research and Theory*. 10 (1): 11 – 35

Radin, B. A. (2006) *Challenging the Performance Movement: Accountability, Complexity, and Democratic Values*. Washington, DC: Georgetown University Press

Rainey, H. (1993) "Toward a Theory of Goal Ambiguity in Public Organizations". *In* Perry, J. (ed.) *Research in Public Administration*. Vol. 2: 121–66. Greenwich, CT: JAI Press

Rachhaus, R. W. (2009) 'Principal-Agent Problems in Humanitarian Intervention: Moral Hazards, Adverse Selection, and the Commitment Dilemma'. *International Studies Quarterly*. 53 (1): 871–884

Radnor, Z. and McGuire, M. (2004) 'Performance Management in Public Sector: Fact of Fiction?' *International Journal of Productivity and Performance Management*. 53 (3): 245-260

Rafferty, A. E., Restubog, S. L. D. and Jimmieson, N. L. (2010). 'Losing Sleep: Examining the Cascading Effects of Supervisors' Experience of Injustice on Subordinate's Psychological Health'. *Work & Stress*. 24: 36-55

Rainey, H. G., Backoff, R. W. and Levine, C. L. (1976) Comparing Public and Private Organizations. *Public Administration Review*. 36 (2): 233–46

- Ramshaw, P. (2012) 'On the Beat: Variations in the Patrolling Styles of the Police Officer'. *Journal of Organizational Ethnography*. 1 (2): 213 – 233
- Rattray, J. and Jones, M. C. (2007) 'Essential Elements in Questionnaire Design and Development'. *Journal of Clinical Nursing*. 16: 234–243
- Reason, J. (2000) 'Human Error: Models and Management'. *British Medical Journal*. 320, 768-770
- Reason TV (2013) Cop Fired for Speaking Out Against Ticket and Arrest Quotas. [Online] <http://reason.com/blog/2013/07/26/online-community-comes-to-whistle-blower> [Accessed 4th August 2013]
- Reiner, R. (2010) *The Politics of the Police* (4<sup>th</sup> Ed.) Oxford: Oxford University Press
- Restubog, S. L. D., Scott, K. L., and Zagencyk, T. J. (2011). 'When Distress Hits Home: The Role of Contextual Factors and Psychological Distress in Predicting Employees' Responses to Abusive Supervision'. *Journal of Applied Psychology*. 96: 713-729
- Rice, P., and Ezzy, D. (1999) *Qualitative Research Methods: A Health Focus*. Melbourne: Oxford University Press
- Richardson, S. (2013) 'Making Routine Performance Information Relevant.' *Public Administration Review*. 73 (2): 276-277
- Ridgway, V. F. (1956) 'Dysfunctional Consequences of Performance Measurements'. *Administrative Science Quarterly* 1 (2): 240-247
- Ritov, I. and Baron, J. (1990) 'Reluctance to Vaccinate: Omission Bias and Ambiguity'. *Journal of Behavioral Decision Making*. 3: 263–277
- Rix, B. (2013) *Approaches to Performance Management in PCC's Draft Police and Crime Plans*. [Online] <http://bernardrix.com/2013/03/11/approaches-to-performance-management-in-pccs-draft-police-and-crime-plans/> [Accessed 8<sup>th</sup> June 2013]
- Roberts, G. C., and Hall, H. K. (1987) 'Motivation in Sport: Goal Setting and Performance'. Department of Kinesiology, University of Illinois. Unpublished manuscript (cited in Locke and Latham, 1990)
- Robertson, J. and Barling, J. (2014) 'Lead Well, Be Well: Leadership Behaviors Influence Employee Wellbeing'. *Wellbeing: A Complete Reference Guide*. pp.1-17
- Ron, A. (1999) *Regression Analysis and the Philosophy of Social Science: A Critical Realist View*. *Critical Realism: Implications for Practice*. Orebro University Sweden: Centre for Critical Realism
- Rosenthal, R. (1991) *Meta-Analytic Procedures for Social Research*. (2<sup>nd</sup> Edn). Newbury Park, CA: Sage

- Rothstein, R. (2008) 'Holding Accountability to Account'. *National Center on Performance Incentives*. Working Paper 2008 – 04. Nashville: Vanderbilt
- Rousseau, D. M. (1997) 'Organizational Behavior in the New Organizational Era'. *Annual Review of Psychology*. 48: 515–546
- Rousseau, D. M., Manning, J. and Denyer, D. (2008) 'Evidence in Management and Organizational Science: Assembling the Field's Full Weight of Scientific Knowledge Through Syntheses'. *Academy of Management Annals*. 2: 475–515
- Royal Statistical Society. (2005) *Performance Indicators: Good, Bad and Ugly*. London: RSS
- Rowson, J., Lindley, E. and Stanko, B. (2012) *Reflexive Coppers: Adaptive Challenges to Policing*. London: RSA
- Rozin, P., and Royzman, E. B. (2001) Negativity Bias, Negativity Dominance, and Contagion'. *Personality and Social Psychology Review*. 5 (4): 296-320
- Rutland CSP (Community Safety Partnership) (2012) *Crime Forecast*. [Online] [http://www.rutland.gov.uk/pdf/Late%20Paper%202\\_Crime%20Forecast\\_Agenda%2006.pdf](http://www.rutland.gov.uk/pdf/Late%20Paper%202_Crime%20Forecast_Agenda%2006.pdf) [Accessed 21<sup>st</sup> July 2013]
- Sanderson, I. (2001) 'Performance Management, Evaluation and Learning in 'Modern' Local Government'. *Public Administration*. 79: 297 - 313
- Sandman, P. M. (1993) *Responding to Community Outrage: Strategies for Effective Risk Communication*. Fairfax, VA: American Industrial Hygiene Association
- Sauro, J. (2007) Should You Use 5 or 7 Point Scales? [Online] <https://www.measuringusability.com/blog/scale-points.php> [Accessed 10th March 2014]
- Savage, S. (2007) *Police Reform: Forces for Change*. Oxford: Oxford University Press
- Sayer, A. (1992) *Method in Social Science: A Realist Approach*. (2nd ed). London: Routledge
- Sayer, A. (2000) *Realism and Social Science*. London: Sage
- Schalock, R. L. (2001) *Outcome Based Evaluation*. New York: Kluwer Academic / Plenum Publishers
- Scheaffer, R. L., Watkins, A. E. and Landwehr, J. M. (1998) What Every High-School Graduate Should Know about Statistics. In *Rejections on Statistics: Learning, Teaching and Assessment in Grades K-12*, Ed. S. P. Lajoie, pp. 3-31. Mahwah, NJ: Lawrence Erlbaum
- Scheurich, J. (1997) *Research method in the postmodern*. London: The Falmer Press
- Schick, A. (1998) *A Contemporary Approach to Public Expenditure Management*. Washington: World Bank Institute

Schmid, C. F. (1983) *Statistical Graphs: Design, Principles and Practices*. New York: Wiley

Schmidtke, K. A., Watson, D. G. and Vlaev, I. (2017a) 'The Use of Control Charts by Laypeople and Hospital Decision Makers for Guiding Decision Making'. *The Quarterly Journal of Experimental Psychology*. 70 (7): 1114-1128

Schmidtke, K. A., Poots, A. J., Carpio, J., Vlaev, I., Kandala, N. and Lilford, R. J. (2017b) 'Considering Chance in Quality and Safety Performance Measures: An Analysis of Performance Reports by Boards in English NHS Trusts'. *BMJ Quality & Safety*. (26): 61-69

Schroeder, W., Martin, K. and Lorensen, B. (2006) *Visualization Toolkit: An Object-Oriented Approach to 3D Graphics*. (4th Edn). Kitware: New York

Schweitzer, M. E., Ordonez, L., and Douma, B. (2004) 'Goal Setting as a Motivator of Unethical Behaviour'. *Academy of Management Journal*. 47 (3): 422 – 432

Scottish Centre for Crime and Justice Research (2014) *Non-Statutory Stop and Search in Scotland. Briefing 6/2014* [Online] <http://www.sccjr.ac.uk/wp-content/uploads/2014/06/SCCJR-Non-statutory-stop-and-search-FINAL-1.4.pdf> [Accessed 22nd March 2015]

Scottish Police Authority (2014) *SPA Policing Performance Report*. [Online] <http://www.spa.police.uk/assets/126884/241558/item8> [Accessed 27th August 2014]

Seddon, J. (2003) *Freedom from Command and Control*. Buckingham: Vanguard

Seddon, J. (2008) *Systems Thinking in the Public Sector*. Axminster: Triarchy Press

See, K. E., Heath, C. Fox, C. R. (2003) *Motivating Individual Performance with Challenging Goals: Is it Better to Stretch a Little or a Lot?* [Online] <http://fox-lab.org/wp-content/uploads/2013/08/SeeHeathFoxWP06.pdf> [Accessed 14th March 2015]

Seidman, I. (2006) *Interviewing as Qualitative Research: A Guide for Researchers in Education and Social Sciences*. (3<sup>rd</sup> Edition) New York and London: Teachers College Press

Seijts, G. H., and Latham, G. P. (2001) 'The Effect of Learning, Outcome, and Proximal Goals on a Moderately Complex Task'. *Journal of Organizational Behaviour*. 22: 291-307

Shane, J. M. (2007) *What Every Chief Executive Should Know: Using Data to Measure Police Performance*. New York: Looseleaf Law Publications, Inc

Shane, J. M. (2008) *Developing a Performance Management Model*. New York: Looseleaf Law Publications, Inc

- Shane, J. M. (2010) 'Performance Management in Police Agencies: A Conceptual Framework'. *Policing: An International Journal of Police Strategies & Management*. 33 (1): 6-29
- Shapira, Z. (1989) 'Task Choice and Assigned Goals as Determinants of Task Motivation and Performance'. *Organizational Behavior and Human Decision Processes*. 44: 141 – 165
- Sharma, S. (2017) 'Definitions and Models of Statistical Literacy: A Literature Review'. *Open Review of Educational Research*. 4 (1): 118-133
- Shaw, D. (2013) *Crime Statistics are Manipulated, Says Police Chief*. [Online] <http://www.bbc.co.uk/news/uk-25022680> [Accessed 21st November 2013]
- Shaw, D. (2020) *'No Excuses for Not Cutting Crime' Patel Tells Police*. BBC. <https://www.bbc.co.uk/news/uk-51645526> [Accessed 14th March 2020]
- Sherlaw-Johnson C, and Bardsley M (2016) *Monitoring Change in Health Care Through Statistical Process Control Methods*. [Online] The Nuffield Trust 2016 <http://www.nuffieldtrust.org.uk/publications/monitoring-change-health-care-through-statistical-process-control-methods> [Accessed 15th June 2018]
- Shewhart, W. A. (1939) *Statistical Method from the Viewpoint of Quality Control*. Washington DC: The Graduate School, US Department of Agriculture
- Shields, P. M. (1998) "Pragmatism as Philosophy of Science: A Tool for Public Administration". In White, J. D. (Ed.) *Research in Public Administration*. (pp.199-230) London: JAI Press
- Shilston, T. G. (2008) 'One, Two, Three, What Are We Still Counting For? Police Performance Regimes, Public Perceptions of Service Delivery and the Failure of Quantitative Management'. *Policing: A Journal of Policy and Practice*. 2 (3): 359 - 366
- Shiple, N. and Chakraborty, J. (2018) 'Using Pinterest to Improve the Big Data User Experience - A Comparative Analysis in Healthcare'. *Trends and Advances in Information Systems and Technologies*. 746: 949-960
- Shorrocks, S. and Licu, T. (2013) 'Target Culture: Lessons in Unintended Consequences'. *Hindsight*. No 17, Summer 2013: 10 - 16
- Silverman, D. (2000) *Doing Qualitative Research*. London: Sage
- Simon, H. A. (1955) 'A Behavioral Model of Rational Choice.' *The Quarterly Journal of Economics*. 69 (1): 99-118
- Simon, H. A. (1978) 'Rationality as a Process and as Product of Thought'. *American Economic Review: Papers and Proceedings*. 68: 1-16
- Simons, R. (1995) *Levers of Control: How Managers Use Innovative Control Systems to Drive Strategic Renewal*. Boston: Harvard Business School Press

- Sitkin, S. B., See, K. E., Miller, C. C., Lawless, M. W. and Carton, A. M. (2011) 'The Paradox of Stretch Goals: Organizations in Pursuit of the Seemingly Impossible'. *Academy of Management Review*. 36 (3): 544-566
- Skakon, J., Nielsen, K., Borg, V. and Guzman, J. (2010) Are Leaders' Well-being, Behaviors, and Style Associated with the Affective Well-being of their Employees? A Systematic Review of Three Decades of Research. *Work & Stress*. 24: 107-139
- Skolnick, J. H. (1966) *Justice without Trial: Law Enforcement in Democratic Society*. New York: Wiley
- Smerecnik, C. M. R., Mesters, I, Kessels, L. T. E., Ruiter, R. A. C., de Vries, N. K. and de Vries, H. (2010) 'Understanding the Positive Effects of Graphical Risk Information on Comprehension: Measuring Attention Directed to Written, Tabular, and Graphical Risk Information'. *Risk Analysis*. 30: 1387–1398
- Smith, P. (1990) 'The Use of Performance Indicators in the Public Sector'. *Journal of the Royal Statistical Society*. 153 (1) pp.53-72
- Smith, P. (1993) 'Outcome-related Performance Indicators and Organizational Control in the Public Sector'. *British Journal of Management* 4 (3): 135-151
- Smith, P. (1995) 'On the Unintended Consequences of Publishing Performance Data in the Public Sector'. *International Journal of Public Administration*. 18 (2 & 3): 277–310
- Smith, A. (2006) *Crime Statistics: An Independent Review*. [Online] <http://webarchive.nationalarchives.gov.uk/20110218135832/http://rds.homeoffice.gov.uk/rds/pdfs06/crime-statistics-independent-review-06.pdf> [Accessed 6th April 2015]
- Smith, G. R., Fischer, E. P., Nordquist, C. R., Mosley, C. L., and Ledbetter, N. S. (1997) 'Implementing Outcomes Management Systems in Mental Health Settings'. *Psychiatric Services*. 48 (3): 364 - 368
- Solomon, D. J. (2000) *Conducting Web-Based Surveys*. [Online] <http://files.eric.ed.gov/fulltext/ED458291.pdf> Educational Resources Information Center. [Accessed 23rd January 2014]
- Sommers, J. (2013) *Furore Erupts Over Crime Figures 'Fiddling'*. [Online] [http://www.policeoracle.com/news/Police+Performance/2013/Nov/20/Furore-erupts-over-crime-figures-fiddling\\_74508.html](http://www.policeoracle.com/news/Police+Performance/2013/Nov/20/Furore-erupts-over-crime-figures-fiddling_74508.html) [Accessed 21st November 2013]
- Spectator (2011) *Target Men*. [Online] <http://www.spectator.co.uk/essays/all/6975453/target-men.shtml> [Accessed 31st July 2013]
- Speier, C. (2006) 'The Influence of Information Presentation Formats on Complex Task Decision-Making Performance'. *International Journal of Human-Computer Studies*. 64: 1115–1131
- Spitzer, D. R. (2007) *Transforming Performance Measurement: Rethinking the Way We Measure and Drive Organizational Success*. New York: Amacom

Staffordshire OPCC (2013) *Information Form OPCC/I/2013/001: Force Performance 1 April to 31 December 2012*. [Online] <http://www.staffordshire-pcc.gov.uk/Document-Library/Agenda-Papers-300113.pdf> [Accessed 14th July 2013]

Staffordshire Police (2019) *Missing People – Performance Data*. Unpublished

Steinbart, P. J. (1989) 'The Auditor's Responsibility for the Accuracy of Graphs in Annual Reports: Some Evidence of the Need for Additional Guidance'. *Accounting Horizons*. 3 (3): 60–70

Stevenson, M. K., Kanfer, F. H., and Higgins, J. M. (1984) 'Effects of Goal Specificity and Time Cues on Pain Tolerance'. *Cognitive Therapy and Research*. 8: 415-426

Strathclyde Police (2010) *Force Performance Report. Appendix 'A'*. Unpublished

Strathclyde Police (2012) *Strathclyde Police Authority: Force Performance Report. Appendix 'A'*. Unpublished

Strauss, A., and J. Corbin (1990) *Basics of Qualitative Research*. Newbury Park, CA: Sage

Suddaby, R., Hardy, C. and Huy, Q. (2011) 'Where are the New Theories of Organizations?' *Academy of Management Review*. 36: 236-246

Suff, P., Reilly, P. and Cox, A. (2008) *Paying for Performance: New Trends in Performance-Related Pay*. Brighton: IES

Sun, P. Y. and Anderson, M. H. (2012) 'The Combined Influence of Top and Middle Management Leadership Styles on Absorptive Capacity'. *Management Learning*. 43 (1): 25-51

Sutton, R. I. and Staw, B. M. (1995) 'What Theory is Not'. *Administrative Science Quarterly*. 40 (3): 371-384

Surrey OPCC (Office of the Police and Crime Commissioner) (2013) *Police and Crime Plan*. [Online] <http://mycouncil.surreycc.gov.uk/mgConvert2PDF.aspx?ID=4513> [Accessed 8<sup>th</sup> June 2013]

Talisse, R. B. and Aikin, S. F. (2011) *The Pragmatism Reader: From Peirce Through the Present*. Princeton, NJ: Princeton University Press

Tapanes, M. A. (2008) *Chi-Square Goodness of Fit Test*. [Online] <http://www.coedu.usf.edu/IT/Flash/FlashShowcase2008/Tapanes/goodnessoffit.html> [Accessed 4th August 2014]

Taylor, F. (1911) *The Principles of Scientific Management*. New York: Harper & Row

Taylor, J. (2009) 'Strengthening the Link between Performance Measurement and Decision-making'. *Public Administration*. 87 (4): 853–871

- Taylor-Gooby, P. (2009) *Reframing Social Citizenship*. Oxford: Oxford University Press
- Tendler, S. (2007) *We Are Making Ludicrous Arrests Just to Meet our Targets* [Online] <http://www.timesonline.co.uk/tol/news/uk/crime/article1790515.ece> [Accessed 18th June 2013]
- Tepper, B. J. (2000) 'Consequences of Abusive Supervision'. *The Academy of Management Journal*. 43: 178-190
- Terpstra, J. and Trommel, W. (2009) 'Police, Managerialization and Presentational Strategies'. *Policing: An International Journal of Police Strategies & Management*. 32 (1): 128-143
- Tesser, A. (1988) "Toward a Self-Evaluation Maintenance Model of Social Behavior". In Berkowitz, L. (Ed.), *Advances in Experimental Social Psychology*. (Vol. 21, pp. 193-355). New York: Academic Press
- Thagard, P. and Shelley, C. (1997) 'Abductive Reasoning: Logic, Visual Thinking, and Coherence'. Philosophy Department, University of Waterloo. Waterloo: Ontario
- Thames Valley OPCC (2013) *Police and Crime Plan*. [Online] <http://www.thamesvalley-pcc.gov.uk/Document-Library/Police-and-Crime-Plan.pdf> [Accessed 8th June 2013]
- Tobler, W. R. (1970) 'A Computer Movie Simulating Urban Growth in the Detroit Region'. *Economic Geography*. 46 (2): 234-240
- Transport Scotland (2018) *Scotland's Road Safety Framework to 2020*. [Online] <https://www.transport.gov.scot/media/29622/j243698.pdf> [Accessed 11th October 2018]
- Trochim, W. M. K. (2006a) *Construct Validity*. [Online] <http://www.socialresearchmethods.net/kb/constval.php> [Accessed 1st August 2014]
- Trochim, W. M. K. (2006b) *Construct Validity*. [Online] <http://www.socialresearchmethods.net/kb/convdisc.php> [Accessed 1st August 2014]
- Trochim, W. M. K. (2006c) *External Validity*. [Online] <http://www.socialresearchmethods.net/kb/external.php> [Accessed 4th August 2014]
- Tsang, E. W. K. (2006) 'Behavioral Assumptions and Theory Development: The Case of Transaction Cost Economics'. *Strategic Management Journal*. 27 (11): 999-1011
- Tufte, E. R. (1990) *Envisioning Information*. Cheshire, Connecticut: Graphics Press
- Tufte, E. R. (1997) *Visual Explanation: Images and Quantities, Evidence and Narrative* Cheshire, Connecticut: Graphics Press
- Tufte, E. R. (2001) *The Visual Display of Quantitative Information*. (2<sup>nd</sup> Edition) Cheshire, Connecticut: Graphics Press

- Tufte, E. R. (2013) *Visual and Statistical Thinking: Displays of Evidence for Making Decisions*. (5<sup>th</sup> Edn) Cheshire, Connecticut: Graphics Press
- Tully, S. (1994) 'Why to Go for Stretch Targets'. *Fortune*. 130 (10): 145-158
- Turney, P. D. (2001) 'Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL'. *Proceedings of the Twelfth European Conference on Machine Learning*. (pp. 491-502) Berlin: Springer-Verlag
- Turney, P. D. (2002) 'Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews'. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*. Philadelphia, Pennsylvania, USA, July 8-10, 2002. pp 417-424. NRC 44946
- Tversky, A. and Kahneman, D. (1974) 'Judgment Under Uncertainty: Heuristics and Biases'. *Science*. New Series, Vol. 185, No. 4157. (Sep. 27, 1974), pp. 1124-1131
- Tversky, A., and Kahneman, D. (1986) 'Rational Choice and the Framing of Decisions'. *The Journal of Business*. 59 (4): 251-278
- Tversky, A., and Kahneman, D. (1991) 'Loss Aversion in Riskless Choice: A Reference-Dependent Model'. *The Quarterly Journal of Economics*. 106 (4): 1039-106
- Udo-Akang, D. (2012) 'Theoretical Constructs, Concepts and Applications'. *American International Journal of Contemporary Research*. 2 (9): 89 – 97
- Uhrig, J. D., Harris-Kojetin L., Bann, C. and Kuo, T. M. (2006) 'Do Content and Format Affect Older Consumers' use of Comparative Information in a Medicare Health Plan Choice? Results from a Controlled Experiment'. *Med. Care Res. Review*. 63: 701–718
- UK Statistics Authority (2010) *Overcoming Barriers to Trust in Crime Statistics: England and Wales. (Monitoring Report 5)*. London: UK Statistics Authority
- United Kingdom Commission for Health Improvement (2002) *Report on the Clinical Governance Review on Surrey and Sussex Healthcare NHS Trust*. London: HMSO
- US Department of Justice (2016) *Audit of the Federal Bureau of Investigation Annual Financial Statements: Fiscal Year 2015*. [Online] <https://oig.justice.gov/reports/2016/a1607.pdf> [Accessed 19th February, 2016]
- Vagias, W. M. (2006) *Likert-Type Scale Response Anchors*. Clemson International Institute for Tourism and Research Development, Department of Parks, Recreation and Tourism Management. Clemson University
- Van Dooren, W., Bouckaert, G., and Halligan, J. (2010) *Performance Management in the Public Sector*. London: Routledge
- Van Maanen, J. (1973) 'Observations on the making of Policemen'. *Human Organization*. 32 (4):407-418

- Van Maanen, J. (1974) 'Working the Street: A Developmental View of Police Behavior'. In H. Jacob (ed.) *The Potential for Reform of Criminal Justice (Sage Criminal Justice System Annual Review, Volume 3)* Beverly Hills, California: Sage Publications. pp. 83-130
- Van Maanen, J. (1975) 'Police Socialization'. *Administrative Science Quarterly*.20 (3): 207-228
- Van Maanen, J. (1979) 'The Fact of Fiction in Organizational Ethnography'. *Administrative Science Quarterly*. 24: 539–550
- Van Niekerk, A. and May, M. (2012) 'Exploring How Middle Managers Experience the Impact of Senior Management on their Integrity'. *South African Journal of Labour Relations*. 36 (2): 42-61
- Vancil, R. F. (1978) *Decentralization: Ambiguity by design*. Homewood, IL: Dow Jones-Irwin
- Vernon, M. D. (1950) 'The Visual Presentation of Factual Data'. *British Journal of Educational Psychology*. 20: 174-185
- Vessey, I. (1991) 'Cognitive Fit: A Theory-Based Analysis of the Graphs versus Tables Literature'. *Decision Science*. 22: 219–240
- Waddell, A. and Pio, E. (2015) 'The Influence of Senior Leaders on Organisational Learning: Insights from the Employees' Perspective'. *Management Learning*. 46 (4): 461-478
- Waddington, P.A.J. (1999) '*Police Canteen Subculture: An Appreciation*'. *British Journal of Criminology*. 32 (2): 287-309
- Wagner, K. C. (1954) 'Latent Functions of an Executive Control: A Sociological Analysis of a Social System under Stress'. *Research Previews*. Vol. 2 (Chapel Hill: Institute for Research in Social Science. March 1954) mimeo.
- Wainer, H. (1984) 'How to Display Data Badly'. *The American Statistician*. 38 (2): 137-147
- Wacker, J. (1998). 'A Definition of Theory: Research Guidelines for Different Theory-Building Research Methods in Operations Management'. *Journal of Operations Management*. 16 (4): 361–385
- Wallman. K. K. (1993) 'Enhancing Statistical Literacy: Enriching our Society'. *Journal of American Statistical Association*. 88 (421): 1-8
- Ward, M., Grinstein G, and Keim, D. (2015) *Interactive Data Visualization: Foundations, Techniques and Applications*. Boca Raton, FL: CRC Press
- Ware, C. (2013) *Information Visualization: Perception for Design*. (3<sup>rd</sup> Edn). Elsevier: Waltham, MA

- Wason, P. C. (1960). 'On the Failure to Eliminate Hypotheses in a Conceptual Task'. *Quarterly Journal of Experimental Psychology*. 12: 129-140
- Watson, J. M. (1997) 'Assessing Statistical Literacy using the Media'. In Gal, Land Garfield, J. B. (eds), *The Assessment Challenge in Statistical Education*. Amsterdam, IOS Press and The International Statistical Institute, pp 107-121
- Watson, J. (1998) 'The Role of Statistical Literacy in Decisions about Risk: Where to Start'. *For the Learning of Mathematics*. 18 (3): 25-27
- Watson, J. M. (2003) Statistical Literacy at the School Level: What should Students Know and Do? *ISI 54 Berlin 2003*
- Watson, J. M. (2006) Issues for Statistical Literacy in the Middle School. In *ICOTS-7 Conference Proceedings. IASE, Salvador (CD-Rom)* (pp. 1-6)
- Watson, J. (2011) 'Foundations for Improving Statistical Literacy'. *Statistical Journal of the IAOS*. 27 (3-4): 197-204.
- Watson, J. M., and Callingham, R. (2003) 'Statistical Literacy: A Complex Hierarchical Construct'. *Statistics Education Research Journal*. 2 (2): 3-46
- Weisburd, D., Mastrofski, S. D., McNally, A., Greenspan, R., and Willis, J. J. (2003) 'Reforming to Preserve: Compstat and Strategic Problem Solving in American Policing'. *Criminology and Public Policy*. 2: 421 – 456
- Welch, W. W. and Barlau, A. N. (2014) *Addressing Survey Nonresponse Issues: Implications for ATE Principal Investigators, Evaluators, and Researchers*. [Online] <http://www.colorado.edu/ibs/decaproject/pubs/Survey%20nonresponse%20issues%20Implications%20for%20ATE%20PIs%20researchers%20%20evaluators.pdf> [Accessed 14<sup>th</sup> February 2014]
- Western, S. (2007) *Leadership: A Critical Text*. New York: London: Sage
- Wheeler, D. (1998) *Avoiding Man-Made Chaos*. Knoxville: SPC Press
- Wheeler, D.J. (2000) *Understanding Variation: The Key to Managing Chaos*. (2<sup>nd</sup> Ed.) Knoxville: SPC Press
- Whitaker, G., Mastrofski, S., Ostrom, E., Parks, R.B. and Percy, S.L. (1982) *Basic Issues in Police Performance*. Washington DC: National Institute of Justice
- Whitehead, T. (2010) *Police Accused of Fiddling Response Times*. The Telegraph. [Online] <http://www.telegraph.co.uk/news/uknews/law-and-order/7324855/Police-accused-of-fiddling-response-times.html> [Accessed 31st July 2013]
- Williams, C. K. and Karahanna, E. (2013) 'Causal Explanation in the Coordinating Process: A Critical Realist Case Study of Federated IT Governance Structures'. *MIS Quarterly*. 37 (3): 933-964
- Willis, J. J., Mastrofski, S. D., and Weisburd, D. (2004) 'CompStat and Bureaucracy: A Case Study of Challenges and Opportunities for Change'. *Justice Quarterly*. 21 (3): 463 – 496

Willis, J. J., Mastrofski, S. D. and Weisburd, D. (2007). 'Making sense of CompStat: A Theory-Based Analysis of Organizational Change in Three Police Departments'. *Law and Society Review*. 41 (1): 147 – 188

Wilson, J. Q. (1968) *Varieties of Police Behavior: The Management of Law and Order in Eight Communities*. Cambridge, MA: Harvard University Press

Wilson, D. and Dixon, W. (2006) 'Das Adam Smith Problem. A Critical Realist Perspective'. *Journal of Critical Realism*. 5: 252–272

Wilson, D. and Piebalga, A. (2008) *Accurate Performance Measure but Meaningless Ranking Exercise? An Analysis of the English School League Tables*. CMPO Working Paper No. 07/176. Bristol: University of Bristol

Wilson, T., Wiebe, J. and Hoffman, P. (2005) 'Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis'. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*. pp. 347–354: Vancouver, October 2005

Wiltshire Police (2014a) *Performance Culture Review*. Unpublished

Wiltshire Police (2014b) *Systems Thinking and Performance Culture*. Unpublished.

Winship, C. and Mare, R. D. (1992) 'Models for Sample Selection Bias'. *Annual Review of Sociology*. 18: 327-350

Wintour, P. and Domokos, J. (2013) *Leaked JobCentre Newsletter Urges Staff to Improve on Sanctions Targets*. [Online]  
<http://www.theguardian.com/society/2013/mar/25/jobcentre-newsletter-sanctions-targets> [Accessed 18th August 2013]

Woodall, W. H. (2006) 'The Use of Control Charts in Health-Care and Public Health Surveillance'. *Journal of Quality Technology*. 38: 89–104

Woodcock, T., Liberati, E. G., and Dixon-Woods, M. (2019) 'A Mixed-Methods Study of Challenges Experienced by Clinical Teams in Measuring Improvement'. *BMJ Qual Saf* 2019 (0): 1-10. doi: 10.1136/bmjqs-2018-009048

Wright, S. (2007) *Nicked for Throwing a Cream Bun: Police Dossier of Dubious Offences*. Daily Mail. [Online] <http://www.dailymail.co.uk/news/article-454889/Nicked-throwing-cream-bun-police-dossier-dubious-offences.html> [Accessed 29<sup>th</sup> July 2013]

Wu, G., Heath, C. and Larrick, R. (2008) *A Prospect Theory Model of Goal Behavior*. [Online]  
<http://faculty.chicagobooth.edu/george.wu/research/papers/wu%20heath%20larrick%20%28prospect%20theory%20model%20of%20goal%20behavior%29.pdf> [Accessed 8<sup>th</sup> March 2015]

Wuensch, K. L. (2014) *Binary Logistic Regression with SPSS*. [Online] <http://core.ecu.edu/psyc/wuenschk/mv/multreg/logistic-spss.pdf> [Accessed 15th July 2014]

Yau, N. (2013) *Data Points: Visualization that Means Something*. Indianapolis: Wiley

Young, M. (1991) *An Inside Job: Policing and Police Culture*. Oxford: Clarendon Press

Zeelenberg, M., Van Den Bos, K., Van Dijk, E. and Pieters, R. (2002) 'The Inaction Effect in the Psychology of Regret'. *Journal of Personality and Social Psychology*. 82 (3): 314–327

Zeidner, M. (1991) Statistics and Mathematics Anxiety in Social science Students: Some Interesting Parallels. *British Journal of Educational Psychology*. 61 (3): 319-328

Zikmund, W., Babin, B., Carr, J., and Griffin, M. (2008) *Business Research Methods* (8th ed.) Mason, OH: Cengage