Binary Comparisons and Police Performance Measurement: Good or Bad?

Simon Guilfoyle*

Abstract  A common method of presenting data in the police performance environment is the use of binary comparisons. This involves comparing two isolated numeric values, then interpreting the difference between them as a trajectory, or assuming it is significant. The practice is commonplace within UK police forces, appearing at face value to be a simple method for interpreting data. However, concerns exist about the efficacy of the approach; furthermore, experience suggests the practice leads to unwarranted assumptions, which impair decision-making and encourage inappropriate behavioural responses. This article outlines key findings from the author’s doctoral research, which utilizes experimental psychometric testing to assess the predictability of adverse behavioural outcomes associated with the practice.

Introduction

Consider this statement:

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... The overall two-year trend shows a decrease in crime on most forms of transport. 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 above paragraph is typical of the way binary comparisons appear to convey apparent trajectories. (Note the choice of language: ‘fell’, ‘down’, ‘decrease’, ‘rose’). Ostensibly, comparing a current data point with a previous one (e.g. last month/last year/last quarter/same period last year) appears to show whether crime is increasing or decreasing. Indeed, binary comparisons seem to be a straightforward way of quickly establishing ‘direction of travel’, or as a ‘starting point for asking questions’.

However, it is argued that the practice is oversimplistic and misleading. It disregards variation (Deming, 1986; Wheeler, 2000) and is incapable of communicating accurate information about trends, trajectories, or outliers, while appearing to do precisely that. Specifically, it encourages over-interpretation of minor differences, potentially leading to unwarranted conclusions:

Very particularly the practice of concentrating on a comparison with the most recent value, this year’s results...

*Warwick Business School, University of Warwick, UK. E-mail: s.j.guilfoyle@warwick.ac.uk
compared with last year’s, may be very misleading for reasons including regression to the mean. (Bird et al., 2005, p. 14)

It is the author’s experience that the use of binary comparisons in the police performance environment consistently acts as a catalyst for unwarranted assumptions and disproportionate operational responses; managers ascribe meaning to differences, then initiate unnecessary activity intended to rectify perceived deficiencies. This has an adverse impact on the overall system, as capacity is utilized to address ‘trends’ which may not actually exist.

Therefore, as part of ongoing PhD research the author (a serving police officer) studied how police officers interpret and react to different types of performance measurement stimuli, and how predictable these responses are. This article provides a summary of the findings of tests relating to a binary comparison stimulus.

Theoretical grounding: Agency Theory

The research draws upon Agency Theory as its foundational theoretical framework, testing the predictability of behavioural phenomena associated with the use of binary comparisons as performance information, at the level of the principal. Foss and Stea (2014) argue, ‘Agency theory is one of the most important foundational theories in management research... ’ (2014, p. 101). Built around the Principal–Agent construct, it is a complex network of assumptions and theoretical mechanisms.

Kassim and Menon (2003) offer the following definition of the Principal–Agent model:

Agency relationships are created when one party, the principal, enters into a contractual agreement with a second party, the agent, and delegates to the latter responsibility for carrying out a function or set of tasks on the principal’s behalf (2003, p. 122).

Agency Theory is underpinned by assumptions about rational choice, risk disposition, and motivation (Jensen and Meckling, 1976; Jensen, 1983; Eisenhardt, 1985, 1989; Davis et al., 1997a,b; Caers et al., 2006; Van Slyke, 2007). Albanese et al. (1997) assert, ‘Fundamentally, agency theory assumes principals and agents are rational actors’ (1997, p. 610). Rational Choice Theory proposes individuals tend to seek the most efficient outcome for themselves, even at the expense of the collective good (Smith, 1776; Coleman, 1990; Abell, 1991; Zey, 1998). Agency Theory assumptions about risk disposition predict that agents tend to be risk averse, whereas principals tend to be risk neutral (Eisenhardt, 1989; Caers et al., 2006, Linder and Foss, 2013).

The model is dominant in hierarchical organizational structures (such as the police service), where it is assumed that its inherently transactional modes of operation promote efficiency and accountability (Hall and Taylor, 1996; Pollack, 1997; Egan, 1998). Multiple Principal–Agent relationships exist throughout the hierarchy, with some actors performing dual roles, which can result in complex, multilevel relationships. However, the construct is vulnerable to destabilizing influences, related to the potentially divergent interests of principals and agents. In particular:

- Information asymmetry: when one party possesses information the other party does not possess.
- Adverse selection: when one party knows more about attributes of a product or service than another and, as a result, the uninformed party runs the risk of purchasing a product or service of low quality.
- Moral hazard: when one party uses information and expertise and acts opportunistically, to the exclusion of the agreed contract (Adapted from Van Slyke, 2007, p. 162).

To mitigate these risks, principals adopt a combination of controls, incentives, and sanctions...
linked to agents’ performance (Jensen and Meckling, 1976; Walsh and Seward, 1990; Van Slyke, 2007). This is deemed necessary, as ‘... when the agent’s behaviour is not controlled or restrained, the goals of the principal are unlikely to be attained’ (Caers et al., 2006, p. 26). In policing, such controls do not always incorporate the types of incentives commonplace in the private sector (e.g. payment by results); nevertheless, there exists implicit pressure exerted by principals in the pursuit of ‘good’ performance.

For example, managers may develop perceptions about performance based on binary comparisons and use this as a basis for criticism, praise, or to demand improvement. The practice can affect how principals perceive ‘good’ or ‘bad’ performance and how they react as a consequence; this can then influence agents’ behaviour and have an impact at the organizational level.

Agency Theory is, therefore, considered to be a useful and well-established theoretical lens through which to explore the issues at hand. Specifically, the mode of information policy deployed within the agency construct is at the centre of this inquiry (Jacobides and Cronson, 2001; Foss and Stea, 2014). Information Policy (IP) is defined as ‘... the collection and use of performance information in situations of delegated authority to support executive decisions and to determine rewards and incentives’ (Jacobides and Cronson, 2001, p. 202).

Different modes of IP support the creation and distribution of value, and in particular, joint agency value; therefore, the use of effective measures can reduce information asymmetry, goal incongruence, and lead to greater efficiency. Conversely, where IP comprises data presented in binary comparison format, there exists the danger that it can easily be misinterpreted, leading to an adverse impact on the system.

Although discussions about IP in Agency Theory tend to focus on direct measures of agents’ performance (Foss and Stea, 2014), it is argued that team and organizational-level data can also be useful for understanding performance. This can even apply where the data are not directly linked to ‘performance’ (e.g. crime rates), but may be useful as a source of information for decision makers (i.e. principals). Therefore, in addition to those direct measures pertaining to agents’ activity, it is suggested that the concept of information policy can be extended to encompass wider data sources in agency models; as demonstrated by Holmström (1979, 1999), the key consideration when determining measures is the degree of their informativeness.

Consequently, Linder and Foss (2013) emphasize the importance of using the right information to assess performance; if data are not informative (due to their type or mode of presentation), this adds ‘noise’ and makes it difficult to draw conclusions about performance. Wheeler (2000) too, reminds us: ‘Data have no meaning apart from their context’ (2000, p. 12). It is argued here that such absence of context and depth are the chief deficiencies of binary comparisons. Conversely, drawing on Fama and Jensen (1983), Eisenhardt (1989) observes that when richer information is available, decision makers are more likely to engage in behaviours consistent with organizational interests.

Furthermore, she asserts that organizations can compensate for high-outcome uncertainty, difficulties in measurability, and the management of complex activity by improving information systems (Eisenhardt, 1985). Therefore, it is suggested that relevance and clarity of performance information, whether directly related to agents’ activities or otherwise, is a critical factor in mitigating disproportionate/unwarranted behavioural responses (DUBR) by principals.

Consequently, the research aims to establish whether adverse behavioural phenomena arising from the use of binary comparisons as a mode of IP in the police performance management context are predictable. The following section outlines the approach adopted to ascertain this.
Research design and implementation

The study was conducted utilizing an electronic survey instrument constructed using Qualtrics software (Qualtrics, 2014), involving police officers of all ranks and from every UK police force. It comprised five blocks of thematic psychometric micro experiments, with the binary comparison tests being presented to respondents randomly among other stimuli. The structure of the binary comparison block was as follows:

1. Visual stimulus: respondents were asked to provide their interpretation of a numeric table displaying crime figures.
2. Emotional response: respondents were asked a question that assessed their level of concern, arising as a consequence of their interpretation.

The survey instrument was piloted and validated using a content validity approach (Litwin, 1995, p. 35). It was established that variables produced nominal- and ordinal-level data, which exhibited patterns of non-normal distribution. Therefore, non-parametric analytical techniques were selected (Kuzon Jr et al., 1996; Lehmann, 1998; Jamieson, 2004; Hoskin, 2014) and an analysis was conducted using IBM SPSS (IBM Corp, 2013).

The instrument’s Cronbach’s coefficient was found to be 0.921, indicating high-internal reliability, well in excess of the conventional threshold of 0.7 (Nunnaly, 1967; Kline, 1999; Brace et al., 2009; Field, 2013). Furthermore, the vast majority of inter-item correlations were >0.3, and a further test of correlation between individual variables (Spearman’s rho) confirmed significant correlations at the p < 0.001 level in all categories (Fink, 1995, pp. 38–39; Field, 2013, p. 685).

Factor analysis ascertained the instrument possessed factorability (Kline, 1994). The Kaiser–Meyer–Olkin Measure of Sampling Adequacy (KMO) = 0.906, indicating compact patterns of correlation (Hutcheson and Sofroniou, 1999). Additionally, Bartlett’s Test of Sphericity produced output of $\chi^2 = 56,190.223$, df = 300, p < 0.001 (Brace et al, 2009, p. 354). These highly significant results confirm that a large proportion of variance is explained by factors. Furthermore, anti-image matrices comparing negative partial covariances and negative partial correlations indicated the presence of a factor structure underlying the variables. Principal Component Analysis (PCA) also confirmed the instrument may be considered highly reliable, and therefore capable of assessing the underlying dimensions necessary to draw meaningful conclusions from the psychometric tests.

To address potential self-selection bias (Knapp, 2014), a reference survey (N = 195) was conducted with officers from a single force, where a fully randomized sample of respondents was selected using systematic sampling (De Vaus, 2002, pp. 72–74). Data were tested for late response bias and non-response bias (Pace, 1939; Miller and Smith, 1983; Babbie, 1990; Ary et al., 1996). Mann–Whitney U tests were then conducted to establish if output from the reference survey differed significantly from that of the national survey. No statistically significant differences were found between the two data sets, thereby indicating the reference sample was indeed reflective of the national sample; this may therefore be considered representative of the UK police officer population, enabling generalization of findings and ‘test–retest’ reliability (Field, 2013, p. 885).

In April 2014, the survey instrument was circulated to all police forces in England, Wales, Scotland, and Northern Ireland, with the support of...
of the Association of Chief Police Officers (ACPO),
the Police Federation of England and Wales, the
Police Superintendents Association of England
and Wales, and their Scottish and Northern Irish
counterparts. It was active for a period of 30 days,
with 4,917 respondents completing all sections of
the binary comparisons tests, and every force and
rank being represented.

This sample size leads to extremely high-
confidence levels and significant statistical power,
meaning inferential statistics may be legitimately
extrapolated to the UK police officer population,
at the following levels:

- Confidence level: 99.99%
- Indicative probability levels (two-tailed):
  \( p < 0.001 \)
- Indicative margin of error: 2.97%
- Statistical power:
  - 0.989 (to detect an effect size of 0.1)
  - 1.000 (to detect effect sizes > 0.15)

(Raosoft, 2004; Faul et al., 2007; Kalpana, 2011).

These confidence and probability levels are well
in excess of the standard criteria of 95% and
\( p \leq 0.05 \) generally used in statistical analysis
(Field, 2013), while power is much greater than the recommended level of 0.8
(Cohen, 1988). In particular, it is notable that any
detected effect size \( > 0.15 \) is confirmed at the 100%
level.

Therefore, it may be stated that the instrument is
of sufficient rigour to be capable of producing
inferential statistics with high levels of significance
and confidence.

Analysis and findings

The stimulus used for the binary comparison tests
is reproduced below, followed by associated ques-
tions, hypotheses, and analysis.

<table>
<thead>
<tr>
<th>Crime Figures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Last month</td>
</tr>
<tr>
<td>1,598</td>
</tr>
</tbody>
</table>

Figure 1: Crime figures table.

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. Although it is argued it is not pos-
sible to derive meaningful inferences from such a
binary comparison, experience suggests this is pre-
cisely what often happens. In this case, the most
obvious assumption might be that crime is increas-
ing (even though the difference may not be signifi-
cant 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 (Fig. 1).

Key avenues of inquiry are:

1. Is there a strong probability that respondents
   assume there is a trajectory, and therefore select
   ‘Crime is increasing’, rather than other available
   options?
2. Will the group of respondents who select ‘Crime
   is increasing’ be more likely to enter higher
   scores in respect of their emotional response
   (i.e. level of concern)?
3. Is this group of respondents more likely to enact
   DUBRs?

---

2 At the time of the survey, the most recent data on UK police officer numbers indicated there were 129,584 full-time
equivalent (FTE) police officers in the 43 English and Welsh police forces, 17,496 FTE officers in Scotland, and 6,885 FTE
officers in the Police Service of Northern Ireland, totalling 153,956 UK FTE officers (Dar and Berman, 2013).
The questions and response options are presented below:

**Question 1**

‘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 hypothesis being tested is:

\[ H_{a1} : \text{The pattern of frequencies is significantly different from a random pattern; there is a greater probability that respondents will select the ‘crime is increasing’ option.} \]

A frequency table (Fig. 2) and bar chart (Fig. 3) containing response data, along with additional descriptive statistics are presented below:\(^3\):

- Measurement level: Nominal
- Sample size: \( N = 5,366 \)
- Mode: 1 (‘Increasing’)

The results show that 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 11,960.827 (df = 3, \( p < 0.001 \)), indicating this pattern of responses is significantly different from that which could be expected through chance alone. The appropriate effect size estimate, Cohen’s \( w \) (Cohen, 1988; Tapanes, 2008) was found to be 1.49, which is categorized as a very large effect. **Therefore, \( H_{a1} \) is likely supported.**

\(^3\) Although \( N = 4,917 \) for fully completed surveys, additional data from partially completed survey responses were also retained for analysis. As SPSS automatically disregards missing values during analysis, it is logical and prudent to make available all valid data. Furthermore, as non-parametric methods are used to analyse the data, differences in sizes of comparator data sets do not destabilize analysis. Consequently, for some tests, \( N \geq 4,917 \).

<table>
<thead>
<tr>
<th>Crime is:</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increasing</td>
<td>4806</td>
<td>89.6</td>
</tr>
<tr>
<td>Decreasing</td>
<td>41</td>
<td>.8</td>
</tr>
<tr>
<td>Stable</td>
<td>189</td>
<td>3.5</td>
</tr>
<tr>
<td>Don’t Know</td>
<td>330</td>
<td>6.1</td>
</tr>
<tr>
<td>Total</td>
<td>5366</td>
<td>100.0</td>
</tr>
</tbody>
</table>

**Figure 2:** Frequency table (Q1).

**Question 2**

‘As a result of the information contained in the table, how likely is it that you would feel concerned about the crime rate?’

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

The hypotheses being tested are as follows:

\[ H_{a2} : \text{The pattern of frequencies is significantly different from a random pattern; there is a greater probability that respondents will enter higher scores, indicating concern.} \]

\[ H_{a3} : \text{The patterns of frequencies produced in Question 2 by the group who selected the ‘Increasing’ option in Question 1, and the remainder of respondents, are significantly different; the former group is more likely to enter higher levels of concern than the latter.} \]

A frequency table (Fig. 4) and histogram (Fig. 5) containing relevant data, along with
The results show that 51.7% of respondents (2,719 of 5,256) recorded they were ‘Likely’ to feel concerned and a further 12.7% (667 of 5,256) were ‘Very likely’ to feel concerned (totalling 3,386 of 5,256, or 64.4%). Both median and mode = 4 (‘Likely’), with the distribution curve exhibiting negative skewness (−0.600), due to a weighting towards the higher scalars on the right hand side (i.e. greater concern). A ‘goodness-of-fit’ chi-square test produced output of 3,772.096 (df = 4, \( p < 0.001 \)), indicating this pattern of responses is significantly different from that which could be expected through chance alone. Cohen’s \( w = 0.85 \), indicating a large effect.

Therefore, \( H_{a2} \) is likely supported.

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, whereas those who chose
any of the remaining options were coded into a separate group. The purpose of this exercise 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 the homogenous group of remaining respondents? The results are displayed below.

Output from the test shows the distribution of ranks in the two groups, as well as the mean rank in each condition. It is apparent that the 'Increasing' group has the higher mean rank (2,757). \(N = 5,256,\ U = 662,898,\ z = -19.745,\ p < 0.001;\) therefore, it can be stated that there is a statistically significant difference between the groups, with members of the group who believed crime was increasing being significantly more likely to experience higher levels of concern than other respondents (Fig. 6).

The appropriate effect size estimate, \(r\) (Rosenthal, 1991, p. 19) was found to be 0.27, which is in the small-to-medium effect size range (Cohen, 1988; Field, 2013, p. 82).

Therefore, \(H_{a3}\) is likely supported.

**Question 3**

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

Respondents were invited to record responses on a series of 5-point Likert-type scales, indicating how likely it was that they would engage in a range of specified behaviours, as below:

- Do nothing.\(^4\)
- Ask for an explanation about performance.

\(^4\) The variable ‘Do nothing’ was reverse coded to enable aggregation of the variables pertaining to this question.
Communicate an expectation there should be an improvement.

Initiate an operational response (e.g. commission further research, change tactics, and deploy resources).

Output from these variables was combined to produce an aggregate variable, denoting the extent of DUBR.

The hypotheses being tested are as follows:

$H_{a4}$: The pattern of frequencies is significantly different from a random pattern; there is a greater probability that respondents will enter higher scores, indicating propensity to enact disproportionate / unwarranted behavioural responses.

$H_{a5}$: The patterns of frequencies produced in Question 3 by the group who selected the ‘Increasing’ option in Question 1, and the remainder of respondents, are significantly different; the former group is more likely to enact higher levels of disproportionate / unwarranted behavioural responses than the latter.

$H_{a6}$: The patterns of frequencies produced in Question 3 by the group who selected the ‘Likely’ or ‘Very likely’ options in Question 2, and other respondents, are significantly different; the former group is more likely to enact higher levels of disproportionate / unwarranted behavioural responses than the latter.

$H_{a7}$: The patterns of frequencies produced in Questions 1 and 2 significantly predict whether respondents are likely to enact disproportionate / unwarranted behavioural responses.

Descriptive statistics and a histogram (Fig. 7) for the aggregate variable are produced below:

- Measurement level: Ordinal
- Sample size: $N = 4,917$
- Range: 17 (4–20)
- Median: 15
- Mode: 16
- Skewness: $-0.759$
- Standard error of skewness: 0.035

Visual inspection of the histogram finds a distinct ‘pile up’ to the right. The median = 15;

---

5 Each variable was also analysed individually, although the results are not reported herein due to limited space.
mode = 16 (midpoint of range = 12). The distribution curve exhibits strong negative skewness (−0.759), demonstrating a weighting towards the higher scalars on the right hand side (heightened propensity to enact DUBR). A ‘goodness-of-fit’ chi-square test produced output of 3,717.634 (df = 16, \( p < 0.001 \)), indicating this pattern of responses is significantly different from that which could be expected through chance alone. Cohen’s \( w = 0.87 \), indicating a large effect.

Therefore, \( H_{a4} \) is likely supported.

A Mann–Whitney U test for independent samples was then conducted to establish if there was a significant difference in the way the two different groups reacted to the options in Question 3. Respondents who selected the ‘Increasing’ option in Question 1 were placed into one group, whereas those who chose any of the remaining options were coded into another group. The purpose of this exercise was to establish if those who interpreted the stimulus to mean crime was increasing are more likely to enact DUBR than the remainder of respondents. The results are displayed below.

Output from the test shows the distribution of ranks in the two groups, as well as the mean rank in each condition. It is apparent that the ‘Increasing’ group has the higher mean rank (2,562.82). \( N = 4,917, \ U = 649,818, \ z = -15.314, \ p < 0.001 \); therefore, it can be stated there is a statistically significant difference between the groups, with members of the group who believed crime was increasing
being more likely to enact DUBR than other respondents. $r = 0.22$, which is in the small-to-medium effect size range (Fig. 8).

Therefore, $H_{a5}$ is likely supported.

A new variable was then computed to enable freshly defined groups to be compared. Respondents who selected the ‘Very unlikely’ or ‘Unlikely’ options for the ‘emotional response’ variable in Question 2 were placed into one group, whereas those who chose the ‘Likely’ or ‘Very likely’ options were coded into a separate group. ‘Don’t know’ responses from Question 2 were discounted. A Mann–Whitney U test for independent samples was then conducted to establish if there was a significant difference in the way two different groups responded to the options in Question 3. The results are displayed below.

Output from the test shows the distribution of ranks in the two groups, as well as the mean rank in each condition. It is apparent that the ‘Likely’/‘Very likely’ group has the higher mean rank ($2,645.41$).

$N=4,431$, $U=3,341, 473.5$, $z=36.36$, $p<0.001$; therefore, it can be stated there is a statistically significant difference between the groups, with members of the group who entered higher levels of concern being more likely to enact DUBR than other respondents. $r = 0.55$, which signifies a large effect (Fig. 9).

Therefore, $H_{a6}$ is likely supported.

Regression

To test whether the variables at Questions 1 and 2 act as predictor variables for behavioural outcomes (see $H_{a7}$), binary logistic regression analysis was conducted (Brace et al., 2009; Field, 2013, pp. 760–813). A dependent variable was generated by recoding the aggregate ‘behavioural response’ variable from Question 3. The dependent variable is conceptualized as respondents’ disposition towards DUBR.

Individual categorical variables were computed for each of the four options in Question 1.
(i.e. ‘Increasing’, ‘Decreasing’, ‘Stable’, or ‘Don’t know’). These variables were placed alongside the ‘emotional response’ variable produced at Question 2, and regression was then conducted against the dependent variable ‘DUBR’. A total of 3,334 cases were analysed and the model was found to significantly predict behavioural responses, with a medium effect size (omnibus $\chi^2 = 995.805$, df = 2, $p < 0.001$, Nagelkerke’s $R^2$ approximation = 0.386).

A Hosmer and Lemeshow test confirmed the model fits the data well ($\chi^2 = 3.045$, df = 3, $p = 0.385$). A

The model underwent three iterations, following the principles of parsimony (Field, 2013, p. 768); non-significant predictors were stripped away to produce the optimal model. It was found that the categorical variable ‘Increasing’ and the ‘emotional response’ variable most reliably predicted the DUBR dependent variable. The final iteration of the model correctly predicted 87.7% of respondents likely to enact DUBR and 68.1% of those unlikely to do so. Overall, 83.0% of predictions were accurate.

Output for this model is displayed in the Fig. 10. It can be seen that the model significantly predicts the likelihood of DUBR being enacted as a consequence of how respondents interpret crime data presented in binary comparison format. Where respondents believed crime to be increasing, the Exp(B) column indicates there is 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 ‘emotional response’ predictor variable, there is an increase in the odds of DUBR being enacted by a factor of 3.223 (95% CI = 2.951 and 3.520).

Therefore, $H_a$ is likely supported.

**Summary and inferences**

This study produces compelling evidence supporting all of the hypotheses, with extremely high levels of statistical significance. It provides an evidence base which demonstrates empirically that the use of binary comparisons in the police performance environment leads to highly counterproductive behaviours.

In summary, the analysis found:

1. An overwhelming proportion of respondents (89.6%) misinterpreted the binary comparison stimulus to mean crime was increasing.
2. Those who believed crime was increasing were significantly more likely to experience higher levels of concern.
3. Those who believed crime was increasing were significantly more likely to enact DUBR.
4. Those who experienced higher levels of concern were significantly more likely to enact DUBR.
5. The model produced during logistic regression confirmed variables at Questions 1 and 2 significantly predicted behavioural responses at Question 3.

---

**Figure 10:** Logistic regression statistics table.

<table>
<thead>
<tr>
<th>Variables in the Equation</th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>Sig.</th>
<th>Exp(B)</th>
<th>95% C.I. for EXP(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower</td>
</tr>
<tr>
<td>Binary_Increasing_only</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(t)</td>
<td>.623</td>
<td>.163</td>
<td>14.528</td>
<td>1</td>
<td>.000</td>
<td>1.564</td>
<td>1.353</td>
</tr>
<tr>
<td>Q49_1</td>
<td>1.170</td>
<td>.045</td>
<td>676.314</td>
<td>1</td>
<td>.000</td>
<td>3.223</td>
<td>2.951</td>
</tr>
<tr>
<td>Constant</td>
<td>-3.248</td>
<td>.187</td>
<td>303.185</td>
<td>1</td>
<td>.000</td>
<td>.039</td>
<td></td>
</tr>
</tbody>
</table>

*This statistic is a test of the null hypothesis that the model is a good fit; therefore it is significant due to the $p$-value being well in excess of 0.05 (Brace et al., 2009, p. 332; Field, 2013, pp. 765, 876).
The research demonstrates that binary comparisons act as a catalyst for groundless assumptions about (potentially non-existent) trajectories within data, leading to undue concern, impaired decision-making, and increased likelihood of DUBR. Critically, due to the rigorous design and implementation of the research, these findings may be generalized to the population.

Furthermore, as respondents were not exposed to external influences likely to exacerbate adverse reactions (e.g. management pressure), the findings suggest the practice of using binary comparisons in the police performance environment is fundamentally illegitimate by design. Therefore, any notion that dysfunction arises only as a consequence of improper application is rejected.

Further considerations

During the design phase of the research, consideration was given to including additional variants of each stimulus, purporting to convey diverse trajectories. However, this format was ultimately discounted as it would have significantly extended the length of the survey. Nevertheless, this may present an opportunity for future research, using an instrument that tests multiple variants of a single thematic stimulus.

Furthermore, it may be instructive to replicate these tests in non-police performance contexts. From a purely speculative standpoint, it is suggested that the use of binary comparisons in non-police performance environments might well lead to similar outcomes. Therefore, further research in other public and private sector settings may elicit deeper understanding of associated behavioural phenomena.

Finally, it is acknowledged that the binary comparison element of the wider doctoral study highlights deficiencies without necessarily providing solutions (save to advise cessation of the practice). A further paper is being written, which reports findings from a parallel study into the predictability of behaviours triggered by the use of Statistical Process Control (SPC) in the police performance environment; these findings may point the way to improved professional practice.

Conclusion: so what?

It is proposed that this research provides a solid starting point from which to argue strongly against the use of binary comparisons in the police performance environment. The study highlights the highly predictable nature of associated adverse behavioural outcomes and should, therefore, serve as a warning to practitioners, as well as to inform future policy and practice.

Pettigrew (1995, 1997), Hodgkinson and Rousseau (2009), and Hodgkinson and Starkey (2011) argue for a research ethic where methodological rigour and practical relevance are both high. Striking this balance ensures that theory informs practice, making research both academically rigorous and socially useful.

It is hoped, therefore, that this study offers fresh insights capable of fostering a richer understanding of predictable behavioural phenomena in the police performance environment, as well as broader management implications for the design and implementation of information policy in agency models.

In conclusion, the following is proposed.

Theoretical Proposition

If binary comparisons are used within the police performance environment, the following tendencies are highly predictable:

- Individuals are highly likely to inappropriately ascribe meaning to the binary comparison (i.e. to assume the two isolated values represent a trajectory), and therefore draw unwarranted conclusions.
- As a consequence, individuals are highly likely to experience an emotional response that influences their decision-making (i.e. to become unduly concerned if the perceived trajectory...
appears to indicate a different orientation to the desired direction of travel).

- Individuals are highly likely to enact DUBR as a consequence of (1) and (2).

References


