Performance Appraisal: The Policy Capturing of Sergeants in the New Zealand Police Service

A thesis presented in partial fulfillment of

the requirements for the degree of

Master of Science in Psychology

at Massey University

Sharon Mary Rippin

1986

ABSTRACT

This study investigated the policy of sergeants for combining and weighting performance appraisal information about constables. The experiment was conducted in several steps. In the first step constables and sergeants were interviewed about performance dimensions that were necessary for the job of constable. Twenty six constable performance dimensions were identified. Sergeants then rated between eight and ten of their constables on each of the 26 performance dimensions. Factor analysis was used to identify the sergeants underlying performance weighting structure. Eight factors were identified that explained 79% of the total variance. In Step Three behavioural examples of constable performance for each of the eight factors were generated. In Step Four sergeants assigned grades to 60 hypothetical constable protocols which were made up of the statements generated in Step Three. Sergeants also estimated how much weight they felt they assigned each of the eight factors when rating the protocols. A multiple regression equation was computed for each sergeant. Sergeants were found to use four to five factors when assessing constable performance with one factor contributing over half the variance. They were not consistent as a group when rating constables, in terms of the factors they used and their corresponding weights. They also had little insight into their rating policies. Implications of the results for the police's current performance appraisal system are discussed.

ACKNOWLEDGEMENTS

First I would like to thank Mike Smith my supervisor for his practical down to earth approach to research, sense of humour and his infuriating ability to dissolve my latest crisis with the most obvious suggestion.

Thanks also to the staff at Police National Headquarters, in particular Chief Inspector Preston Shaw, Dr Ian Miller and Superintendent Joe Farrow for their comments and help in getting the study underway.

I'm also greatly indebted to Dougal Stewart, (which he won't let me forget) for his criticisms and ideas.

Josie Smith definitely deserves a mention, mainly because she never hid when I continually badgered her about SPSS-X.

Thanks also to Mum and Dad. Mum- for lending me the cats, your encouragement and for constantly reminding me that all work and no play makes Sharon a dull girl. Dad- for your continuing assistance even though times have been tough.

Perry, what can I say but thanks for your enduring patience and support.

TABLE OF CONTENTS

Abstract i

Acknowledgements ii

List of Tables and Figures

Chapter One : Overview of Performance Appraisal 1

Chapter Two : Police Performance Appraisal 12

Chapter Three : Overview of Present Study 31

- Hypotheses 50

Chapter Four : Method 51

- Subjects 51

- Experimental Steps 51

- Step One 51

- Step Two 53

- Step Three 55

- Step Four 56

Chapter Five : Results 60

Chapter Six : Discussion 72

Chapter Seven : Summary and Conclusions 84

References 87

Appendicies

Appendix A : Police 204 Rating Form 104

Appendix B : Patrol Officer Performance Dimensions 109

Appendix C : Patrol Officer Factor Loadings 116

Appendix D : Multiple Regression Weights (United States Study) 117

Appendix E : Constable Job Analysis 119

Appendix F : Constables' Performance Appraisal 125

Appendix G : Definitions of the Eight Performance Factors 156

Appendix H : Generation of Constable Behaviour Statements 158

Appendix I : Performance Evaluation Questionnaire 170

Appendix J : Presentation of Factor Loadings for the Present Study and the United States Study 194

LIST OF TABLES AND FIGURES

TABLES

Table 1: Studies that have captured judgemental policies 36

Table 2: Constable performance factors generated in Step 1(A) 61

Table 3: Factor loadings (>0.5) for constable performance dimensions from principal components factor analyses after varimax rotation 63

Table 4: Unstandardised regression weights for the constable performance factors for 57 sergeants 65

Table 5: Significance levels of the B weights for the eight constable performance factors across 57 sergeants 68

FIGURES

Figure 1: Process model of performance rating (Landy & Farr, 1980) 8

Figure 2: Plot of the eigenvalues against factors (Scree Test) 54

Figure 3: Sergeants' unstandardised mean b weights and standard deviations for the eight factors. 67

Figure 4: Sergeants' mean estimates and standard deviations of weights (out of 100) they assigned the eight factors 67

Figure 5: Correlations between estimated factor importance and b weights across the eight factors. 71

CHAPTER ONE

OVERVIEW OF PERFORMANCE APPRAISAL

There is no escape from performance appraisal. It is impossible to go through life without being assessed in some way. Famous examples of assessment come readily to mind; Henry the Eighth judged his wives on their ability to produce male heirs (Fisher, 1913). Spartans assessed their new born babies on their ability to withstand a night in the cold (Eaton, 1970). Today, we are assessed early in life by the plunket nurse, kindergarten and school teachers, and later, by the bank manager, lecturers, dance instructors and many others. We often assess people who provide us with services such as doctors, chefs and hairdressers and act on our judgement of their effectiveness to decide whether we continue to use their services.

To appraise anything is to set a value on it. The purpose is to find out how a person measures up when compared with some standard of performance. The most common and frequent type of performance appraisal takes place in the work setting. Performance appraisal systems are constructed with the understanding that performance evaluations represent meaningful distinctions among employees that correspond to actual behavioural differences (Wendelken and Inn, 1981). The overall aim of the appraisal is to remove the influence of extraneous factors from the evaluation process in order to focus solely on aspects of performance that are related to some specific criterion.

There are a number of uses for the assessment of work performance. The most general is for administrative personnel decisions such as promotions, salary increases, and layoffs. Cummings (1973) has termed this as one of providing structure for a reward/punishment system. He also suggests that there are at least three other uses for performance appraisal systems - (1) providing criterion information for the selection process (2) providing objectives for training programmes (3) providing elements for supervisory feedback and control. Overall, performance appraisal plays an important role in all personnel decisions.

Organisations continue to express disappointment in performance appraisal systems despite advances in technology (Banks and Murphy, 1985). Reliability and validity remain major problems and new appraisal methods are often met with substantial resistance. In essence, effective performance appraisal in organisations continues to be a compelling but unrealized goal.

Over the past 35 years, researchers have developed several methods to assist performance appraisal in organisations. Contributions fall within three general categories: appraisal formats, training programmes for raters, and appraisal processes (Banks and Murphy, 1985). Researchers developed numerous formats such as checklists, rating scales, narratives, and work samples that help structure the appraisal (Bernardin and Beatty, 1983; Carroll and Schneir, 1982). Formats aid actual appraisals by determining the type and number of dimensions assessed, the types of judgements made, appraisal length,

and comprehensiveness. Some researchers also argue that particular formats guide appraisal judgements (Bernardin and Smith, 1981). Rater training programmes were designed to promote proper utilization of appraisal systems and to improve rating skills. Some of these training programmes incorporate learning principles such as practice, feedback, and active participation (Spool, 1978) and emphasize behavioural observation (Boice, 1983; Thornton and Zorich, 1980). Various approaches were developed to assist the appraisal process. Examples of these approaches are the critical incident method (Flanagan, 1954), diary-keeping (Bernardin and Walter, 1977), participation in format development (Friedman and Cornelius, 1976), and goal setting (Latham and Locke, 1979). These approaches, as well as others, consist of a set of techniques appraisers can use to help them generate valid ratee data.

Such methods are useful in an ideal sense because they promote (but do not guarantee) systematic, job-related, and relatively error free evaluation. However, they have not been adopted widely (DeVries, Morrison, Shullman, and Gerlach, 1981). For the most part, the appraisal systems actually used in organisations have failed to draw on this body of research.

Landy (1985) states that ideally, complete performance measurement should include the combination of three indices of performance - objective data, personnel data and judgemental data . The multi-dimensionality of job "performance" only becomes apparent when these categories are considered simultaneously. For example, is a successful worker one who

turns out the greatest number of units (objective data), one who has not been absent for 27 years (personnel data), or one who is rated highly on quality of work by a supervisor (judgemental data)?

There are several problems with objective data. It is difficult to measure reliably in that each objective measure probably has an unstable observation period. For example, if we take the total number of tickets issued by a traffic officer over a one week period, the relationship between one week and another could depend on a number of factors eg. what shift was assigned, what area was patrolled, time of year etc. The fact that the nature of work is also changing makes it difficult to collect objective data. For example, a major change is the increase in automation in industry where workers who were once operators are now observers. If only objective data is considered for people who observe machines then no differential performance data on these individuals could be obtained unless a machine malfunctions. Another problem is that many workers tend to work in groups such as in car assembly plants, making it difficult to collect individual data. There are also many jobs for which no good objective measures are available eg., manager. There are no clear indicators of what makes one manager better than another?

Personnel data also has weaknesses. This data includes variables such as tardiness, absences, type of salary adjustment, number of accidents etc. Almost all these measures tend to reflect the climate of the organisation, but are rather global in nature. Often the classifying and recording of personnel data is poorly performed. One such example

is the recording of absences, in that they may be either absolute number of days not at work or number of absences regardless of the length of each absence. Latham and Pursell (1975) suggest it may make more sense to measure attendance rather than absences. Overall, personnel data tends to fall prey to the potential confounding effects of other variables in much the same manner as described for objective data.

All this does not imply that objective and personnel data have no value as criteria, but rather, if they are to be useful, a careful analysis of the relationship between the elements of the job as identified by job analysis and elements of behaviour as related to performance appraisal is necessary (Landy, 1985).

Judgemental data is the most frequently used form of measurement. Landy (1985) reported that a literature review of validation studies in the Journal of Applied Psychology between 1965 and 1975 revealed that ratings were used as the primary criterion in 72% of the cases. These judgements can take several forms. They may be a simple comparison of one employee with another, a list of statements which are applied to each employee, or some form of rating by which the employee is placed on a continuum depending on their level of proficiency.

By far the most widely used judgemental measure is the rating scale. These scales can be distinguished from each other on three different dimensions (Guion, 1965). The first dimension is the degree to which the meaning of the response category is defined. This deals with how

the rating scale is marked off into units, whether it is numerical or descriptive. The second dimension is the degree by which the person interpreting the scales can tell what response was intended by the ratee. Response clarity is largely determined by the structure of the scales. The third dimension is the degree to which the performance dimension being rated is defined for the rater. Scale anchors that are defined precisely are less open to misinterpretation and give the rater a reasonable idea of what performance dimensions are being considered.

In spite of the different forms and widespread use of judgemental indices of performance there has been constant dissatisfaction with these measures on the part of the researcher and practitioner. The major source of dissatisfaction can be largely attributed to three types of rating errors- halo, central tendency and leniency errors (Anastasi , 1982). Halo errors occur when a rater has a generally favourable or unfavourable impression of the person to be rated. Ratings are therefore assigned which are consistent with that impression. No method has been devised that effectively eliminates halo errors , and research on alternative solutions still continues (King, Hunter and Schmitt 1980; Landy, Vance, Barnes-Farrell and Steele 1980). The second type of error central tendency, is characterised by an unwillingness by the rater to assign extreme ratings, both high and low. Leniency error, the third type of error refers to the reluctance on the part of many raters to assign favourable and unfavourable ratings. This results in ratings being bunched up towards the lower and upper ends of the scale. Both leniency and central tendency errors reduce the effective width of the

scale and make ratings less discriminative (Anastasi , 1982). An enormous amount of research has been conducted in an attempt to minimize the effects of these errors by using alternative evaluation schemes. A brief examination of the research demonstrates that the process of appraising performance is incredibly complex, with many opportunities for the ratings to be influenced by factors other than the performance of the ratee.

Researchers in the area of performance appraisal have concluded that a model is necessary before any significant advances can be made in understanding judgemental performance measures (DeCotiis, 1977; Kane and Lawler, 1978; Zedeck, Jacobs and Kafry 1976). Landy and Farr (1980) proposed a process model that suggests the effects of various components on the overall accuracy of ratings (see Figure 1). It is important to keep in mind that the goal of performance rating is to provide an accurate performance description of the ratee. In this model, it is represented as the box on the right hand side labelled "Performance Description". All the other boxes may be thought of as potential obstacles to accurate performance appraisal. They act as filters, systematically distorting the attempt by the rater to accurately describe the job-related behaviour of the ratee.

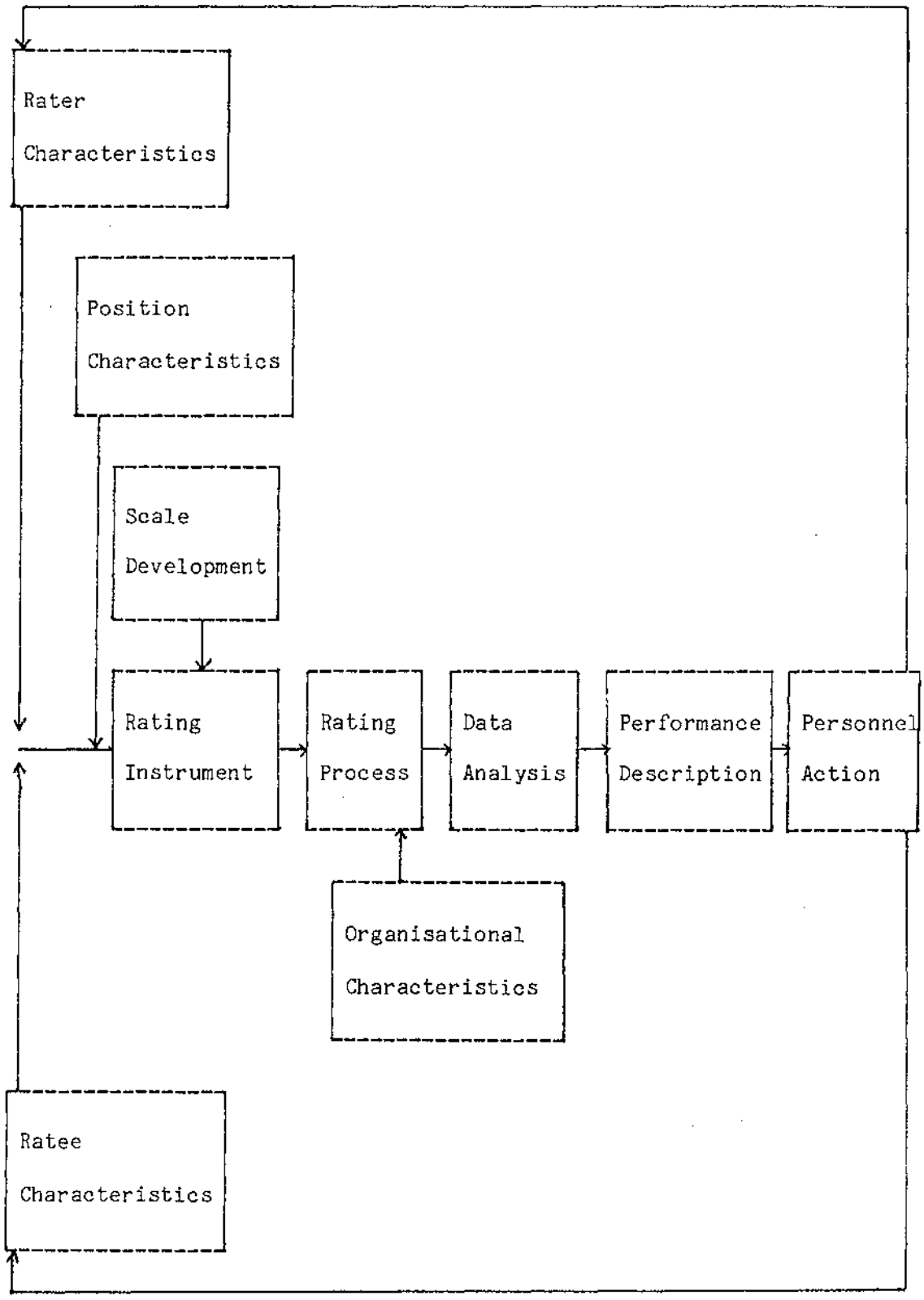

Figure 1: PROCESS MODEL OF PERFORMANCE RATING (Landy and Farr, 1980)

The model assumes that there are certain characteristics brought to the rating task that are properties of the ratee and the rater. For example, the rater introduces biases that may be related to age, sex, race, leadership style, personal relationship with the ratee, and so forth. In addition to the main effects of raters and ratees, there are undoubtedly interactions of rater and ratee characteristics. For example, DeJung and Kaplan (1962) and Hamner, Kim, Baird and Bigoness (1974) found that ratees who were the same race as the rater received higher ratings than ratees of a different race. Other rater-ratee characteristics that interact may include factors such as education, previous experience with performance rating, and tenure in the organisation. Several factors interact to influence the overall accuracy of performance description as seen in Figure 1. The position the person to be rated holds in the organisation is a factor that affects the choice and/or development of a rating instrument, and the purpose for which rating is done. It is not uncommon to see ratings used to make administrative decisions at one level in an organisation but used for counselling at another level.

A conceptually independent variable in the system is the instrument actually used to gather the performance information. Through a process of scale development, or selection, an instrument is identified that presumably is capable of helping raters make distinctions among ratees with respect to the various categories of behaviour. The scale development may involve developmental groups as in the case of the Behaviourally Anchored Rating Scales (BARS) methodology, or item analysis derived from a study of current employees, as in the case of

Summated Ratings or Forced-Choice Inventories. Regardless of the method of development, an instrument will be selected or constructed to produce judgements about performance. The component labelled rating process refers to the constraints placed on the rater by requests or demands. For example, when raters are faced with a short length of time to make a judgement about someone, individuals tend to use fewer sources of information and to weigh unfavourable information more heavily in making evaluations (Wright, 1974). The organisation in which the ratings are gathered might have certain characteristics that also influence the accuracy of ratings eg., turnover levels, part-time to full-time employee ratio, and seasonal variation in the work force. After ratings have been gathered, the data are analysed to produce accurate and reliable performance descriptions. Various analytic techniques have been shown to be more successful at reducing or eliminating rating errors than other techniques (Landy and Farr, 1980).The combination of all these elements discussed above produce a performance description. On the basis of this information certain personnel actions are implemented either actively or by default eg., selection systems are maintained or changed, salaries or work force levels are altered, employees are told of weaknesses and strengths.

While this model does not offer much in the way of an explanation as to why these elements may have adverse effects on the accuracy of performance appraisal, it does provide a view of the complexity of the rating process.

One area of performance appraisal that is particularly complex and important, is the rating of police men and women. The Police Service is one of the larger employers in New Zealand, and its performance needs to be carefully monitored to ensure the well-being and protection of society. In the next section an attempt will be made to look at the work that has already been done in the area of police performance appraisal. Difficulties that have been encountered in the assessment of police performance will be highlighted.