An Exploration of the Validity and Reliability of 'Managerial Reading Assessment' – A Cognitive Ability Test

A thesis presented in partial fulfilment of the requirements for the degree of

Masters of Arts in Psychology

at Massey University

Mary Alice O'Hare

1999

ABSTRACT

Cognitive ability tests are generally considered in the empirical literature to be one of the most valid predictors for selecting managerial level staff. However, very few of these tests have been specifically designed and developed for managers. Managerial Reading Assessment (MRA) is an original cognitive ability test which has been created for this purpose. Because critical thinking skills, particularly the ability to draw inferences, are regarded as being crucial to the successful performance of a manager's job, this test specifically targets this skill. The present study investigated the validity and reliability of Managerial Reading Assessment (MRA) to assess its potential as a selection test for managers. A total of 97 voluntary participants, the majority of whom were drawn from junior to senior levels of management, were recruited from their place of work to take part in this research. Respondents were asked to complete the test and return it by mail. To evaluate the validity of the MRA, two criterion measures (salary and highest educational level achieved) were adopted. When education was utilised as the criterion, a validity coefficient of rho=0.39 was obtained, significant at the 0.01 level. The size of this correlation is comparable to those obtained for other cognitive ability tests. The internal consistency of the test was computed using the alpha coefficient. The results indicate that this test is also reliable. More study would need to be conducted to further assess the psychometric properties of this test.

ACKNOWLEDGEMENTS

Many thanks to my supervisor Associate Professor Douglas Paton for granting me the freedom to pursue independent research whilst at the same time always being available for interesting discussions, to answer questions and for making hilarious jokes! Thank you. I would also like to thank Dr Duncan Hedderely for his help with the statistical consultation. Thanks also to friends and flatmates for their support and kindness. I am also deeply grateful for the generosity of all of the participants who took part in this research – without their contribution this Masterate would not have been possible.

TABLE OF CONTENTS

LIST OF TABLES

INTRODUCTION

The goal of management selection is to hire the applicant most likely to succeed on the job. To achieve this, managers are typically appraised using a multiple hurdle approach involving several selection devices. These devices can include interviews, aptitude tests and reference checks (Cascio, 1991). Of these, cognitive ability tests are one of the most superior instruments to use (Salgado, 1999). However, few of these cognitive ability tests have been expressly tailored to meet the needs of managerial selection.

The present research builds on a project begun in 1997 in which an original cognitive ability test, Managerial Reading Assessment (MRA), was designed and developed (O'Hare, 1997). Because decision-making, and the concomitant skill of drawing valid conclusions, is a vital component of the manger's job, the MRA specifically targets this ability. The aim of the present study is to explore the validity and reliability of the MRA test to assess it's potential for selecting managerial level staff.

The Literature Review, which follows, is divided into two parts. The first part comprises a description and evaluation of the more common selection devices used to select managerial level staff. The predictors that are presented here are interviews, assessment centres, work sample tests, and cognitive ability tests. Of the cognitive ability tests currently used to assess potential managers, the Watson

Glaser Critical Thinking Appraisal is probably the most widely used (Ryan & Sackett, 1987). This test has been used to predict the performance of executives, managers and other technical and professional employees who are required to think critically or analytically in the course of their job (Watson & Glaser, 1994).

The next chapter then leads onto a more detailed analysis of the construct of critical thinking and inferencing. These skills, in particular the ability to draw inferences, are considered crucial for analysing information and arriving at good conclusions. Because decisions are the outcomes of critical thinking and inferencing and are a key component of the manager's job, the next chapter explores the research regarding how people make decisions and what constitutes a "good" decision.

In the construction and development of the Managerial Reading Assessment (MRA) test, which assesses decision-making (specifically the ability to draw inferences), a number of psychometric factors had to be considered. The second part of the Literature Review outlines these factors. The points covered are the reliability and validity of the test, as well as an assessment of the criterion measures used to assess the statistical properties of the test. These are important factors to be considered in the construction of any selection test if it is to be used with any degree of confidence for predicting future performance on the job.

LITERATURE REVIEW

PART 1
MANAGEMENT SELECTION

The selection of managerial level employees typically requires candidates to undergo a series of selection tests, which are used to predict their future performance on the job. The most common predictors included in this multiple hurdle approach are interviews, cognitive ability tests and personal history forms (Cascio, 1991). Due to the high economic utility of hiring managerial level staff, Rudman (1991) asserts that employers wish to obtain as full a picture as possible about prospective employees, hence more comprehensive assessment procedures are undertaken for this calibre of staff, to assist them in the decision-making process. The selection devices that are utilised have varying degrees of validity and usefulness for this task.

INTERVIEWS

Until very recently, the majority of the research that has been conducted on the employment interview has concluded, almost unanimously, that this predictor does not show much evidence of validity for predicting future performance on the job. Despite this, interviews continue to be used extensively for staff selection.

Reilly and Chao (1989) state that although interviews are extensively used to select staff, they have uses other than selection. For example, they are often used as a communication vehicle between employer and employee, enabling the employer to

provide information about the job and seek additional information from the potential employee, as well as allowing candidates to ask questions.

Practitioners' Perceived Validity of Interviews and Their Reported Use of Interviews

Despite the low validities attached to the interview, many human resource personnel view it as being more valid than it actually is and ranked it as their most frequently used selection device (Dakin & Armstrong, 1989). Taylor, Mills and O'Driscoll (1993) undertook two surveys in New Zealand to ascertain which selection devices were utilised by Senior Human Resource personnel. They too found that interviews were among those employed the most often and that they were regarded as having higher validities than they actually do.

In the U.S. Harris & Dworkin (1990) reported that although the human resource managers in their sample did not rate unstructured interviews as being amongst the 3 most valid predictors, they were the second most used screening tool. The researchers suggest that this finding may indicate that interviews are regarded more as a "communication device rather than as a screening procedure" by human resource managers

Structured and Unstructured Interviews

Structured interviews have consistently returned higher validities and are more reliable than unstructured interviews. Wright, Lichentenfels and Pursell (1989) suggest a number of reasons to account for the higher validity of structured interviews: the interview questions are highly job related (being based on a job analysis), candidates are all asked the same questions, and their responses are scored against answers that have been previously agreed upon as indicative of different levels of performance.

Unstructured interviews, on the other hand, are more free ranging and non-directive. This could introduce error as candidates may be assessed on answers to questions that are not related to the job (Wright et al., 1989). In addition, each candidate may be asked different questions (Latham, Saari, Russell & Campion, 1980) which may elicit different types of information, and this could result in biased evaluations.

Two specific types of structured interviews that have been investigated are the situational interview and the behavioural index interview. The situational interview requires candidates to indicate how they would react or respond to a hypothetical situation. The hypothetical situation that they are presented with is usually derived from critical incidents, which are job related. Situational interviews are based on the premise that people will act according to their intentions to act. Latham, Saari, Russell and Campion (1980) reported validity coefficients in the 30's for this type of structured interview.

The behaviour description interview, also based on critical incidents obtained from a job analysis, requires the job applicant to recall similar events from their past and describe how they behaved in response to them. Janz (1982) reported a validity coefficient of 0.54 for this type of interview (compared with .07 for a standard interview). Behaviour description interviews are based on the supposition that past behaviour predicts future behaviour (Keenan, 1989). However, this assumption does not take into account that people may learn from past experience, or that major life events may have also occurred in the intervening time perhaps significantly altering the person's outlook, and that both of these factors may influence future behaviour. In addition, the environment in which the past behaviour occurred and the circumstances that surrounded it, may differ significantly from the critical incident presented in the interview, making comparisons inappropriate.

Meta-Analyses of the Validity of Interviews

Reilly and Chao (1982) conducted a meta-analysis of the validity and fairness of a number of different selection methods and compared them with cognitive ability tests. They looked at 12 research studies investigating the validity of interviews and concluded that, in line with other research, interviews were neither valid nor reliable enough to be used for selection purposes. Reilly and Chao (1989) calculated that interviews carried a validity coefficient of 0.19 (using supervisor's ratings as a criterion) which is much lower than the validity coefficients reported for cognitive ability tests.

Hunter and Hunter (1984) in their meta-analysis of various predictors found a validity of 0.14 for interviews using supervisor's ratings as the criterion. When interviews were used to predict training success the obtained coefficient was 0.10. When the criterion was promotion the validity coefficient dropped to 0.08. Wright, Lichentenfels and Pursell (1989) also conducted a meta-analysis of interviews but drew a distinction between unstructured and structured interviews, which Hunter and Hunter (1984) had not done. Their study focused on entry-level jobs. They calculated the mean validity coefficient of structured interviews as being 0.39.

A more recent meta-analysis (Huffcutt & Arthur, 1994) revisited Hunter and Hunter's (1984) study and made some methodological improvements such as including more studies, differentiating between levels of structure in the interview and correcting for restriction of range. This research also comprised only entry-level jobs. Their results indicate that structure moderates the validity of the interview. Although the validity increases as the structure does, this happens only up to a certain point. After this ceiling has been reached, validity more or less remains the same. This study suggests that structured interviews have higher validities than had been previously thought.

Marchese and Muchinsky (1993, cited in Salgado 1999) also conducted a meta-analysis of the interview. They too found that the level of structure in the interview attenuates validity. They calculated the mean corrected validity of structured interviews to be 0.38, which is considerably less than that reported by Huffcutt and Arthur (1994).

Salgado (1999) comments on the variability of the results obtained from these meta-analyses regarding the degree to which interview structure impacts on the validity coefficient. Because of this inconsistency there would need to be further investigation and replication before definite conclusions could be drawn regarding the interview. It would be interesting to know on what basis studies were chosen for inclusion in these meta-analyses. Further, as these meta-analyses have been conducted on entry-level jobs, it would be interesting to conduct similar studies on managerial level studies to ascertain if the validity coefficients would differ significantly, as managerial level jobs are more cognitively complex than entry-level jobs.

Cognitive Ability Level and Interviews

A study by Huffcutt, Roth & McDaniel (1996) revealed that an interviewer's assessment of an interviewee is correlated with the applicant's cognitive ability. Huffcutt et al., (1996) found a mean corrected correlation coefficient of 0.40 for this relationship. In addition, the researchers mention that interviews that do show evidence of the ability-interview correlation have better predictive ability for later job performance. Huffcutt et al., (1996) identify several ways in which an applicant's mental ability could have an impact on the outcome of the interview. Individuals high in cognitive ability can think in more sophisticated and complex ways and have a larger amount of retained knowledge at their disposal than those with lower ability levels. This enables them to understand and answer difficult, technical or abstract questions with greater ease and competence. In addition,

persons with higher mental ability may also behave differently in an interview and may be better at presenting themselves in a favourable light than those of lesser ability.

ASSESSMENT CENTRES

Another selection device increasingly used to select managerial level staff is the Assessment Centre (Gaugler, Rosenthal, Thornton & Bentson, 1987). Spychalski, Quinones, Gaugler & Pohley (1997) conducted a survey of the use of Assessment Centres in the U.S. and found that they were primarily used for staff selection, promotion, and development. They can also be utilised for the identification of managerial potential (Moses, 1973). While assessment centres have been occasionally used to assess some non-managerial staff, they are predominantly used to assess managers (Gaugler et al., 1987). To assess the managers, a variety of evaluation tools such as in-basket exercises, leaderless group discussions, simulations, structured interviews, cognitive ability tests and personality tests are typically used (Goldstein, Yusko, Braverman, Smith & Chung, 1998). During the assessment period candidates are assessed across a number of dimensions that relate to successful managerial performance. These include such traits and abilities as leadership skills, planning and organisational expertise (Campbell & Bray, 1993), oral and written communication skills, forcefulness and decision-making capabilities (Moses, 1973). Assessment Centres allow raters to accumulate new and extra information, which may not be evident from the more traditional forms of staff selection, to assist them when making decisions about candidates (Campbell & Bray, 1993).

While assessment centre practices differ widely according to their intended purpose, as well as across industries (Spychalski et al., 1997), there are a number of moderator variables which can contribute to higher validities. Gaugler et al., (1987) in their meta-analysis of 50 assessment centres identified these factors as including

using a high number of different evaluation exercises, having a psychologist rather than a manager rate the candidates, ensuring that the gender composition of the candidate group includes a higher rather than a smaller percentage of women, and incorporating peer assessments into the rating given to candidates.

Assessment centres have been found to be quite good predictors of later management success (Moses 1973; Campbell & Bray, 1993). Gaugler et al., (1987) in their meta-analysis reported a mean corrected validity coefficient of 0.37 for assessment centres when they were used as a selection device. When assessment centres were used to predict management potential, however, their validity coefficient jumped to 0.53.

Klimoski & Brickner (1987) suggest that the reason assessment centres have predictive validity for managers may be due to the candidates' cognitive ability. They suggest that, in addition to intelligence playing a decisive role in managers' behaviour at assessment centres, their intelligence influences assessors' estimation of them. Klimoski and Brickner (1987) further state that these managers' later performance on the job will reflect these trends seen in the assessment centre. They believe that due to the mental demands of managers' work (eg analysis, reasoning, planning), intelligence is of major importance in determining success in this role. Ten years earlier, Klimoski and Strickland (1977) reported that intelligence tests predicted managerial success, more so than assessment centres. Goldstein et al., (1998) point out that some of the evaluation devices used in assessment centres, such as in-basket exercises, place a high cognitive demand on candidates due to the written content, which requires thoughtful action. In addition, some of the criteria that managers are assessed against, such as decision-making skills and problem solving, are highly cognitive in character and therefore require candidates to have a certain level of cognitive ability to perform them. Goldstein et al., (1998) suggest that since some of these assessment centre exercises do appear to tap specific

cognitive abilities, and if this is the only ability each is assessing, then perhaps it may be more appropriate to use cognitive ability tests which have been specifically designed to for that particular skill.

Smither, Reilly, Millsap, Pearlman & Stoffey (1993) investigated candidate's reactions to some of the evaluation exercises used in assessment centres. In particular, the researchers questioned candidates about their perceived face validity and predictive validity of these assessment devices. They found that applicants regarded cognitive ability tests as being very job related and as having good predictive ability, although their face validity was seen as being less than their predictive validity. Certain of the other evaluation devices, such as in-basket exercises and leaderless group discussion, were also viewed as having a high overlap with the task requirements of a manager's job. In comparison, personality tests and biodata tools were regarded as lacking in job relatedness. The researchers believe that candidates' perceptions of the predictive ability and job relatedness of the assessment procedures is important because this will colour the way applicants regard the organisation. If an organisation is viewed as attractive it will attract and retain higher calibre staff. In addition, the perception of face validity is important as it appears that face validity may have an effect (albeit small) on motivation and this may then influence cognitive ability test scores (Chan, Schmitt, DeShon, Clause & Delbridge, 1997).

There are several drawbacks associated with the use of assessment centres. One of these hindrances is their prohibitive cost to develop and run. Two researchers (Hoffmann & Thornton, 1997) found that they were approximately ten times more expensive per person than aptitude tests. In addition, assessment centre validities were smaller than those obtained on the aptitude test. Other limitations of assessment centres are the small number of people who can be effectively assessed during them (Moses, 1973) and their reported lack of construct validity. However,

despite this lack of construct validity, research has indicated their predictive ability (Joyce, Thayer & Pond III, 1994). In addition, assessment centres are usually conducted over a period of several days, which may preclude their use in terms of time and practicality, for both organisations and candidates.

WORK SAMPLE TESTS

Asher and Sciarrino (1974) specify work sample tests as being complex tests which are a "miniature replica of the criterion task" (p519). They draw a distinction between motor and verbal work sample tests. Motor tests refer to the physical manipulation or operation of objects, for example, operating a lathe. Verbal work sample tests assess both verbal and written language skills and /or relationship or people skills. An example of a verbal work sample test is the in-basket exercise given to potential managers which requires them to deal with an array of problems that one might reasonably expect to find in a manager's in-tray. Verbal work sample tests are more appropriate for managerial and administrative type positions whereas motor work sample tests are suitable for jobs having more of a psychomotor or manual component.

Asher and Sciarrino (1974) reported that verbal work sample tests obtained a mean validity of 0.45 for predicting job proficiency. They noted that work sample tests were better at predicting training success, with a mean validity of 0.55, than job proficiency. Of the eight predictors listed, verbal work sample tests ranked as the 4th most valid method of predicting job performance (they were preceded by intelligence tests, motor work sample tests and biographical information).

In their meta-analysis of the validity of various selection devices, Hunter and Hunter (1984) cite work sample tests as having the highest validity (0.54) for selection into jobs "on the basis of current performance" (ie not potential performance). Asher & Sciarrino (1974) attribute the high validity of work sample

tests to their close point to point relationship with the job in question. This occurs because the work sample test contains many elements in common with the job. Gordon and Kleiman (1976) suggest an alternative hypothesis to explain the high validity of work sample tests. They suggest that work sample tests, due to their high face validity, may increase the motivation and interest levels of candidates, which could contribute to better performance. They suggest that this may have an accumulative effect and may account for the high validity for predicting training success. They posit that this may also be a contributing factor in the lower validity, observed in the training context, for intelligence tests, which generally do not have high face validity.

Work sample tests, however, only sample current skill or ability levels (Landy, 1989). Thus, for the purposes of promotion, they would be of value if the person was being promoted to a job where the task requirements were the same or similar to that of their current job (Hunter & Hunter, 1984). In addition, care needs to be taken that the sample of work behaviour included in the test is representative. This can sometimes be "time consuming and difficult" (Dunnette & Borman, 1979). Two other considerations relate to the amount of time required for the administration of work sample tests and deciding on who would be the most appropriate "expert" to assess the candidate's work. Guion (1978) also cautions against scoring bias when grading work sample tests. Work sample tests can be assessed by either observing the process of making the product and marking the candidate on this basis, or by rating the final product. Either way, bias can be present, particularly for subjective ratings, and great care must be taken to avoid this.

COGNITIVE ABILITY TESTS

Salgado (1999) states that there are 2 avenues of inquiry in the field of cognitive ability research. The first approach is that taken by the "psychometric g

proponents" (p7) who hold that intelligence is measured by a single factor g. According to this line of thinking, specific cognitive ability tests and traditional IQ tests both appear to measure the same thing - general intelligence- referred to as g. That is, they believe that all cognitive ability tests comprise the general factor of intelligence g (to varying degrees) as well as the specific ability being tested for. Ree, Earles & Teachout (1994) found that 'g' had the highest predictive validity for performance on a work sample for US Air Force applicants. 's' (which is a combination of intelligence and experience) added incremental validity to 'g' (the amount was small but significant).

The second line of reasoning regards intelligence as comprising several factors such as personality, interest, intelligence as process, and intelligence as knowledge (known as the PPIK theory). This theory looks at intelligence from a typical performance perspective rather than one of maximal performance. Several of its advocates posit that tests of this kind should correlate with occupational performance, which also reflects typical performance. They argue that IQ tests, on the other hand, assess maximal performance (Salgado, 1999).

Notwithstanding the above arguments, tests of critical thinking which assess different skills or abilities associated with critical reasoning, such as the Watson Glaser Critical Thinking Appraisal, report moderate to high validities for predicting future performance on the job (Watson & Glaser, 1994).

Validity of Cognitive Ability Tests

Schmidt, Hunter & Pearlman (1981), using a sample of almost 400,000 people, investigated the validity of aptitude tests for positions that were within the same job family but had differing task requirements, and also for positions from different job families. Their results indicate that task differences between jobs in the same family do not affect the validity of the tests. In addition, they found that there were

only small differences in the validities of aptitude tests across job families. Schmidt et al., (1981) concluded, therefore, that aptitude tests were valid across all jobs.

Hunter and Hunter (1984), who included hundreds of research studies in their meta-analysis which investigated the validity of various selection devices, concluded that ability tests were valid across all job types. For entry level jobs Hunter and Hunter (1984) used a composite of cognitive ability tests and psychomotor tests which they labelled ability tests. They state that ability tests are the best predictor for these jobs with a mean validity of 0.53. Tests of cognitive ability were found to be more valid for 'thinking' jobs (such as managers' jobs) and that the predictive validity of cognitive ability tests increased as the cognitive complexity of the job increased. Anastasi (1988) notes that the cognitive complexity of a job relates to the increased amount of information processing and decision-making that is required to successfully complete the job. Hunter and Hunter (1984) state that cognitive ability tests have an average validity of 0.54 for predicting training success on all jobs and a mean validity of 0.45 when used alone as a predictor of future successful job performance.

Ghiselli (1973) who conducted a meta-analysis of published (and some unpublished) research studies which investigated the validity of various predictors from 1920 to 1971, reported that tests of intellectual abilities had the highest overall validity for predicting proficiency for executives, managers and administrators. They had slightly lower validity for predicting training success for this group. Hunter and Hunter (1984) present a table of Hunter's (1981) reanalysis of Ghiselli's data which place the mean validity of cognitive ability tests for managers at 0.53.

In his review of the selection research for the period 1991 – 1997, Salgado (1999) summarised by stating that research on cognitive ability tests indicates that of all selection devices they remain the single best predictors of future job performance.

Utility

Using valid selection devices (or not using them) can have a large financial impact on an organisation in terms of staff placement (or replacement), employee effectiveness and worker productivity, all of which influence organisational productivity and goal achievement. The fiscal aspect of employing valid tests is naturally, therefore, of concern to organisations (Raju & Burke, 1986). Hunter & Hunter (1984) calculated the utility of employing various selection devices. They estimated that the use of cognitive ability tests could result in savings of millions of dollars to organisations and, that substituting them with predictors of lesser validity, may result in the accrual of hefty costs. Schmidt, Hunter, McKenzie & Muldrow (1979) also attest to the large economic benefits to be gained from using valid selection tests, in terms of increased worker productivity.

Another large scale study, conducted by Schmidt, Hunter, Outerbridge and Tratnee (1986), empirically investigated the savings to an organisation that a valid cognitive ability test, used as a selection device, would create. Their sample consisted of nearly all of the white-collar workers employed by the US Federal Government. As a direct result of utilising a cognitive ability test, more effective and suitable applicants would be hired. This would result in savings from either of two outcomes. In the first situation, the longer the new hires remain on the job, the higher the overall work output is, due to their more advanced abilities. This translates into an estimated $86 million over 13 years. Conversely, employers may choose or prefer to sustain their present output level, in which case, fewer employees would be required. Thus, they could choose not to replace those who are fired, made redundant or resign. Savings from this option are calculated at $272 million per year. In addition Schmidt et al., (1986) forecast that there would be a 61 % decrease in poor performers entering the government's ranks (they define poor

performers as the lowest 10th percentile of workers). This figure represents nearly 14,000 people per year.

Notwithstanding the above impressive results, Latham and Whyte (1994) discovered that, in practice, managers are negatively influenced by utility analysis findings. When presented with the substantial savings and benefits associated with using a valid selection method, the 143 managers who were included in the study reduced their advocacy of it. When only the validity and reliability of a particular test were mentioned, however, the managers were more inclined to react positively to the psychologist's recommendations regarding its use.

Unfortunately, at present the high validity associated with cognitive ability tests is not fully realised or made use of by human resource professionals or managers responsible for hiring staff. Dakin and Armstrong (1986) surveyed 21 Human Resource consultants from around New Zealand, who hired both senior and middle management, about their use of and beliefs regarding the validity of a number of selection methods. It was discovered that cognitive ability tests were ranked as the second least valid predictor and consequently their lack of use reflected this low ranking. Another New Zealand study by Taylor, Mills & O'Driscoll (1993), showed that cognitive ability tests were infrequently used in selection by senior human resource personnel. Some of the more frequently cited reasons for this included the cost of using cognitive ability tests relative to the expected benefits, lack of support from managers and the lack of relevance of the tests to the industry that the human resource staff were recruiting for.

An overseas study by Harris and Dworkin (1990) found that cognitive ability tests were ranked at 8th place for reported use with only 32% of Human resource Managers indicating that they used them. However, of all of the selection tools that

were mentioned, cognitive ability tests were viewed as the device that was the least susceptible to 'faking'.