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

ABSTRACT	ii
ACKNOWLEDGEMENTS	iii
Table of Contents	iv
List of Tables	v
INTRODUCTION	1
LITERATURE REVIEW	
PART 1 Chapter 1: Management Selection Methods	3
Chapter 2: Critical Thinking and Inferences	18
Chapter 3: Decision Making	27
PART 2 Psychometric Properties of Tests	
Chapter 4: Reliability	32
Chapter 5: Validity	36
Chapter 6: The Criterion	41
AIMS	51
METHODOLOGY	52
RESULTS	60
DISCUSSION	67
CONCLUSIONS	76
REFERENCES	77
APPENDICES	87

LIST OF TABLES

Table 1: Summary of Mean Test Scores and Standard Deviations for Employees and Students 61

Table 2: Composition of Market Sectors 63

Table 3: Summary of Validities, Means and Standard Deviations for each Sub-sample 65

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'.

CHAPTER 2
CRITICAL THINKING

The Watson-Glaser Critical Thinking Appraisal

Of the cognitive ability tests used to select managerial level staff, the Watson-Glaser Critical Thinking Appraisal (WGCTA) is possibly the most widely used test (Ryan & Sackett, 1987). The WGCTA has been used for the selection and categorisation of managers from a wide array of industries as well as for other professional and technical jobs that require critical thinking skills (Watson & Glaser, 1994). The WGCTA is an expanded version of Watson's (1925) tests, which measure fair-mindedness. Watson devised six tests (A-F) to measure an individual's tendency to be prejudiced in their opinions. Test C is an inference test in which the person is asked to indicate whether a conclusion is definitely true or false or whether there is no data to support it. Individuals are to base their answers solely on the information provided in the text.

Glaser (1937) expanded these tests to incorporate the five skill areas currently assessed in the WGCTA. These skills are the ability to draw inferences, the recognition of assumptions, deduction, interpretation, and the evaluation of arguments. The inference section was expanded to include 'probably true' and 'probably false' options. In addition, test-takers were instructed to draw on their general knowledge when assessing conclusions that fitted these two extra options (Watson & Glaser, 1980).

Glaser (1937) originally developed the WGCTA for use in an educational context, to assess students' ability to think critically. Since then, the test has undergone several revisions. The most recent form of the test, Form S, which was published in

1994, was shortened to facilitate its use as an employment selection test (Watson & Glaser, 1994). The WGCTA is now used extensively to select staff (Ryan & Sackett, 1987) as well as assess critical thinking (Norris & Ennis, 1990).

Watson and Glaser (1980) operationally define critical thinking as consisting of five different skills. These skills are the ability to draw inferences, evaluate arguments and recognise assumptions, as well as interpretation and deduction.

DEFINITIONS OF CRITICAL THINKING

Johnson (1994) states that there are many different definitions of critical thinking. He believes that though varied, those put forward by Ennis, Siegel, and McPeck have some commonalties. All of their definitions regard critical thinking as incorporating various cognitive skills and tendencies or dispositions, and requires knowledge or information.

Ennis (1987) defines critical thinking as "reasonable and reflective thinking that is focused upon deciding what to believe or do" (p10). Ennis lists the 12 abilities and 14 dispositions that are required for critical thinking. The abilities include focusing on a question, analysing arguments, deducing and judging deductions, inducing and judging inductions, making value judgements, identifying assumptions, defining terms, asking for clarification and judging the credibility of a source or observation report. The dispositions include seeking reasons and clarity, open-mindedness, maintaining relevance, seeking alternatives, considering other people's point of view, and altering one's own stance when the facts no longer support that position.

Norris and Ennis (1990) elaborate on Ennis' definition by adding that good thinking is based on good reasons and leads to the best conclusions. Critical thinkers are people who are focused in their thinking and actively and purposively search for good reasons when assessing how reasonable their own or others' conclusions are.

Siegel (1988) defines a critical thinker as "one who is appropriately moved by reasons" (p32). He points to the connection between a critical thinker and a rational person. A rational person is someone who believes in the reasons (which have been assessed critically) and acts according to them. In addition, Siegel includes underlying character of the critical thinker – what he terms the critical spirit – as a crucial component of critical thinking. The critical spirit describes a character trait and includes the tendency, habit, willingness and commitment to think critically.

McPeck (1981) adds that a certain amount of scepticism is inherent in the idea of critical thinking in that one does not accept a given conclusion, assumption or established procedure purely because it is standard practice. A critical thinker is one who considers and weighs up alternatives. McPeck (1981) calls this 'reflective scepticism' (p7) because it is not applied indiscriminately to every situation but is used appropriately to the subject under scrutiny.

Blair (1992) includes the ability to accurately evaluate sources of information as being a component of critical thinking (as does Ennis (1987)). Blair (1992) states that we receive most of the information that forms the basis of our beliefs from other sources, rather than from our own direct experience. Therefore, this skill is of vital importance. Essentially it requires people to assess the degree of reliability and credibility of their sources of information. To do this it is necessary to ascertain whether the source had the opportunity to witness what was reported, or have the requisite knowledge or expertise to make or interpret the claim. Norris and Ennis (1990) add several other criteria such as assessing any conflict of interest the source may have, the degree to which they agree with other sources and the extent to which they follow accepted procedures. Blair (1992) asserts that the quality of an individual's judgements and decisions would be greatly enhanced if the credibility and reliability of sources of information were routinely critically appraised.

INFERENCES

Norris and Ennis (1990) posit that critical thinking has 3 principle components. The first part comprises the information and data that form the support for a decision. The second part, which forms a bridge between the first and third parts, is the ability to draw inferences. The third component is the final decision that is reached. Their model of critical thinking demonstrates the importance of the process of inference when arriving at a decision as it forms the link between the basic support for a decision (this comprises information obtained from others or personal observation, background knowledge and previously held conclusions) and the actual decision itself. Other factors, such as clarity and the disposition to think critically, also play a role in the critical thinking process.

Norris and Ennis (1990) define inference as "the process of reaching conclusions based upon reasons" (p191). They suggest that thinkers must be able not only to assess the quality of pre-existing inferences, they also need to be able to form new valid inferences. They identify three types of inferences that are used when reaching conclusions: deductive, inductive and value judging. Colberg (1985) defines deductive inferences as those that refer to conclusions that necessarily follow from the information that is given whereas inductive inferences are conclusions that are probabilistic in nature. The third type of inference, value judging, requires the person to assess the outcomes of the decision and to judge the alternatives, as well as evaluating its merits in terms of ethics and principles (Norris and Ennis, 1990).

Markman (1981, cited in Phillips 1992) regards inference as the crux of understanding as it requires a person to interpret, convert and extrapolate from the information they have been given. To do this it is important to incorporate background knowledge when making or assessing inferences (Ennis, 1987). For

instance, when a person is reading, the information contained in the text is not complete enough for adequate conclusions to be drawn about its meaning. The absence of background knowledge means that alternative hypotheses cannot be formulated to fit the situations described in the text, if the obvious interpretation does not stand up to scrutiny. Finally, without background knowledge, the reader cannot assess the evidence for or against the arguments being presented in the text (Govier, 1985 cited in Phillips 1992).

McPeck (1981) states that for comprehension to occur it is necessary for the person to not only decode the information being presented (ie recognise the words), but also to draw inferences from that material. Critical thinking also specifically requires the application of the same skill - that of drawing inferences.

People's Inferential Strategies

Collins, Brown and Larkin (1980) proposed the theory that people use a progressive-refinement strategy to enable them to understand what is written in the text. This process enables readers to refine their interpretations of the text until they reach an explanation that is in accord with the facts contained in the passage. Collins et al., (1980) identified eight inferential strategies that people use in this process. These problem solving techniques are: rebinding, questioning a default interpretation and any direct or indirect conflicts, shifting focus, case analysis and most likely case assignment. Rebinding is used when the person suggests an interpretation of the text and then realises that it is contrary to previous information. They then generate another interpretation, which is more suitable.

Phillips (1988) investigated the inferential strategies used by young readers to determine if they applied the same principles of critical thinking as adults did. Her sample consisted of 40 children with high ability in reading and 40 youngsters with low ability. Phillips (1988) found that the good readers utilised productive

strategies that were in accordance with the rules of critical thinking but the poor readers did not. Four of the ten strategies utilised by the children were very similar to seven of the adult strategies identified by Collins et al., (1980). The productive strategies employed by the better readers included 'rebinding', shifting focus; assessing alternative explanations; confirming previous interpretations; challenging default interpretations, and 'empathising with the experiences of others'.

The poor readers, on the other hand, used strategies that were counter-productive and which therefore, did not enable them to interpret the text correctly. These strategies included 'assuming a default interpretation' in which the reader, after incorrectly interpreting the story, misinterprets new (conflicting) information to fit in with the incorrect version of the story. A second counter-productive strategy is 'withholding or reiterating information'. In this instance, the reader does not respond to questions about the text or they merely repeat or paraphrase what they have already said.

Humane Aspects of Critical Thinking

Martin (1994) cautions that when distancing oneself from a situation in order to analyse it critically, one must not lose compassion and become uncaring. She contends that otherwise the critical thinker will end up becoming a "spectator citizen". Ennis (1987) includes "sensitivity to others" as the fourteenth disposition which critical thinkers ought to possess. He classifies this attitude as important but not vital.

The Generalisability Debate

A great deal of debate in the critical thinking arena has centred on the generalisability of critical thinking skills. Two points of view have emerged with the generalists on one hand and those who argue for the subject specificity of critical thinking skills on the other.

Subject specifists assert that what constitutes a good reason in one field may not be adequate in another. They point to the fact that mathematics and art, for example, require different types of proof to substantiate conclusions or claims (Siegel, 1994). In addition, McPeck (1981) states that because the skills required to think critically in one field are tied up with the knowledge structures in that arena, it cannot be assumed that the skills themselves are transferable. Thus, the skills necessary to critically assess the validity of arguments in one field may be completely different from those required in another arena.

The generalisability advocates believe that the skills of critical thinking can be applied across different fields. They posit that, typically, critical thinking requires a person to assess the validity, adequacy, and strength of reasons and that these criteria and principles of reason are applicable across disciplines as well as in everyday affairs (Siegel, 1994).

CREATIVE THINKING

Ennis and Norris (1990) present their thesis that there are overlaps between critical and creative thinking and that each are a subset of good thinking. They state that creative thinking comes into play, for example, when critical thinkers brainstorm alternatives to an already existing hypothesis. This view is similar to that of Shank (1988) who regards one of the underlying processes of critical thinking to be that of searching for alternative explanations to problems. Weisberg (1988) also argues for the role of creative thinking, specifically insight, in problem solving. Critical thinking, however, is required to assess the viability of these alternative explanations and hypotheses. Creative thinking, unlike critical thinking, is non-evaluative and therefore, of itself, cannot provide suitable answers to as to which decision to adopt when problem solving (Norris and Ennis, 1990).

The Current State of Critical Thinking in Schools Today

Sternberg and Baron (1987) comment that in the United States students are receiving very low scores on tests dealing with higher order reasoning. This is prompting great concern amongst educators as well as at governmental level. It appears that students can understand the literal meaning of material presented to them but "their performance drops substantially when they are asked to infer, integrate, and evaluate" (px). Glaser (1984) likewise points out that although students are acquiring greater mastery over the more elementary aspects of their education, they lack higher order skills associated with reasoning, logic, critical thinking and inference. He attributes this situation to the pervasive influence of the mechanistic and drill oriented learning theories of the past.

Glaser (1984) gives a resume of the history of learning theories from Thorndike to the present time. Thorndike believed that what was learned in one subject was not really transferable to another and consequently proposed a rote, mechanistic style of learning. Because of his focus on drill, Thorndike's critics felt that children were not developing the use of higher order thinking skills.

The stance of John Dewey in 1896, on the other hand, was more oriented to problem solving and stressed the importance of the mental process whereby a person reaches solutions. His approach was not empirical, but philosophical. He espoused the view that the solution to a problem was found when the person thinks it has been found (Ennis, 1962 cited in McPeck 1981). Another opponent of the Thorndike approach was William Brownell. Brownell believed that to be successful at subjects like arithmetic, it was necessary for children to understand the underlying principles and concepts, and be adept at manipulating these. His 1928 and 1935 studies indicated that Thorndike's method of instruction did not foster these types of skills but instead only made children better at "immature and cumbersome procedures" (Glaser, 1984, p93). George Katona, in 1940, proposed

that in order for meaningful learning to occur, individuals need to cognitively organise information. This enables new data to be fitted into the existing structure and facilitates the later retrieval of this information. He suggested that rote memorisation is only of merit when there are no organising principles underlying the material to be learned.

More recently, programmes have been implemented which focus on the process of learning and the acquisition of reasoning skills. Some of these approaches teach these skills in a general manner or in the context of the subjects covered in the curriculum.

Why the Ability to Think Critically is so Important

Within the context of the workplace, critical thinking is important in terms of worker effectiveness and the effects this has on organisational performance. An individual who can critically evaluate information and draw the appropriate inferences will be able to make effective decisions, which will ultimately impact on organisational performance. In addition, such people will be able to communicate clearly and effectively which with other employees which, in collaborative work situations, will lead to enhanced productivity.

In general, critical thinking skills have become of paramount importance today in this "Information Age" as people must process increasing amounts of complex information in relatively shorter amounts of time. As a result of this, these cognitive skills are likely to be of increasing salience as we move forward into the 21st century.

CHAPTER 3
DECISION - MAKING

Decisions are the outcomes of critical thinking and inferences. One of the key components of a manager's job is decision-making. The decisions that a manager makes affect both the organisation in terms of productivity, and the staff who work there. It is therefore of paramount importance that managers make decisions that are competent. A great deal of research has been conducted into the decision-making process and the factors which influence decisions. In particular, researchers have attempted to describe how individuals make decisions and what constitutes a good decision.

Decision making research has been classified in a number of different ways. Abelson & Levi (1985, cited in Maule & Svenson 1993) identified a number of approaches that have been adopted to study decision-making. These include process approaches to decision-making which look at the cognitive aspects involved in the process of reaching decisions. Other approaches evaluate decisions using normative, prescriptive and descriptive models of decision-making behaviour. Descriptive approaches are concerned with how people make decisions in real life. Prescriptive and normative approaches point to how people should make decisions. Prescriptive approaches pay attention to the way the people assess and combine information whereas normative approaches ignore this (Maule & Svenson, 1993). Although not a complete categorisation, these approaches have provided a framework within which researchers have evaluated how decisions are made.

HOW DO PEOPLE MAKE DECISIONS?

However, the question still remains, how do people make decisions? Janis and Mann (1977) reviewed some of the different approaches to decision-making. These include the decision optimising approach, the saticficing approach and the elimination by aspects method. In the optimising approach, decision-makers are presumed to take the alternative that offers the largest benefits relevant to costs. To optimise or maximise their decision choice, individuals need to consider and weigh up every alternative course of action before making a decision. However, in the workplace such a strategy may not be viable due to time constraints and the amount of information that would have to be evaluated.

Another method is that of "satisficing". This approach specifies that a decision or course of action is chosen because it meets a minimum standard of conditions. Decision-makers typically take more of a superficial approach when assessing the alternatives open to them and usually make a decision that "will do" (Janis & Mann, 1977, p 26). However, the satisficing approach to decision-making is somewhat haphazard and different courses of action are only reviewed once before being discarded.

In the "elimination by aspects" approach the decision-maker sifts through a number of alternatives according to whether they contain a salient aspect, and discards those that don't. This process is repeated with the next important aspect until they have all been considered and there is only one alternative left. However, the person may run out of alternatives before all of the important aspects are considered or, conversely, they may run out of aspects before the alternatives have been assessed.

However, Mitchell and Beach (1990) note the growing dissatisfaction amongst researchers with behavioural decision theory because the way people make decisions in organisational contexts often does not have much bearing on the

expectancy models that have been formulated by theorists. Because many managers rarely rely on formal models of decision-making, but instead use intuitive decision-making strategies, research has also looked at naturalistic or intuitive models of decision-making. Unfortunately, the decisions that people make intuitively may be prone to a wide array of biases and errors (Tversky & Kahneman, 1982).

Bias and Error in Decision-Making

A number of factors can interfere with an individual's information processing capabilities, which in turn can threaten the quality of decisions made. These factors include stress (Shanteau & Dino, 1993; Janis & Mann, 1994), time pressure (Maule & Hockey, 1993), complexity and amount of information, task difficulty, distractions and emotional arousal (Bodenhausen, 1990 cited in Kaplan, Wanshula & Zanna 1993).

In such situations people tend to resort to simple strategies to help them to reach decisions. These strategies include using general knowledge, stereotypes, or simple rules of thumb (known as heuristics). However, problems can arise as a result of utilising these simplified stratagems because they do not allow for a careful consideration of the salient features or details of the information or situation (Kaplan ,Wanshula & Zanna, 1993). Thus, these cognitive shortcuts can lead to errors and incorrect conclusions and ultimately faulty decisions.

Jagacinski (1991) comments that often people must reach decisions using information that has key parts missing. Brodt (1990) outlines some of the heuristics or inferential strategies that people may rely on when reaching decisions in situations where there is incomplete information or other conditions of uncertainty. 'Availability' is one of these heuristics. Tversky & Kahneman (1982) state that people employ this rule of thumb "to assess the frequency of a class or the probability of an event by the ease with which instances or occurrences can be

brought to mind" (p11). Thus, the decision-maker uses the ease of retrieval to decide if the event is frequent or probable, rather than the actual number of instances of that event (Taylor, 1982). However, such a strategy may result in systematic bias because events that are more easily recalled may be viewed as occurring more frequently than those which are more difficult to remember, but occur as frequently (Tversky & Kahneman, 1982). Wagner (1991) comments that, managers, due to the large amount of information that they often need to assimilate to solve problems and to arrive at decisions, may fall prey to the errors and distortions that can result from the availability heuristic.

The person who uses the availability heuristic may also be open to error as a result of the salience bias (Brodt, 1990) and the vividness effect (Nisbett & Ross, 1980). The salience bias may occur when people or events stand out from the ordinary. These people and events have more of an impact and are therefore more easily recoverable from memory and this in turn could affect the inferences that are drawn about this information (Brodt, 1990). The vividness effect refers to information that is interesting emotionally, captures the attention and excites the imagination. This characteristic of the information makes it more easily remembered and this could influence the availability of the information for making inferences (Nisbett & Ross, 1980).

Although in some situations these heuristics, or inferential strategies, may be helpful, they very often lead to errors in judgement (Brodt, 1990). Tversky and Kahneman (1982) note that these biases are also evident amongst experienced researchers (not just lay people) when they make decisions intuitively.

WHAT MAKES A DECISION "GOOD"?

Using an intuitive model, Zakay (1984) investigated the criteria that middle managers use to assess the 'goodness' of other managers' decisions. The 145

managers in this study, who were from 3 electronic technology industries, were asked to complete a 25-item questionnaire. The most important criterion listed by the managers for assessing the quality of decisions was "goodness of outcomes". In descending order of salience, the other factors were "correctness of the decision process", "information utilisation", "realism" and resources", "ethics", "subjective rationality," "acceptance" and lastly "feelings and social compromise".

Other models have been used to identify the features that make a decision "good". Some researchers have assessed the quality of a decision in terms of how much it diverges from a normative model. However, such a divergence may not in actual fact be an error. Winkler and Murphy (1973) state that real life decisions cannot be assessed according to only one normative model as the model may not accurately or adequately reflect the situation.

Another approach defines optimal decisions as those that meet a certain criterion level (eg profit) while taking into consideration environmental variables and timeframes (Einhorn & Hogarth 1981). A different method for evaluating decisions is according to their outcomes. However, poor results may not necessarily be attributable to poor decisions (Zakay, 1984).

Improving Decision-Making

To improve the quality of inferences that people make, Nisbett, Kranz, Jepson and Fong (1982) suggest the utilisation of 'inferential maxims' or slogans such as "it's an empirical question" to guide the way individuals approach information. Other researchers (Gadzella, Hartsoe & Harper, 1989), have found that instruction about critical thinking, can improve this skill, particularly the ability to draw inferences, for people with average and above average intelligence levels.

PART 2
PSYCHOMETRIC PROPERTIES OF TESTS

The MRA test assesses an individual's decision-making ability. In particular, it focuses on the ability to draw inferences because this skill is considered central to the decision-making process and the ability to think critically. In order to assess the ability of the MRA to function as a selection test, a number of psychometric factors had to be considered in its construction and development. These factors are the reliability and validity of the instrument as well as the criterion used to assess the test's validity. These factors are discussed in the following three chapters.

CHAPTER 4
RELIABILITY

The reliability of an instrument refers to the consistency with which it measures the construct it was designed to measure (Kline, 1989) and its "relative freedom from unsystematic errors" (Aiken, 1988, p95). Unsystematic errors, such as differing test conditions and the test takers state of health or mind, are unpredictable and can lower a test's reliability. Classical test theory posits that an individual's obtained score is likely to contain some degree of error, that is, the amount that the obtained score varies from their true (hypothetical) score (Aiken, 1988). The actual score that an individual achieves on a test is, therefore, an approximation of what their true

score would be (Kidder, 1982). The smaller this component of error is, the more reliable the instrument.

The reliability of a test can be assessed in several ways: test-retest reliability, split-half reliability and equivalent forms. Each of these methods yields a coefficient of reliability which indicates how reliable the measuring device is (Landy, 1989).

Test-retest Reliability

Test-retest reliability refers to the test's reliability over time. In this type of reliability the test is administered to a group of people and, after a suitable lapse of time, administered to the same collection of people again. The interval of time between each test administration cannot be so long that major life events have occurred within the individual's life which could change them, and thus alter or affect their performance on the test (Kidder, 1982). Nor can it be of such short duration that the person can remember the answers to the test, as this too could affect their responses (Berry & Houston, 1993) and spuriously raise the correlation coefficient (Kline, 1986). Test-retest reliability is concerned with the relative standing of each person's scores on each administration of the test (Kidder, 1982). A coefficient of stability is obtained from this type of reliability estimation (Landy, 1989). This type of reliability does not give any indication of the internal consistency of the test.

Split Half Reliability

This type of reliability is related to the internal consistency of the test rather than the stability of the test scores over time (Anastasi, 1986). Kline (1989) states that internal consistency reliability is of paramount importance. The reason being that if a test is assessing a particular construct, then all parts of the test should be consistent with each other (i.e. all assessing aspects of that construct). If the parts

are not consistent, they cannot be measuring the same variable. And if the test is not consistent it cannot be valid.

With split half reliability the test is given to candidates only once. Afterwards, the test is split into two equivalent parts and scores are calculated for each half of the test. These two scores are then correlated with each other, giving a reliability coefficient. Because this coefficient is only for a test half the length of the original test, the Spearman Rank Prophecy formula is used to give an estimation of the reliability of the entire test (Aiken, 1988).

Anastasi (1990) advises that when splitting the test, it is important to ensure that the two parts are equivalent. This is often best achieved by dividing the test into odd and even numbered questions rather than just partitioning it down the middle. This helps to ensure a good mix of easy and difficult items, as well as minimising any of the other variables that might come into play such as practice effects, boredom and tiredness, when a test is attempted from start to finish. In addition, when items in the test relate to one particular piece of information such as a section of prose, these items should remain together in one half of the test. This helps to avoid falsely inflating the correlation coefficient, as items in both halves would be affected by even a single comprehension error.

Equivalent or Alternate Forms Reliability

In the alternate forms type of validity two versions of the same test are constructed. The two tests have to be carefully matched to ensure that they both contain items that are comparable in terms of difficulty and content (Landy, 1989). Both of these tests are then administered at different times to candidates and the scores from each are correlated. A coefficient of equivalence is realised from this method of assessing reliability. The problems associated with this kind of reliability are the

difficulty, cost and time involved in developing 2 equivalent forms of the same test (Berry & Houston, 1993).

CHAPTER 5
VALIDITY

Validity refers to "extent to which the test measures what it has been designed to test" (Aiken, 1988, p103). As Aiken notes, different validities can be attributed to a test depending on the criterion used, the type of validity investigated and the way in which the test will be used. There are several different methods that can be utilised to assess a test's validity. Depending on which design is used, information is obtained concerning the face validity, content validity, criterion-referenced validity and construct validity of a test. Wainer & Baum (1988) state "test validation is a process of accumulating evidence to support the inferences made from test scores"(p21).

Face Validity

Face validity pertains to the extent to which a test appears to measure what it is supposed to measure and is essentially a subjective process (Kidder, 1981). This form of validity is important for test takers in terms of the test's acceptability to them (Berry & Houston, 1993) and also for marketing purposes (Aiken, 1988).

Content Validity

Content validity assesses the extent to which the test items adequately sample the domain of skill or knowledge areas that the test is intended to measure. Again, this type of validity involves the judgement of subject matter experts who decide if the test performs in this respect (Aiken, 1988).

Criterion-Related Validity

Criterion-related validity is empirically derived and involves the correlation of test scores with some measure of job performance, called the criterion. There are two types of criterion-related validity: concurrent and predictive (Aiken, 1988). In concurrent validity, the criterion is collected at the same time as the predictor (test) scores and the sample comprises current employees. In predictive validity, the criterion is collected at some stage after the test scores. The sample usually consists of job applicants (Berry & Houston, 1993). The predictive validity of an instrument is investigated to enable predictions to be made about test-takers on the basis of their scores on the test.

Construct validity provides an indication of the degree to which the test measures the theoretical construct or trait it is meant to be assessing. Construct validity comprises both convergent and divergent validity. Convergent validity involves correlating test scores on the new test with scores obtained on an already existing test, which has been validated and measures a similar construct. If the new test measures a similar construct to the pre-existing test, then the correlation should be high. Divergent validity refers to the extent that the new test does not agree or correlate with a previously validated test which measures a dissimilar construct (Berry & Houston, 1993).

To justify the construction and development of a new test when there is already a validated test that measures the same or a similar construct, it has been suggested that the new test be either shorter or easier to administer (Kline, 1989), more thoroughly measure the construct in question (Berry & Houston, 1993) or used in place of tests that applicants were already familiar with (Turban, Saunders, Francis, & Osborn, 1989).

Issues Surrounding the use of Concurrent versus Predictive Validity Designs

Since the primary purpose of selection tests is to assist employers in choosing the 'right' employee, the question that arises is whether the concurrent validity design is "as good as" the predictive validity design and if the two can be used interchangeably. A great deal of the discussion which has surrounded the use of either predictive or concurrent validity designs, has focussed on the superiority of the predictive method of estimating validity and the comparative shortcomings of the concurrent validity studies (Barrett, Phillips & Alexander, 1981). Barrett and his colleagues list the four central criticisms levelled against concurrent validity as being "missing persons", restriction of range, motivational and demographic differences between present employees and job applicants, and confounding by job experience" (p1). They cite the official APA stance that the predictive validity design is more scientific than the concurrent validity design and therefore to be preferred.

However, Barrett et al., (1981) then go onto present their case that the conceptual differences between concurrent and predictive validity are not as great as had previously been thought, and that some of the aforementioned drawbacks apply to both the concurrent and predictive validity designs, although not in equal degrees. In addition, they believe that the differences that do exist do not have an appreciable effect on the accuracy or size of the resulting validity coefficient.

Guion and Cranny (1982) in their reply to Barrett at al., (1981), argue that concurrent and predictive validity cannot be substituted one for the other due to their conceptual and practical differences. They distinguish between five subcategories of predictive validity to elaborate their point concerning the differences between concurrent and predictive validity studies.

In concurrent validity designs the sample consists of present employees. This sample is not considered to be representative of the pool of applicants, which form the sample in predictive validity studies, as it does not include those who were not hired and those who have since been promoted. This constitutes the "missing persons" problem. The range of scores obtained on both predictor and criterion measures will therefore be restricted in range as they will not include scores from both the very effective and the less effective individuals. This in turn will influence the obtained validity coefficient (Barrett et al., 1981). Guion and Cranny (1982) state that due to this lack of data at the extreme edges, it is not possible to make "informed estimates of the population parameters" (p239). They also note that it will increase the risk of Type II error occurring. However, Guion and Cranny also admit that this problem is of concern in two of the five kinds of predictive validity that they identified.

Barrett and his colleagues contend that the sample obtained in the predictive validity design is restricted in range anyway as a result of the organisation's hiring procedures. Therefore, like the concurrent validity sample, it too cannot be representative of the potential applicant population. The issue then becomes one of deciding if the range restriction prevalent in concurrent validity studies has a significant effect on the resulting validity coefficients.

Lent, Aurbach, & Levin (1971) conducted a review to establish which form of validity (concurrent or predictive) yielded the highest frequency of significant results. Their study revealed that the predictive design did so. However, they did not attribute this to the (supposed) superiority of the predictive validity design but rather to more careful selection of criteria and predictors. In addition, they speculated that some concurrent validity designs might have been used as pilot studies for predictive studies, and wondered whether the results might also be a reflection of researcher experience. Bemis (1968, cited in Barrett et al. 1981) in

their review, on the other hand, found that the validity coefficients obtained for both predictive and concurrent validity studies were similar.

Barrett et al., (1991) note that job applicants have a higher degree of motivation to do well on the predictor (selection test) than job incumbents due to their belief that their performance on the test may influence the hiring decision. This may affect their scores. Both sets of researchers point out that this effect may plague personality tests and interest inventories, but has little or no effect on scores obtained on cognitive ability tests. Barrett and colleagues further reason that contamination by motivational differences must be minimal for personality tests and interest inventories because the obtained validity coefficients for both types of validity designs are of the same size.

Differences in job experience or training are evident in concurrent validity designs but are controlled for in predictive validity designs, due to the different composition of the two samples (present employees and job applicants). Barrett et al., (1981) argue that this may affect predictors such as work samples and interviews, but not cognitive ability tests. In any instance, they do not believe the effect of these two variables has been investigated fully enough at this stage to warrant firm conclusions to be drawn.

CHAPTER 6
THE CRITERION

In order to assess the validity of a selection test it is necessary to evaluate it against some index of job success. This index is called the criterion. Bingham (1926, cited in Austin & Villanova 1992) defined the criterion as "something which may be used as a measuring stick for gauging a workers relative success or failure." Scores on the test are then correlated with the scores obtained on the job performance index to provide a measure of test validity (Murphy & Schiarella, 1997). Difficulties arise, however, in deciding what actually constitutes an appropriate or sufficient measure of job performance.

THE ULTIMATE CRITERION

The ultimate criterion is a theoretical construct, which can be viewed as the perfect or ideal criterion for measuring true success (Blum & Naylor, 1968). Since the ultimate criterion is a theoretical ideal only and therefore extremely difficult to obtain (James & Ellison, 1973), the 'actual' criterion (which is an approximation of the ultimate criterion) is used instead. The degree to which the actual criterion overlaps with the ultimate criterion is called criterion relevance. Criterion deficiency relates to that portion of the actual criterion that does not overlap with the ultimate criterion, and criterion contamination (which consists of error and bias) represents variance in the actual criterion which is not present in the ultimate criterion. Criterion deficiency and contamination distort the actual criterion (Blum & Naylor, 1968).

Many different criteria have been used to measure job performance. The most commonly used indicator is some kind of subjective rating (Viswesvarau, Ones, & Schmidt, 1996). Supervisor's ratings are employed the most frequently (Landy & Shankster, 1994) with peer ratings being the next most often used (Cascio, 1991). Other criteria have included salary and promotions (Judge, Cable, Bougreau & Bretz, 1995); university grades (Roth, BeVier, Switzer, Schippamnn, 1996); the amount of education a person has completed, and scores obtained on other already existing tests (Anastasi, 1986).

UNIVERSITY GRADES

Notwithstanding the fact that university grades vary from one university to another (Reilly & Warech, 1993) and are influenced by the marking strategy employed by the examiner and the subject being assessed, they have been found to be reasonably valid predictors of future job performance. Roth et al., (1996) conducted a meta-analysis of the relationship between university grades and job performance. After corrections for statistical artifacts had been made they found a fairly sizeable correlation in the 0.30's. Interestingly, they also discovered that grades obtained for undergraduate and masters degrees were more valid than those acquired for doctorates or in medical school. Roth and colleagues concluded therefore, that grades are better predictors of job performance than had previously been thought.

THE AMOUNT OF EDUCATION COMPLETED

Another criterion utilised in the validation of selection devices is the amount of education that a person has completed. This variable is often used for "out of school adults" (Anastasi, 1986, p148) and is related to the criterion of grades. The rationale behind the use of the highest educational level achieved is that those of higher intelligence pursue their education the furthest while the less capable students fall away earlier. However, it is to be noted that sometimes other factors

such as motivation, financial considerations and social concerns may also feature in decisions to continue or abandon higher education (Anastasi, 1986).

Judge et al., (1995) explored the validity of various predictors of executive career success. Their investigation revealed that the quantity of education an individual had completed played a significant role in the amount of salary they earned. The researchers especially noted the difference in accumulated earnings between those who obtained postgraduate degrees and executives with undergraduate degrees. Over a 20-year span they calculated that the higher qualified people earned an average of $320,000 more than their lower educated counterparts, indicating that the amount of education a person completes is a factor in later job success.

SALARY

Salary has been used as an index of job performance with some researchers describing it as an external and objective measure of job success (Judge et al, 1995). It is believed to be an indication of the value that a particular person has to an organisation. For example, subordinates are more likely to be awarded higher salary increments when their immediate supervisor is reliant upon them for their expertise or superior ability / knowledge in a particular area (Bartol & Martin, 1990).

There are, however, some non-performance-related factors which may influence pay rates. Geographical location can be a source of variance in salary figures (Judge et al., 1995) as can gender (Treiman & Hartmann, 1981). In addition, salary ranges may differ across industries making inter-organisational comparisons difficult. Sometimes, too, pay allocation may be motivated by political considerations (Bartol & Martin, 1990).

However, as salary level does increase the further up the organisation's hierarchy one goes and, in principle, those who are more skilled and able occupy higher positions (although aberrations to this can occur), there may be justification in using salary as an index of job performance. It's use as a criterion measure is perhaps also appropriate these days as performance based pay is more of a prominent feature of the workplace.

ALREADY EXISTING TESTS

Already existing tests that measure a similar or related construct are also often employed as a criterion for assessing a new test's validity (Anastasi, 1986). The benefit of using pre-existing tests is that they have already been validated and information regarding their psychometric properties is easily accessible. Anastasi (1986) states that an already existing test is a suitable criterion only when the new test is a shorter or simpler version of it.

PERFORMANCE RATINGS

Performance ratings are "subjective evaluations that can be obtained from supervisor's, peers, subordinates, self or customers" (Viswesvarau et al., 1996, p557). Because ratings are based on human judgements, and are therefore subjective in nature, they can be prone to error and bias. This limits their usefulness as indices of work performance. However, ratings (particularly those from supervisors) are still widely used in the validation of tests (Wedge & Kavanagh, 1988). Saal, Downey & Lahey (1980) suggest that performance ratings are used mostly because of the lack of other objective measures of performance. Other factors influencing the use of ratings may be related to their availability (many companies have some kind of performance assessment on file) and the (relative) ease with which they can be acquired. The unreliability of ratings could account for some of the rather low validity coefficients that are reported for some selection tests, indicating criterion deficiency rather than test invalidity.

Supervisor's Ratings

It has been found that of all sources of rating, supervisor's ratings are the most widely accepted by subordinates. This is probably because evaluation is viewed as one of the functions of a supervisor. Berry and Houston (1993) point out that the immediate supervisor probably has a fairly good idea of the duties and tasks of their subordinate's job and presumably can therefore rate the person more accurately. They also note, however, that there are some employees who have a high degree of autonomy, such as managers, technical staff and managers, and consequently limited close contact with their immediate supervisor. In these instances, the supervisor may not have direct knowledge of their subordinate's work performance and may therefore be unable to rate them with accuracy

Interpersonal affect can also influence supervisor's ratings. Varma, DeNisi and Peters (1996), in a study of 112 first line supervisors, found that affect played a part in all performance appraisals with the effect being more noticeable for ratings pertaining to traits rather than those relating to task performance.

With supervisor's ratings, it has been found that interrater reliability (ie the extent of agreement between raters) is lower than intrarater reliability but that interrater reliability is higher for supervisors' ratings than for those given by peers (Rothstein, 1990; Viswesvarau et al.,1996).

Peer evaluations

These have been proposed as a source of performance ratings on the basis that peers have the opportunity to observe a person's work behaviour more closely than supervisors. Because they work alongside their co-worker, they are likely to have a fairly good understanding of what their work entails. Amir, Kovavsky and Sahran (1970) found peer ratings to be valid and reliable for promotions, although this was

in a military setting. Other researchers, however, have found that these types of ratings can introduce bias, possibly due to competitiveness on the part of the peer. Perceived personal attributes of the ratee also influence the rating assigned to them Borman, White & Dorsey (1995)

DeNisi, Randolph & Blencoe (1983) studied the after-effects of negative and positive peer ratings. They postulated that negative peer ratings could lower cohesiveness amongst workers, have a detrimental effect on subsequent performance, and lead to reprisals from co-workers who might give lower ratings in return. Their study confirmed these hypotheses. The effects for positive ratings were small and not significant. The researchers did mention, however, that their study was limited by the short timeframe that it was conducted in, and speculated that some of the effects they observed might have been ameliorated with time.

Other researchers (Fox, Ben-Nahum, & Yinon 1989) have found that the accuracy of peer ratings was very low when the rater and ratee were dissimilar. Ratings were more accurate when the person being rated was similar to the rater.

Subordinate's ratings

This type of rating is of limited usefulness for several reasons. Subordinates typically only see a restricted proportion of their supervisor's work - they are unlikely to observe their superior's interactions with those from other levels of the organisational hierarchy, with other professionals, or with clients – and therefore cannot reliably rate these particular functions. Bernardin (1980, cited in DeNisi et al. 1983) found that subordinates who had received negative ratings from supervisors may give their superior poor ratings on dimensions relating to leadership style. In his study, Antonioni (1994) gave subordinates the option of rating their supervisor anonymously or with identification. His results show that when subordinates gave inflated ratings to their supervisors when their identity as

the rater was known. On the positive side, Smither, London, Vasilopoulos, Reilly, Millsap and Salvemini (1995) reported that some managers' performance improved over a 6-month period after receiving upward feedback on their work.

Self ratings

Research has found that when individuals rate themselves, the ratings tend to be inaccurate. Mabe and West (1982) in their review and meta-analysis of 55 research papers investigating self evaluations reported a low mean validity of .29 (corrected for statistical artifacts) for this type of appraisal. They also concluded that raters who had a high intelligence level, an internal locus of control, or previous experience with self-rating were more accurate than those who did not. Self-evaluations tend to exhibit leniency errors possibly due to the inability of raters to view themselves objectively or because of a desire to enhance oneself (Mabe & West, 1982). DeNisi and Shaw (1977) in a study of 114 students in an introductory psychology class, found that self-assessments of ability were inaccurate and could not be used in place of ability tests.

Rater Errors and Biases

Landy (1989) distinguishes between 3 classes of errors which may influence performance ratings: leniency-severity errors, central tendency errors and halo errors. Leniency-severity errors occur when the rater is either too harsh or too lenient in their gradings relative to the actual performance of the employee. This tendency may be a reflection of the personal expectations of the rater (Berry & Houston, 1993) or different conceptions of what constitutes exceptional, good or poor behaviour (Landy, 1989).

Central tendency errors occur when raters are habitually loath to use the extreme ends of a scale. Consequently, their ratings are mostly clustered around the middle of the scale. Test validation becomes a problem when this type of error is

encountered in performance evaluations because of the resulting restriction of range of the ratings. This in turn may shrink the size of the validity coefficient (Landy, 1989).

The third type of rater error is halo error. Cooper (1981) states that halo error may occur as a result of ill-defined or very descriptive rating dimensions and cognitive errors on the part of the rater, for example, inadvertently omitting details but including their own views of the rating dimension. Other factors are insufficient knowledge of, or access to, the ratee's performance and 'engulfing' – when an overwhelming characteristic of the ratee influences ratings along the other dimensions. True halo occurs when the rating categories are correlated.

The effects of implicit personality theory on performance ratings have also been investigated. Landy & Farr (1980) define implicit personality theory as "the assumed values on performance dimensions that are independent of the actual behaviour of the ratee on those dimensions". Krzystofiak, Cardy and Newman in their 1988 study discovered that implicit personality theory did have an effect on appraisals, and that this effect was still in evidence even when the ratee vignettes were composed of purely behavioural ratee descriptions.

In their review of performance ratings, Landy and Farr (1980) concluded that rater gender does not affect ratings.

Rater errors and biases, however, can be minimised through the use of appropriate and clearly defined rating scales and by rater training (Landy & Shankster, 1994) although the latter appears to depend on the type of training (Wedge & Kavanagh, 1988). Wedge and Kavanagh tested the effect of different types of training on rater errors (halo, leniency and restriction of range) and rater accuracy. They found that while traditional rater training reduced the amount of errors it also lessened the

accuracy of the ratings. Other types of training (observational and decision-making) increased the accuracy of ratings but had little effect in curtailing rater errors. Rater experience, however, does improve ratings although it is not understood precisely what underlying factors contribute to this (Landy & Farr, 1980).

Murphy and Schiarella (1997) argue in favour of employing several criteria, which cover different aspects of job performance, when assessing test validity. They suggest that this is a "better and more realistic" way of evaluating validity due to the multifarious nature of job performance. They also note that employers frequently utilise several selection devices during the hiring process indicating that this too is a "multivariate process".

MULTIPLE VERSUS COMPOSITE CRITERIA

When several criteria are used to assess a construct such as job performance the question arises as to how the resulting information will be treated. One option is to combine the criteria into one single global measure of job success. This single index is called the composite criterion (Blum and Naylor, 1968). There are several ways of combining the data and each involve assigning different weights to the individual subcriteria according to their relative importance, using either evaluative or statistical procedures. For instance, expert judgements may be used to do this or the subcriteria may be weighted according to their respective reliabilities. Alternatively, equal weights may be allocated to each subcriteria.

The problems associated with the composite approach concern how to decide what weight to assign to which criteria, the availability of the experts and their familiarity with the job in question. Even with the statistical procedures, issues of relevancy must be taken into account. In addition, opponents to the composite criteria point

out that although job success is multidimensional, the separate dimensions may not be additive (Blum & Naylor, 1968).

Consequently, another approach has been advocated instead. This involves considering each of the multiple criteria separately with no attempt to combine them. Landy (1989) suggests that there are times when the composite criteria is an appropriate measure to use and other times when the multiple criterion approach is more suitable.

Brogden (1946, cited in Viswesvarau et al., 1996) states that in the end the most important requirements for criteria are that they are reliable, relevant and practical. Viswesvarau and colleagues note that reliability is the one characteristic that is uniformly demanded by all researchers.

AIMS

This research is an extension of a project begun last year in which the MRA test was designed, constructed and then subjected to an item analysis using a sample of intermediate to senior level employees. The present study will look at the further development of the psychometric properties of the test, its and its potential usefulness as an employment selection tool. Specifically, this research has 2 components:

1. **The investigation of the criterion-related validity (concurrent design) of the MRA.** The validity of the instrument will be explored using two different criteria: salary and highest educational level achieved.
2. **The assessment of the MRA's reliability**. The reliability of the tool will be calculated using the alpha coefficient, which is a measure of internal consistency.

METHODOLOGY

SAMPLE

The overall sample comprised 96 voluntary participants recruited from around New Zealand. This sample consisted of two distinct groups of people. The first group comprised 86 employees, the majority of whom were managers. The second sub-sample was composed of 10 university students from both undergraduate and postgraduate levels. The results from these 2 subgroups were analysed separately due to the marked differences in the composition of the two samples. Group 1 participants were recruited from their respective companies and organisations, which were randomly selected from the telephone book. Respondents came from a diverse range of industries, as well as from large corporations and small businesses. The occupational level of participants ranged from the junior to senior echelons of the organisational hierarchy. Group 2 subjects came from several tertiary institutes around New Zealand.

MEASURES

Managerial Reading Assessment (MRA)

This is an original cognitive ability test, which assesses the ability to draw inferences. The present researcher designed and constructed the test, and then conducted an item analysis on it in 1997, using a small sample (40 subjects) of intermediate to senior level employees. Originally, the MRA comprised 55 questions with a total of 384 inferences. The test was reduced to its present length (8 questions with a total of 45 inferences) as a result of the item analysis.

Each of the 8 questions describes a workplace scenario or situation followed by a number of inferences that could be drawn from the preceding information. Candidates are required to assess the degree of truth or falsity of each of the inferences, based on the information presented. The candidates select an answer from one of the following options: 'true', 'probably true', 'false', 'probably false', and 'insufficient data'. For the categories of 'probably true' and 'probably false', respondents can incorporate their general knowledge in determining their answer.

All of the scenarios described in the test are fictitious although some draw on information contained in newspapers, magazines, business weeklies or research articles. References were included in the original long version of the test, but these were subsequently removed.

MRA is a paper and pencil test, which takes approximately ½ hour to complete. This length was chosen in a bid to increase response rate as people are more likely to complete shorter rather than longer questionnaires (Roszkowski & Bean, 1990).

Managerial Rating Scale

Supervisor's ratings of participants' on-the-job work performance were one of the 3 criteria employed to explore the validity of the MRA. To enable meaningful comparisons to be made between the ratings from supervisors in different organisations (and fields), the present researcher created a Managerial Rating Scale (See Appendix A). The purpose of utilising this scale was to ensure that participants were assessed along the same dimensions, rather than trying to compare evaluations from the different rating scales used by the various participating organisations.

The Managerial Rating Scale assesses the major managerial core competencies, which the literature has indicated as being essential to the successful performance of managers (e.g. Rippin, 1995). This scale lists 10 competencies and asks supervisors

to rate the employee along each of these using a 7 point Likert Scale. The Likert scale ranges from outstanding to incompetent. The supervisor is then asked to provide a global rating of their subordinate's performance, again using the Likert Scale.

Statistical Data Sheet

To analyse the statistical properties of the MRA test, participants were asked to complete an anonymous statistical data sheet . To assess the validity of the MRA, respondents were asked to supply information about the following: market sector, occupational status, highest educational level achieved, salary range and age. Due to the sensitive nature of information regarding salary, participants were asked to tick a salary range rather than provide an actual figure. This was done in a bid to increase response rate and to provide participants with extra security regarding confidentiality.

PROCEDURE

To recruit volunteers to participate in this research an initial telephone call was made to either the Manager or the Human Resources Manager within each organisation. During this conversation the research and its objectives were briefly outlined and, if this met with an expression of interest, an information letter (see Appendix B) was posted or e-mailed to the person. Sometimes, the contact people agreed to participate themselves or recommended others who could be approached. When individuals assented to take part in the study, a covering letter, the test, the statistical data sheet and the information sheet were posted or e-mailed to them. Participants were also provided with a Freepost envelope to return the completed forms. Respondents were given the option of answering anonymously or identifying themselves (for feedback purposes).

Initially, subjects were given a choice of taking part in one of two validity studies. The first validity study requested permission to obtain a confidential on-the-job performance appraisal of the participants from their immediate supervisor. If participants agreed to this, they were asked to complete a consent form. However, this option had to be dropped as too few people wished to take part in it and it was felt that it was hampering the response rate, due to the sensitivity of the information requested. Instead, all participants were asked to take part in the second validity study in which salary level and 'highest educational level achieved' were utilised as the criteria for assessing the validity of the MRA test.

If participants wished to receive confidential individual feedback on their results, they were asked to provide an address where these could be posted to them. Otherwise, this research was conducted on an anonymous basis.

This research employed a survey methodology. There are a number of benefits and drawbacks associated with this method. One of the advantages of the survey methodology is that it is relatively cost efficient requiring only the extra expense of stamps and telephone calls (Landy, 1989). This provides the opportunity for a large number of people to be contacted to take part in the research. It also enables respondents from many far-flung geographical locations to take part. In addition, during the initial telephone call it is possible for a good rapport to be established with the other party which could enhance response rate (Kidder, 1982).

A postal survey also places less pressure on the participants as they can schedule time to complete the test according to their commitments. In the present study, this enabled very busy managers, who may not have otherwise agreed, to participate in this research. This aspect of postal surveys. potentially also allows respondents more time to think over their answers, perhaps improving the quality of their responses (Simon, 1969).

The survey methodology provides participants with the option of answering anonymously. Because information can be supplied anonymously it increases the likelihood that it will be more accurate (Miller, 1983) and this makes it more valuable when analysing the statistical properties of the test. For the present study this feature of postal surveys possibly made it more feasible to collect data regarding individual's salary levels, educational level etc. Anonymity also contributes to the independence of the research (Kidder, 1982).

Another advantage of the survey methodology is that it completely excludes the effects of interviewer bias (Kidder, 1982). Surveys also cut down on administration time (Miller, 1983). However, this also means that the conditions under which different participants complete the test will vary. For example, some people may have been able to retire to a quiet room whereas others may have been in a noisy environment surrounded by distractions. Some individuals may have completed the entire test in one sitting whilst others may have done finished it after successive attempts. It is also conceivable that some may have conferred with others over some of the answers. All of these factors could have affected the quality of responses and contributed to unsystematic variance in test conditions which could decrease validity (Anastasi, 1986).

Postal surveys may also be prone to response bias as potential participants can elect not to complete the questionnaire. However, follow up letters or phone calls can help to improve response rate (Shaugnessey & Zechmeister, 1994). Another related drawback is the slow response rate as participants may take their time to return the completed questionnaire (Kidder, 1982).

Overall, for the purposes of this research, the benefits of using the survey methodology far outweighed the drawbacks.

ETHICAL ISSUES

A number of ethical considerations required to be taken into consideration with this research. These issues were: confidentiality and anonymity, informed consent, feedback, and independence of the research (from the participants' employment environment).

Confidentiality and Anonymity Issues

Participants were asked to complete a statistical data sheet which requested the following information: respondents' market sector, occupational status, highest educational level achieved, salary range and age. It was necessary to gather this data to analyse the statistical properties of the test. This was explained (in written form at the top of the sheet) to participants. Participants could decline to provide any or all of this information. Due to the sensitive nature of some of this information, issues of confidentiality and privacy had to be addressed. To protect the privacy of individuals and to provide as much anonymity as possible, participants were not asked for their name. In addition, the statistical information was stored apart from test scores and any correspondence that may have been entered into, in a secure place, until it's destruction at the completion of the research. Only the researcher and the researcher's academic supervisor had access to this data.

All of the participants were randomly assigned a confidential code number and this was entered onto their test sheets, statistical data sheets and any correspondence they may have entered into. The list which contained the names with their assigned numbers was kept separate and in a secure place to maximise confidentiality. Only the researcher and the researcher's academic advisor had access to this list. Kidder (1982) recommends this as an advisable and commendable procedure to follow.

For those who chose to respond anonymously 'anon' was simply entered on the sheet next to their delegated number.

In addition, respondents' individual test score results were kept confidential to the participant, the researcher and the researcher's academic supervisor. These scores were not disclosed to participants' employers (or anyone else).

Independence of the Research

As all of the participants were recruited from their place of employment, it was explicitly stated in the information sheet that participation in this research was completely independent of their work and had no ramifications with respect to this. Further, respondents were informed of their right to accept or decline to take part, or refuse to answer particular questions without prejudice. This was included in an effort to remove any perceived coercion to participate and to provide a climate in which the potential participant could exercise freedom of choice (Kidder, 1982).

Feedback / Evaluation

Participants were given the option of receiving confidential, individual feedback on their test results, if they so desired. To obtain this, the participant was asked to supply an address to which the results could be posted. Employers were provided with a general summary of the psychometric properties of the test.

Informed Consent

Full and informed consent was sought from respondents prior to participation in this research. This was achieved firstly by a brief explanation about the research and it's aims during the initial telephone contact. If individuals expressed an interest in taking part, a more detailed information sheet was either posted, e-mailed or hand delivered to them. In addition, participants who elected to take part in the mini-study requiring their supervisor's rating of their-on-the-job performance, signed a

consent form giving permission for this rating to be given (Kidder, 1982). All of the participants were given contact details of the researcher and the researcher's academic supervisor in case they wished to receive further information or clarification. The supervisors who provided the evaluation were informed in writing (at the top of the rating scale) of the purpose for which the rating was to be used (Landy, 1989).

RESULTS

SAMPLE

The total sample comprised 97 participants. This sample was broken down into employees (n=86) and students (n=11). Of the employee sample, 78 elected to take part in the validity study. The remaining respondents (10 in total) completed the test but did not provide any criterion information and therefore could not be included in the validity study.

1 employee and 1 student participant returned their completed forms too late to be included in the validity study. Their test scores, however, were included in the overall descriptive statistics (ie the total mean and standard deviation calculations).

A number of the large organisations who were contacted about taking part in this research were undergoing restructuring and therefore could not take part. Others, due to workload, were unable to contribute bigger (or any) samples. Some organisations had adopted the policy of not taking part in research. Other companies, because of their own preferred employment selection methods, did not wish to introduce new methods, or take part in any research, which used different selection tools. Therefore, of necessity, many different companies (large and small) from all over New Zealand were asked for either only one, or a few possible volunteers. This had the advantage of obtaining a very interesting sample from a wide array of companies from many different market sectors.

DESCRIPTIVE STATISTICS

For the initial data analysis, the mean, median, mode and standard deviation were computed for both the employee sample and the student samples (see Table 1 for a summary of the mean scores and standard deviations).

Table 1. Summary of Mean Test Scores and Standard Deviations for Employees and Students

Sample	**N**	**Mean Test Score**	**Standard Deviation**
Employees	86	50	10
Students	11	49	7

RELIABILITY

The reliability of the test was computed for the entire sample (employees and students combined). Split half reliability was calculated using the alpha coefficient. The average reliability of every way the test could be split was 0.51.

VALIDITY

The validity of the MRA test was calculated for each of the 2 sub-samples: the employee group and the student group using different criteria for each. For the employee group total test scores were correlated first with highest educational level achieved and then with salary. In the case of the student sample the median grade received for the preceding year's exam marks was utilised as the criterion. Validity was calculated using Spearman's rho because each of the criteria were reported as a range or a set of ordered categories rather than an actual figure, making Spearman's

rho more appropriate than the usual Pearson's correlation. (See Table 3 for a summary of all of the validity coefficients).

The validity coefficient for the employee sample using the criterion of educational level was rho=0.39. This figure is statistically significant at the .01 level. The magnitude of this coefficient, although moderate, is within the range of others obtained for cognitive ability tests. When salary was the criterion there was little indication of any correlation.

The validity coefficient for the student sample was rho=0.55. This is comparatively higher than others reported for cognitive ability tests, but due to the small size of this sub-sample it is not statistically significant. Therefore, this figure is probably best regarded as an indication of a trend.

Partial Correlation

It was thought that there might be a link between salary and age, and education and age. Consequently, a partial correlation was computed for the employee sample controlling for age, in case this was a lurking variable (Moore & McCabe, 1993). There was very little correlation between salary and percent. There was however, a significant relationship between education and percent ($r=0.39$, $p<0.001$).

Validity across Market Sectors

The employee sample was then subdivided into 3 loosely related market sectors (see Table 2) and validities were calculated for each of the 3 categories. The first industry group comprised government and tertiary education employees, the second 'sales and clerical' employees and the third group was composed of 'agricultural / manufacturing / service industry' workers. The final groupings of industries in the second and third categories were suggested from the categorisations found in the

Dictionary of Occupational Titles. Again, candidate's total scores were correlated with educational level and then with salary.

Table 2. Composition of Market Sectors

Market Sector 1 Government /Education	**Market Sector 2 Clerical / Sales**	**Market Sector 3 Agriculture/Manufacturing/ Service**
Public Service	Retail	Agriculture
Tertiary Education	Sales	Horticulture
Government	Marketing	Manufacturing
Local Government	Information Provider	Forestry
Research / Technology / Science	Information / Technology	Production
	Telecommunications	Transport
	Communications	Vehicle Servicing
	Media	Towage & Salvage
	Real Estate	Tourism
	Insurance	Entertainment Hire
	Finance	Service Industry
	Human Resources	Hospitality

The validity coefficient for industry group 1 (government and education employees) was rho=0.36 with educational level as the criterion. With salary as the criterion, again, there was very little correlation.

The second industry group (sales and clerical) obtained a validity coefficient of rho= 0.39 with education as the criterion. Using the index of salary yielded a very small correlation rho=0.15.

The validity coefficients for Industry group 3 (agriculture / manufacturing / service) indicated a low correlation for educational level (rho=0.15) and none for salary.

Table 3. Summary of Validities, Means and Standard Deviations obtained for each sub-sample

Sample Group	**N**	MRA Test Scores		**Correlation with Criterion**		
		Mean	**S.D.**	**Education Level**	**Salary**	**Median of Last Year's Exam Marks (Students Only)**
All Employees	78	50	10.3	0.39**	-0.72	-
Occupational Group 1 (Junior Level)	25	59	9.5	0.3	0.39	-
Occupational Group 2 (Middle Level)	34	48	9.7	0.41*	0.15	-
Occupational Group 3 (Senior Level)	19	46	8.8	0.02	-0.18	-
Market Sector 1 (Govt / Education)	28	54	9.8	0.36	0.04	-
Market Sector 2 (Clerical / Sales)	23	48	11.2	0.39	0.15	-
Market Sector 3 (Ag /Mfg /Service)	26	47	8.6	0.15	-0.13	-
Students	10	50	7	-	-	0.55

* Correlation is significant at the 0.05 level
** Correlation is significant at the 0.01 level

Partial Correlation for Market Sector 1

A partial correlation was calculated for the government and education employees, controlling for age. There was a very low correlation between salary and percent. The correlation between education and percent was not significant ($r=0.36$, $p<0.06$) but it would probably be worthwhile to investigate this further with a larger sample.

Validity across Occupational Levels

The employee sample was then partitioned according to the approximate occupational level of the participants and validity coefficients were calculated for each of the resulting 3 categories. The three sub-samples were 'junior level' (included in this category were junior level managers as well as occupations of a similar or lower level), 'middle management', and 'senior management'. The senior management category also included small business owners because some of these people listed their occupations as 'Company Director' or 'Business Partner', and although some indicated their business was a small concern, others did not. Consequently, it was impossible to differentiate between individuals who held these positions in large corporates and those who operated a small family business.

Occupational level group 1 (junior management and other junior level jobs) obtained a validity coefficient of rho=0.3 with education as the criterion. When salary was utilised as the index, the correlation coefficient was rho=0.39. The correlation coefficient obtained using salary as the criterion is interesting. This sub-sample is the only group to yield a coefficient of moderate magnitude. This may indicate that salary is more linked to performance at the lower levels of the organisational hierarchy where quality of output is possibly easier to monitor.

For the middle management level group the validity was rho=0.41, using education as the criterion. There was evidence of a small correlation when salary was used as the criterion (rho=0.15).

The third occupational level group showed no correlation with either education or salary. This result is not surprising considering the diversity of this group (ie senior level management and small business owners and partners).

Correlations of the 5 Sub-scales Included in the Test for the Total Employee Sample

Participants' answers for each of the five answer options on the test (ie 'true', 'probably true', 'insufficient data', 'probably false' and 'false') were totalled and then correlated with each other and also with the salary and education, to discover if there were any relationships amongst them. This was done for the full employee sample.

There appears to be a moderate correlation between PT and PF categories indicating that people who do well on PT also do well on PF. Conversely, those who score poorly on PT also score poorly on PF. This result is expected as these 2 options both deal with incomplete information and ambiguity.

A similar significant relationship also exists between T and F. Again, this result is expected because, in this instance, the information dealt with in both types of questions is unambiguous and mostly straightforward. This suggests that people either perform well in ambiguous situations or in unambiguous situations. Or, that the dimensions of T and F are separate from the dimensions of PT and PF.

There is also a significant relationship between ID and PF. These results indicate that individuals who can identify conclusions that are PF also score fairly well on the ID option.

With educational level as the criterion, there appears to be a significant correlation between the ability to handle PT, PF and inferences that have insufficient data to support them. Given that education is meant to foster more advanced reasoning skills, this link is expected.

DISCUSSION

Validity of the MRA using Educational Level as the Criterion

When the criterion was the highest educational level achieved, the validity coefficient for the entire employee sample was 0.39, significant at the 0.01 level. The sample was then divided along occupational level lines. Junior level employees yielded a validity coefficient of 0.3, and middle level management employees returned a significant correlation (rho=0.41, $p< 0.05$). The lack of a correlation for senior level employees may be due to the diversity of this group, which included small business owner/operators as well as employees from the senior echelons of their respective organisations. The sample was also partitioned according to market sectors. Market sectors 1 and 2 obtained moderate validity coefficients (0.36 and 0.39 respectively) indicating that the test may be of some use for selecting staff from these industries. However, neither of these correlations was significant, due to the small sample sizes, therefore these findings would need to be confirmed using larger sample sizes.

The magnitude of these validity coefficients (apart from those obtained for occupational level 3 and industry group 3) indicates that the test has moderate validity, using the amount of education completed as the criterion. In addition, these coefficients are within the range of those reported for other cognitive ability tests.

The amount of education completed was chosen as a performance index because it can be viewed as the adult out-of-school version of school grades (Anastasi, 1988) as well as being linked to later career success (Judge, et al 1995). In addition, the

amount of education that a person has completed should have an impact on the ability to draw inferences.

Salary as the Criterion

Overall, the correlation between salary and performance on the test was not significant. In fact, apart from Occupational Group 1 (junior level employees), there appeared to be a very low relationship between the two. There could be a number of reasons for this result.

Salary may not be a good or reliable index of work performance. Bartol and Martin (1988) comment that although performance related pay is often viewed with approval, it is very rarely effectively put into practice in organisations. In fact, when managers (who were responsible for making salary increment recommendations) were polled on their salary-raise criteria, Sherer, Schwab, and Heneman (1986) discovered that factors other than performance influenced their decisions. These factors included length of tenure, current salary, other pending job offers and consistency of employee's performance to date. One manager rated another job offer almost as highly as performance. This research also pointed to the very different weights managers assign to different criteria and the potential effect this would have on pay allocation decisions.

Other factors that may limit the usefulness of salary as a criterion could relate to the differences in the structure, range, and levels of pay and also salary ceilings across industries. This would make it very difficult to make performance comparisons on the basis of salary level between a high paying industry and a market sector with relatively lower pay levels. Thus, a junior level manager in one industry may receive a similar salary to a middle-level employee in another. It is also possible that due to these remuneration differences, a low performer in one industry may receive a similar salary to a middle-level performer in another industry. At this

point in time there are still differences in salary rates between the private and public sectors, as well as gender differences in remuneration. The sample used in the present research comprised individuals, both male and female, from a very wide array of private and public market sectors throughout New Zealand. This could very well have played a significant role in the low correlations obtained when salary was utilised as the performance index.

However, in Occupational Level Group 1 (junior level managers and employees) a moderate correlation between salary level and test scores was apparent (rho=. 39). This relationship suggests that, at the junior level, performance may be linked to salary. It may well be that the work output quality is more immediately apparent at the lower levels of the organisation than at the senior levels, and salary increments are awarded accordingly. At the senior levels of management, factors other than performance may be influencing salary levels. In addition, it may be that performance at the more elevated levels of an organisation may be difficult to define and measure, or that the outcomes of performance may not be manifested for many years.

Interestingly, the Watson-Glaser Critical Thinking Appraisal Manual (1994) reported similar results when salary was used as a criterion. When the sample comprised lower level management applicants, the correlation between salary and test scores was r=. 25. For middle level management participants r= -. 06 and for senior level management applicants, r=. 18. It is to be noted however, that their sample sizes were larger.

Age did not appear to have an effect on percentage scores for either salary or education.

Validity of the Student Sample

To assess the validity of the student sample, the median grade received for the previous year's exams were used as the criterion. This yielded a correlation coefficient of rho=0.55. This figure is at the higher end of the spectrum for cognitive ability tests. However, due to the small sample size, this coefficient is not significant and should therefore be regarded as an indication of a trend which would be interesting to investigate further.

Reliability

Reliability was computed using coefficient alpha, which is a measure of internal consistency. Kline (1986) cites Cronbach and Nunally as stating that this is the "most important index of test reliability". This study obtained a reliability coefficient of 0.51.

It has been recommended that a sample size of no less than 200 be used for assessing reliability, and that larger samples are even better (Kline, 1986). This study had a total of 96 participants in the reliability study. Consequently, any effects that may influence test scores, and thus reliability, could have been magnified. For example, as a person progresses through the test they may "warm-up" mentally, or begin to think about their answers or thought processes. One could speculate that this could alter response patterns and perhaps show evidence of an alpha-beta effect, with candidates performing better on the latter half of the test, due to learning or contemplation of the inference – drawing process. Alternatively, due to time constraints or distractions some individuals may not be able to attend to the test, or their answers, and this could have an impact on their results. With very large samples, errors from these sources can be minimised.

In addition to the size of the sample, the composition of the sample can also have an effect on the reliability. The sample should mirror the target population (Kline, 1986). The present sample was very heterogeneous comprising participants from many market sectors, different occupational levels, large corporations and small businesses, and from both the private and public sector. The composition of this sample was governed by the respondents who completed and returned their tests. However, the majority of the participants in the sample were managers, albeit from diverse backgrounds.

A perusal of the results shows that if question 17 were deleted from the test, the reliability would become 0.54. Participants appear to have found this question difficult as a high percentage of people answered it incorrectly.

Correlations between Test Items

Participants' scores on the 5 sub-sets of the MRA test (ie the 'true', 'probably true', 'insufficient data', probably false' and 'false' options) were totalled and correlated with each other. The results suggest a relationship between the ability to identify 'true' and 'false' items and a correlation between the ability to identify 'probably true' and 'probably false' conclusions. These results are expected as the 'true' and 'false' options both deal with unambiguous information, whereas the 'probably true' and 'probably false' choices both deal with ambiguous information. This suggests that the dimensions of 'true' and 'false' are distinct from the dimensions of 'probably true' and 'probably false' or, that perhaps people either have a facility for dealing with ambiguous situations or a proficiency in dealing with unambiguous information. However, these findings can be viewed as indications only and would need to be verified by using a larger sample. There also appears to be a link between education and the ability to deal with ambiguous situations. This relationship is expected because it is to be hoped that education endows people with the skills to enable them to reason in a subtler manner.

Concurrent Validity Design versus Predictive Validity Design

The present research investigated the validity of the MRA using a concurrent validity design. Several of the advantages associated with using this design include the (relative) ease of access to participants, as the sample is drawn from current employees. Also, the criterion measures, which are collected at the same time as the predictor scores, are perhaps easier to obtain. However, there are also a number of drawbacks concomitant with the concurrent design. One of these is restriction of range, as performers at the extreme ends of the spectrum are unlikely to be part of the employee sample (ie they may have been fired or promoted). However, Barrett, Phillips and Alexander (1981) point out that predictive validity designs may also be subject to restriction of range, albeit to a lesser extent, due to the hiring practices of the organisation, as not all of the applicants (who comprise the predicative validity sample) are hired. Some researchers did not find any difference between the magnitude of the obtained validity coefficients in predictive and concurrent validity designs, indicating that concurrent validity designs are useful in a predictive setting. The present concurrent validity study, therefore, may be viewed as a pilot study for a more comprehensive predictive validity study to be undertaken at a doctoral level.

Rationale behind the Present Research

Research has indicated that cognitive ability tests have predictive validity for selecting staff, and that the validity of these tests increases as the cognitive complexity of the job increases. Landy, Shankster and Kohler (1994) note that there is little research currently being done on developing and testing new cognitive ability tests. The present study is an attempt to fill some of this gap. The MRA test is an original cognitive ability test, which was designed to assess the ability to draw inferences. This skill is central to both the critical thinking process and decision-making. The MRA was specifically designed for managers to assess their skills in this area, as decision-making is one of the core competencies required for

successful managerial performance. In addition, the test utilised workplace scenarios in an effort to increase the test's face validity. Further study of this test will be necessary to evaluate its potential as a selection test.

Further Study

To further assess the psychometric properties of the MRA a more homogeneous sample could be used. This would involve drawing a sample from either a large organisation, which was willing and able to participate, or from one market sector e.g. finance. The benefit of utilising one large organisation is the greater consistency that would result from using criterion measures such as salary or supervisor's ratings, which can often be influenced by context, industry standards and organisational climate. Conversely, sampling from one market sector would probably provide a more stable base for comparison with the criterion of salary, as there would be more likely to be a coherent structure to pay rates across the range. This would effectively mean comparing 'apples with apples' rather than 'apples with oranges'. In addition, if a large enough sample were taken from each of several different market sectors, or large corporates, this would enable the inter-organisational validity of the instrument to be explored and valid comparisons to be made. Norms could also be constructed for these groups.

To help diminish the effects of restriction of range (one of the drawbacks of concurrent validity designs) it would be useful to perform a predictive validity study on the MRA. Although a true predictive validity study is extremely difficult to do (Landy, 1989), should an opportunity present itself, this would be a valuable method of exploring the predictive validity of the MRA.

Another worthwhile study would be to correlate the test scores obtained on the MRA with those gained on an IQ test. One would expect a fairly good correlation

due to the apparent link between IQ level and the ability to draw inferences (Gadzella, Hartsoe & Harper, 1989).

It would also be interesting to investigate the validity and reliability of the MRA using samples from different countries to assess it's potential as a selection instrument in an international setting. Again, the above points about sampling would also need to be taken into consideration. If the test were to be translated into another language, careful consideration would need to be taken to ensure that the nuances present in the text were maintained and none extra added, as nuances play a role in the detection of inferences. The place names mentioned in most of the scenarios outlined in the MRA are Australasian - these could easily be changed to suit the particular country the sample was being drawn from, or, the place names could be altered to completely fictitious locations.

In terms of assessing the reliability of the MRA, the test-retest form of reliability could be used to measure the stability of the test over time.

Although the ability to draw inferences is a crucial component in the decision making process, and is the skill which most differentiates between above average and poor readers, it is not the only capability required to make rational and logical decisions. It may, therefore, be of value to construct further tests, which assess other aspects of the ability to think critically and effective decision-making. These could be used in tandem with the MRA for selection purposes. Conversely, one test could be designed, which contains sub-sections, each evaluating different dimensions of this process.

The MRA was primarily designed and constructed as an employment selection device. However, there could be some place for its utilisation in the context of professional development. As the ability to draw inferences is a skill which can be

taught, particularly by those with average and above average IQ levels (Gadzella et al, 198?) perhaps scores could be compared before and after such training is given, to assess any benefits that may have accrued from such training. The MRA would have to be subjected to the appropriate testing first to ascertain its merits in this regard.

Limitations of the Present study

The limitations of the present study arise from the sample size. Ideally, in test development very large samples are required to properly validate a test. Large samples enable sources of error to be minimised (Anastasi, 1988) and thus give more accurate estimates of reliability. Although the sample was randomly selected, participation was voluntary, therefore there is the possibility of some bias as a result of the self-selection of participants into the study as well as non-response.

CONCLUSIONS

The results of this research indicate that for the employee sample, the MRA test has moderate (statistically significant) validity when 'highest educational level achieved' was utilised as the criterion. It would appear that this test may also be valid for students, however this avenue would need to be confirmed using a larger sample. In addition, the test has been found to be reliable. Overall, the results of this research indicate that the MRA test may indeed be suitable for the selection of managerial level staff. Again, however, these findings would need to be further evaluated using a larger sample.

REFERENCES

Aiken, J. (1988). *Psychological Testing and Assessment.* (6th ed). Massachusetts: Allyn & Bacon Inc.

Amir, Y., Kovavsky, Y., & Sharan, S. (1970). Peer nominations as a predictor of multistage promotions in a ramified organisation. *Journal of Applied Psychology*, 54, 462-469.

Anastasi, A. (1988). *Psychological Testing.* (6th ed). New York: Methuen.

Antonioni, D. (1994). The effects of feedback accountability on upward appraisal. *Personnel Psychology,* 47, 348-356.

Arvey, R.D., & Campion, J.E. (1982). The employment interview: a summary and review of recent research. *Personnel Psychology*, 35,281-322.

Asher, J.J., Sciarrino, J.A. (1974). Realistic work samples: A review. *Personnel Psychology,* 27, 519-534.

Austin, J.T., & Villanova, P. (1992). The Criterion Problem 1917-1991. *Journal of Applied Psychology,* 77(6), 836-874.

Barrett, G.V., Phillips, J.S., & Alexander, R.A. (1981). Concurrent and predictive validity designs: A critical reanalysis. *Journal of Applied Psychology,* 66(1), 1-6.

Bartol, K.M. & Martin, D.C. (1988). Influences on managerial pay allocations: A dependency perspective. *Personnel Psychology*, 41, 36-377.

Berry, L.M., & Houston, J.P. (1993*). Psychology at Work.* Dubuque: Brown and Benchmark.

Blair, J.A. (1992). The evaluation of sources. In P.Norris (Ed), *The Generalisabiltiy of Critical Thinking – Multiple Perspectives on an Educational Ideal*. Columbia: Teachers College Press.

Blum, M.L., & Naylor, J.C. (1968). *Industrial Psychology: Its Theoretical and Social Foundations*. New York: Harper & Row.

Borman, W., White, L., & Dorsey, D. (1995). Effects of ratee task performance and interpersonal factors on supervisor and peer performance ratings. *Journal of Applied Psychology*, 80(10, 168-177.

Brodt, S.E. (1990). Cognitive illusions and personnel management decisions. *International Review of Industrial and Organizational Psychology*, 5, 229-279.

Campbell, R.J., & Bray, D.W. (1993). Use of an assessment centre as an aid in management selection. *Personnel Psychology*, 46, 691-699.

Campion, J.E. (1972). Work sampling for personnel selection. *Journal of Applied Psychology*, 56(1), 40-44.

Carretta, M.J., & Ree, T.R. (1998). General cognitive ability and occupational performance. *International Review of Industrial and Organisational Psychology*, 13, 159-184.

Cascio, W.F. (1991). *Applied Psychology in Personnel Management.* (4th ed). Englewood Cliffs, NJ: Prentice Hall.

Chan, D., Schmitt, N., DeShon, R.P., Clause, C.S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationship between race, test performance, face validity perceptions and test taking motivation. *Journal of Applied Psychology*, 82(2), 300-310.

Colberg, M. (1985). *Logic based measurement of verbal reasoning: A key to increased validity and economy.* Personnel Psychology, 38(2), 347-360.

Collins, A., Brown, J.S., & Larkin, K.M. (1980). Inference in text understanding. In R. J. Spiro, B. C. Bruce & W. F. Brewer (Eds), *Theoretical Issues in Reading Comprehension. Perspectives from Cognitive Psychology, Linguistics, Artificial Intelligence, and Education.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Cooper, W.H. (1981). Ubiquitous halo. *Psychological Bulletin*, 90(2), 218-244.

Dakin, S., & Armstrong, J.S. (1989). Predicting job performance: A comparison of expert opinion and research findings. *International Journal of Forecasting*, 5, 187-194.

DeNisi, A.S., & Shaw, J.B. (1977). Investigation of the uses of self-reports of abilities. *Journal of Applied Psychology*, 62(5), 641-644.

DeNisi, A.S.. Randolph, W.A., & Blencoe, A.G. (1983). Potential problems with peer ratings. *Academy of Management Journal*, 26, 457-464.

Drasgow, F. & Miller, H.E. (1982). Psychometric and substantive issues in scale construction and validation. *Journal of Applied Psychology,* 67(3), 268-279.

Dunnette, M.D., & Borman, W.C. (1979). Personnel selection and classification systems. *Annual Review of Psychology,* 30, 477-525.

Einhorn, H.J., & Hogarth, R.M. (1981). Behavioural decision theory: Processes of judgement and choice. *Annual Review of Psychology*, 32, 53-88.

Ennis, R.H. (1990). A taxonomy of critical thinking dispositions and abilites. In J.B. Baron and R.J. Sternberg (Eds*), Teaching Thinking Skills: Theory and Practice.* New York: W.H. Freeman & Co Ltd.

Fox, S., Ben-Nahum., & Yinon, Y. (1989). Perceived similarity and accuracy of peer ratings. *Journal of Applied Psychology,* 74, 781-786.

Gadzella, B.M., Hartsoe, K., & Harper, J. (1989). Critical thinking and mental ability groups. *Psychological Reports,* 65, 1019-1026.

Gaugler, B.B., Rosenthal, D.B., Thornton III, G.C., Bentson, C. (1987). Meta-analysis of assessment centre validity. *Journal of Applied Psychology,* 72(3), 493-511.

Ghiselli, E.E. (1973). The validity of aptitude tests in personnel selection. *Personnel Psychology,* 26, 461-477.

Glaser, E.M. (1937). *An experiment in the development of critical thinking.* Contributions to Education, No. 843. New York: Bureau of Publications, Teachers College, Columbia University.

Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist,* 34(1), 93-104.

Goldstein, W., Yusko, K.P., Braverman, E.P., Smith, D.B., & Chung, B. (1998). The role of cognitive ability in the subgroup differences and incremental validity of assessment centre exercises. *Personnel Psychology,* 51(2), 357-374.

Gordon, M.E., & Kleiman, L.S. (1976). The prediction of trainability using a work sample test and an aptitude test: A direct comparison. *Personnel Psychology*, 29(2), 243-253.

Guion, R.M. (1978). Scoring of content domain samples: The problem of fairness. *Journal of Applied Psychology*, 63(4), 499-506.

Guion, R.M., & Cranny, C.J. (1982). A note on concurrent and predictive validity designs: A critical reanalysis. *Journal of Applied Psychology*, 67(2), 239-244.

Harris, M.M., and Dworkin, J.D. (1990). Preemployment screening procedures: How human resource managers perceive them. *Journal of Business and Psychology*, 4(3), 279-292.

Hoffman, C.C., & Thornton III, G.C. (1997). Examining selection utility where competing predictors differ in adverse impact. *Personnel Psychology*, 50(2), 455-470.

Huffcutt, A.I., & Arthur, Jr. W. (1994). Hunter and Hunter revisited: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, 79, 599-616.

Huffcutt, A.I., Roth, P.L., & McDaniel, M.A. (1996). Constructs assessed in interviews. *Journal of Applied Psychology*, 81(5), 459-473.

Hunter, J.E., & Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, 96, 72-98.

Jagacinski,C.M. (1991). Personnel decision making: The impact of missing information. *Journal of Applied Psychology*, 76, 19-30.

James, L.R., & Ellison, R.L. (1973). Criterion composites for scientific creativity. *Personnel Psychology*, 26, 147-161.

Janis, I.L., & Mann, L. (1977). *Decision Making. A Psychological Analysis of Conflict Choice and Commitment*. The Free Press: New York.

Janz, T. (1982). Initial comparisons of patterned behaviour description interviews versus unstructured interviews. *Journal of Applied Psychology*, 67, 577-580.

Johnston, R.H. (1994). The problem of defining critical thinking. In P. Norris (Ed), *The Generalisabiltiy of Critical Thinking – Multiple Perspectives on an Educational Ideal by Teachers*. Columbia: College Press.

Joyce, L.W., Thayer, P, Pond III, S.B. (1994). Managerial functions*: An alternative to traditional assessment centre dimensions*, 47, 109-122.

Judge, T.A., Cable, D.M., Bougreau, J.W., & Bretz Jr, R.D. (1995). An empirical investigation of the predictors of executive career success. *Personnel Psychology,* 48, 485-519.

Kaplan, M.F., Wanshula, T., & Zanna, M.P. (1993). Time pressure and Iiformation integration in social judgement. The effect of need for structure. In A.J. Maule and O. Svenson (Eds), *Time Pressure and Stress in Human Judgement and Decision-making.* New York: .Plenum Press.

Keenan, T. (1989). Selection and interviewing. *International Review of Industrial and Organizational Psychology*, 4, 1-23.

Kidder, L.H. (1982. *Research Methods in Social Relations*. (4th ed). New York: Holt-Saunders.

Klimoski, R., & Brickner, M. (1987). Why do assessment centres work? The puzzle of assessment centre validity. *Personnel Psychology*, 40, 243-260.

Klimoski, R.J., & Strickland, W.J. (1977). Assessment centres – valid or merely prescient? *Personnel Psychology*, 30,353-361.

Kline, P. (1989). *A Handbook of Test Construction.* New York: Methuen.

Krzystofiak, R., Cardy, R.L., & Newman, J. (1988). Implicit personality and performance appraisal: The influence of trait inferences on evaluations of behaviour. *Journal of Applied Psychology,* 73, 515-521.

Landy, F.J. (1989). *The Psychology of Work Behaviour.* (3rd ed). California: Brooks/Cole Publishing Company.

Landy, F. J. & Farr, (1980). Performance rating. *Psychological Bulletin*, 87(1), 72-107.

Landy, F.J., Shankster, & Kohler (1994). Personnel selection and placement. *Annual Review of Psychology* ,46,261-296.

Latham, G.P. Saari, L.M., Russell, E.P.,& Campion, M.A. (1980). The situational interview. *Journal of Applied Psychology,* 65, 422-427.

Latham, G.P., & Whyte, G. (1994). The futility of utility analysis. *Personnel Psychology*, 47, 31-46.

Lent, R.H., Aurbach, H.A., & Levin, L.S. (1971). Predictors, criteria, and significant results. *Personnel Psychology*, 24, 519-532.

Mabe, P.A., & West, S.G. (1982). Validity of self evaluation of ability. *Journal of Applied Psychology*, 67, 280-296.

Martin, J.R. (1994). Critical thinking for a humane world. In P.Norris (Ed), *The Generalisability of Critical Thinking – Multiple Perspectives on an Educational Ideal*. Columbia: .Teachers College Press.

Maule, A.J., & Svenson, O. (1993). Theoretical and empirical approaches to behavioural decision-making and their relation to time constraints. In A.J. Maule and O. Svenson (Eds), *Time Pressure and Stress in Human Judgement and Decision-making*. New York: Plenum Press.

McDaniel, M.A., Whetzel, D.L., Schmidt, F.L. & Maurer, S. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, 79, 599-616.

McPeck, J.E. (1981). *Critical Thinking and Education*. Oxford: Martin Robertson & Company Ltd.

Miller, D.C. (1983). *Handbook of Research Design and Social Measurement*. (4th ed). New York: Longman.

Mitchell, T.R., & Beach, L.R. (1990). ...Do I love thee? Let me count...Toward an Understanding of intuitive and automatic decision-making. *Organisational Behaviour and Human Decision Processes*, 47, 1-22.

Moore, D.S. & McCabe, G.P. (1993). *Introduction to the Practice of Statistics*. (2nd Ed). New York: W.H. Freeman & Company Ltd.

Moses, J.L. (1973). The development of an assessment centre for the early identification of supervisory potential. *Personnel Psychology*, 26, 569-580.

Murphy, K.R., & Shiarella, A.H. (1997). Implications of the multidimensional nature of job performance for the validity of selection tests: Multivariate frameworks for studying test validity. *Personnel Psychology*, 50, 823-854.

Nisbett, R., & Ross, L. (1989). *Human Inference Strategies and Shortcomings of Social Judgement*. Englewood Cliffs, New Jersey: Prentice Hall Inc.

Nisbett, R.E., Krantz, Jepson, C., & Fong, G.T. (1982). Improving inductive inference. In Kahneman, D., Slovic, P., & Tversky, A. (1984). *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge University Press:

Norris, S.P., & Ennis, R.H. (1990). *Evaluating Critical Thinking*. Victoria: Hawker Brownlow Education.

O'Hare, M.A. (1997). *Reading and Understanding Assessment - A Selection Test for Managers*. Unpublished Honours Project, Victoria University, New Zealand.

Phillips, L.M. (1988). Young readers' inference strategies in reading comprehension. *Cognition and Instruction*, 5(3), 193-222.

Raju, N.S., & Burke, M.J. (1986). Utility analysis. In R. A. Berk (Ed), *Performance Assessment*. John Hopkins Press.

Ree, M.J., Earles, J.A., & Teachout, M.S. (1994). Predicting job performance not much more than g. *Journal of Applied Psychology*, 79(4), 518-524.

Reilly, R.R., & Chao, G.T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology*, 35, 1-62.

Rippin, S.M. (1995). *The Competencies Used to Assess the Effectiveness of New Zealand Managers*. Unpublished Thesis, Victoria University, New Zealand.

Roth, P.L., BeVier, C.A., Switzer III, F.S., 7 Schippmann, J.S. (1996). Meta-analyzing the relationship between grades and job performance. *Journal of Applied Psychology*, 81(5), 548-556.

Rothstein (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology*, 75, 322-327.

Roszkowski, M.J., & Bean, A.G. (1990). Believe it or not! Longer questionnaires have lower response rates. *Journal of Business and Psychology*, 4(4), 495-509.

Rudman, R. (1991). *Human Resource Management in New Zealand*. Auck, NSZ: Longman Paul.

Ryan, A.M., & Sackett, P.R. (1987). A survey of individual assessment practices by I/O Psychologists. *Personnel Psychology*, 40,455-488.

Saal, F.E., Downey, R.G., & Lahey, M.A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin,* 88, 413-414.

Salgado, J.F. (1999). Personnel Selection methods. *International Review of Industrial and Organisational Psychology,* 14, 1-54.

Schmidt, F.L., & Hunter, J.E. (1981). Employment testing: Old theories and new research findings. *American Psychologist*, 36(10), 1128-1167.

Schmidt, F.L., Hunter, J.E., & Pearlman, K. (1981). Task differences and validity of aptitude tests in selection: A red herring. *Journal of Applied Psychology,* 66, 166-185.

Schmidt, F.L., Hunter, J.E., McKenzie, R.C., & Muldrow, T. (1979). The impact of valid selection procedures on work force productivity. *Journal of Applied Psychology*, 64, 609-626.

Schmidt, F.L., Hunter, J.E., Outerbridge, A.N., & Tratnee, M.H. (1986). Economic impact of job selections on size, productivity and payroll costs of the federal workforce: an empirically based demonstration. *Personnel Psychology,* 39, 1-32.

Shank, R.C. (1988). Creativity as a mechanical process. In R.J. Sternberg (Ed), *The nature of Creativity: Contemporary Psychological Perspectives*. Cambridge University Press: Cambridge.

Shanteau, J., & Dino, A.D. (1993). Environmental stressor effects on creativity and decision making. In A.J. Maule and O. Svenson (Eds), *Time Pressure and Stress in Human Judgement and Decision-making.* New York: Plenum Press.

Shaugnessey, J.J., & Zechmeister, E.B. (1994). *Research Methods in Psychology.* 3rd ed. New York: McGraw-Hill.

Sherer, P.D., Schwab, D.P., Heneman III, H.G. (1986). Managerial salary-raise decisions: A policy capturing approach. *Personnel Psychology*, 40, 27-38.

Siegel, H. (1988). *Educating Reason – Rationality, Critical Thinking and Education.* London: Routledge.

Siegel, H. (1992). The generalisability of critical thinking skills, dispositions and epistemology. In P.Norris (Ed), *The Generalisability of Critical Thinking – Multiple Perspectives on an Educational Ideal.* Columbia: Teachers College Press.

Simon, S.C. (1969). *Research Methods in Social Science – The Art of Empirical Investigation.* New York: Random House.

Smither, J.W., London, M., Vasilopoulos, N.L., Reilly, R.R., Millsap, R.E. & Salvemini, N. (1995). An examination of an upward feedback programme over time. *Personnel Psychology*, 48, 1-34.

Smither, J.W., Reilly, R.R., Millsap, R.E., Pearlman, K., Stoffey, R.W. (1993). Applicant reactions to selection procedures. *Personnel Psychology*, 46, 49-76.

Spychalski, A.C., Quinones, M.A., Gaugler, B.B., Pohley, K. (1997). A survey of assessment centre practices in organisations in the U.S. *Personnel Psychology*, 50, 71-90.

Sternberg, R.J., & Baron, J.B. (1987). *Teaching Thinking Skills.* New York: W.H. Freeman & Co.

Taylor, P., Mills, A., & O'Driscoll, M. (1993). Personnel selection methods used by New Zealand organisations and personnel consulting firms. *New Zealand Journal of Psychology*, 22, 19-31.

Taylor, S.E. (1984). The availability bias in social perception and interaction. In D. Kahneman, P. Slovic, & A. Tversky (Eds). *Judgement Under Uncertainty: Heuristics and Biases.* Cambridge University Press.

Treiman, D.J., & Hartmann, H.I. (1981). Women, work and wages. In D.J. Treiman & H.I. Hartman (Eds), *Women, Work and Wages.* Washington: National Academy Press.

Turban, D.B., Sanders, P.A., Francis, D.J., & Osborn, H.G. (1989). Construct equivalence as an approach to replacing validated cognitive selection tests. *Journal of Applied Psychology*, 74, 62-71.

Tversky, A., & Kahneman, D. (1984). Heuristics and biases. In D. Kahneman, P. Slovic, & A. Tversky (Eds). *Judgement Under Uncertainty: Heuristics and Biases.* Cambridge University Press.

Varma, A., DeNisi, A.S., & Peters, L.H. (1996). Interpersonal affect and performance appraisal: A field study. *Personnel Psychology,* 49, 341-360.

Vinchin, A., Schippmain, J.S., Switzer III, F.S., & Roth, P.L. (1998). A meta analytic review of predictors of job performance for sales people. *Journal of Applied Psychology*, 83(4), 586-597.

Viswesvarau, C., Ones, D.S., & Schmidt, F.L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, 81(5) 557-574.

Wagner, R.K. (1991). Managerial problem solving. In R. J. Sternberg & P.A. Frensch (Eds), *Complex Problem Solving: Principles and Mechanisms*. Hillsdale, New Jersey: Lawrence Erlbaum Associates Inc.

Wainer, H., & Baum, H.I. (1988). *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum Assocs.

Watson, G. B. (1925). *The measurement of fairmindedness*. Contributions to Education, no 176. New York: Bureau of Publications, Teachers College, Columbia University.

Watson, G.B., & Glaser, E.M. (1980). *Watson-Glaser Critical Thinking Appraisal Manual*. San Antonio, TX: Psychological Corporation.

Watson, & Glaser, E.M. (1994). *Watson-Glaser Critical Thinking Form S Manual*. San Antonio: Psychological Corporation.

Wedge, J.W., & Kavanagh. M.J. (1988). Improving the accuracy of performance evaluations: Comparison of three methods of performance appraiser training. *Journal of Applied Psychology*, 73 ,68-73.

Weisberg, R.W. (1988). Problem solving and creativity. In R.J. Sternberg (Ed), *The Nature of Creativity: Contemporary Psychological Perspectives*. Cambridge: Cambridge University Press.

Winkler, R.L., & Murphy, A.H. (1973). Experiments in the laboratory and the real world. *Organisational Behaviour and Human Performance*, 10, 252-270.

Wright, P.M., Lichtenfels, P.A., & Pursell, E.D. (1989). The structured interview: Additional studies and a meta-analysis. *Journal of Occupational Psychology*, 62, 191-199.

Zakay, D. (1984). The evaluation of managerial decisions' quality by managers. *Acta Psychologica*, 56, 49-57.

APPENDIX A
MANAGERIAL READING ASSESSMENT RESEARCH PROJECT

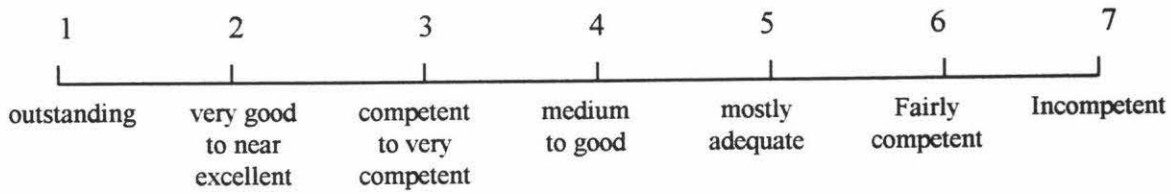
_____ has agreed to take part in the above research project. The aim of this research is to explore the reliability and validity of a test which assesses managers' ability to draw accurate conclusions from a set of given data. _____ is completing a questionnaire which tests this skill.

As part of this research s/he has given me permission to obtain a rating of his/her performance on the job. Could you please assist me by giving a rating of his/her performance?

Your rating will be completely confidential to yourself, my Massey academic supervisor and myself. No-one else will have access to it and it will be destroyed at the completion of this research (November, 1998).

Listed below are 10 managerial core competencies. Next to each competency please enter your rating of his/her performance, using the 7 point scale below.

RATING SCALE



COMPETENCIES	YOUR RATING
1. Communication skills	_____
2. Problem solving	_____
3. Planning/organising	_____
4. Networking	_____
5. Team building	_____
6. Decision making	_____
7. Managing/directing subordinates	_____
8. Analytical skills	_____
9. Technical skills/knowledge	_____
10. Monitoring/evaluating operations and work performance	_____

What is your overall assessment of this person's performance? _____

Thank you very much for taking the time to complete this rating scale. If you have any questions relating to this research, please do not hesitate to contact me at (e-mail) or write C/- Psychology Department, Massey University, Private Bag 11222, Palmerston North, or telephone My academic supervisor can be contacted at the same address or Tel (Palmerston North).

Mary O' Hare

APPENDIX B

Managerial Reading Assessment

INFORMATION SHEET

My name is Mary O'Hare and I am conducting some post-graduate research in Psychology under the supervision of Associate Professor Doug Paton from Massey University.

I have designed a cognitive ability test called Managerial Reading Assessment (MRA) which in its final form will be used as a selection test for people who need to make decisions in their job or who supervise others. This test targets the ability to draw inferences from a set of given data. The aim of this research is to explore the reliability and validity of the MRA test.

In this study you will be asked to complete the test (it takes approximately ½ hour to complete) and a statistical data sheet.

Your name will not appear anywhere on the MRA test sheet, or the statistical data sheet so this information will be anonymous. **ALL** information and data provided by you will be destroyed at the completion of the research (end of March, 1999). During the research, the only people who will have access to this data will be my academic supervisor and myself. The data will be stored apart in a secure place in a locked filing cabinet until it is destroyed.

If you would like to receive confidential individual feedback on your test results these will be mailed to you (please indicate the address you would like them sent to).

This research is voluntary and completely independent of your work. You can consent, decline, withdraw or refuse to answer particular questions without any repercussions.

If you would like to take part in this research please contact me at (e-mail) or write, or ring . My academic supervisor can be contacted at School of Psychology, Massey University, Private Bag 11222, Palmerston North.