Cognitive ability and job performance in a New Zealand service organisation

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

Master of Science

in

Industrial/Organisational Psychology

at Massey University, Manawatu,

New Zealand.

Catherine Mann

2011

Abstract

This study investigated cognitive ability and job performance theoretically and empirically. A New Zealand government organisation tested job candidates' verbal, numeric and abstract abilities during their selection procedure and appraised employees' task, contextual and team performance as part of their performance management system. The service-based organisation provided scores on these variables for 43 recently hired employees. There was partial support for the hypothesis that cognitive abilities were related, as numeric reasoning scores correlated significantly with verbal reasoning (r = .38, p = .018) and abstract reasoning scores (r = .36, p = .023). Verbal reasoning scores did not correlate significantly with abstract reasoning scores (r = .24, p = .128), though this was probably due to low power. Individual task and contextual performance ratings correlated with each other as hypothesised (r = .32, p = .036), supporting the theory that some performance processes relate to both task and contextual performance. Team dynamics were expected to obscure simple linear relationships between team performance and individual-level variables and, as hypothesised, team performance did not correlate significantly with task or contextual performance, or cognitive abilities. Abstract reasoning did not show significant positive correlations with task or contextual performance, contrary to expectations, indicating that participants already had job-related experience. Numeric reasoning was not expected to relate to task or contextual performance as work was service based and not likely to require numeric ability, which was borne out in the non-significant correlations. Verbal ability scores correlated positively with task performance ratings (r = .44; p <.001), supporting the hypothesis that verbal ability would be associated with task performance in a service organisation. Verbal reasoning scores did not correlate with contextual performance ratings. Implications of these results for criterion-related validity, as well as cognitive ability and job performance theories are discussed along with limitations of the study and suggestions for future research.

Acknowledgements

I am grateful to my supervisor, Dr Gus Habermann, for all his support and advice. My thanks also go out to the participating organisation, in particular the Human Resources manager, for all her time and effort in providing me with data and information about the organisation. OPRA Consulting Ltd and Psytech International Ltd, especially Paul Englert, Sarah Burke, Sue Sommerville and Paul Wood, have been very accommodating, introducing me to the participating organisation and providing access to test materials and supporting me with other miscellaneous assistance with many aspects of the project.

I would also like to thank in particular Eleanor and Ben Sutton, as well as Michael Woodward, Sheila and Stewart Mann for their incredibly constructive comments and suggestions on drafts. Finally, I am very grateful to the Mann family, Sutton family and Woodward family for their love and support that helped me complete this thesis in the wake of the Christchurch earthquakes.

Catherine Mann
7th April, 2011

TABLE OF CONTENTS

List of figures

List of tables

1 Introduction

1.1 Overview

When selecting personnel, there are a number of ways for organisations to compare job applicants, including collecting curriculum vitae or biographical data ('biodata') to compare applicants' relevant work experience and qualifications (Salgado, Viswesvaran, & Ones, 2002; Taylor, Mills, & O'Driscoll, 1993). There are also psychometric assessments for estimating individual differences in personality, integrity or cognitive ability (J. C. Hogan, Hogan, & Gregory, 1992; J. C. Hogan, Hogan, & Murtha, 1992; Mumford, Connelly, Helton, Strange, & Osburn, 2001; Ones & Viswesvaran, 2007). Most employers will want to interview candidates, one objective being to assess how well each will 'fit' in the job, team, or organisation (Borman, Hanson, & Hedge, 1997; Edenborough, 2005; Edwards, 1991; Hollenbeck, 2000; Kristof, 1996; Taylor, Keelty, & McDonnell, 2002; Vogel & Feldman, 2009).

Organisations want to know which assessments lead to a higher likelihood of a selection decision that benefits them (Ackerman & Humphreys, 1990; Hough & Oswald, 2000; Robertson & Smith, 2001; Sackett & Lievens, 2008). Examining the criterion-related validity of each assessment is one way of finding this out; in other words, examining how well each method of selection 'predicts' job performance (Binning & Barrett, 1989; Cronbach, 1990; Shultz & Whitney, 2005). Criterion-related validity of selection procedures is usually estimated by statistically analysing the strength of the relationship between the scores job applicants receive from a selection assessment and their subsequent ratings of job performance by a supervisor (e.g. Barrick & Mount, 1991; Hunter & Hunter, 1984; Schmidt & Hunter, 1998).

Meta-analytic research (Schmidt & Hunter, 2002; Schulze, 2004) has shown that scores on cognitive ability tests have the highest criterion-related validity of the

assessments mentioned above (Salgado, et al., 2002; Schmidt & Hunter, 1998, 2004). This is the case internationally and over a range of jobs (Bertua, Anderson, & Salgado, 2005; Hulsheger, Maier, & Stumpp, 2007; Salgado, et al., 2003). On average, differences in employees' cognitive ability test scores account for roughly 25% of the variance in their job performance ratings (Ones, Dilchert, Viswesvaran, & Salgado, 2010; Salgado, et al., 2002)[1]. The practical implication of this finding is that, in order to maximise the likelihood of selecting the 'best' candidate (in terms of subsequent job performance ratings), organisations should place particular importance to scores on cognitive ability tests in selection processes (Hunter & Hunter, 1984; Kuncel, Ones, & Sackett, 2010; Schmidt & Hunter, 1998, 2004).

Examination of the relationship between cognitive ability scores and job performance ratings is often focused at a very general level. That is, research tends to examine general cognitive ability and/or overall job performance (e.g. Gottfredson, 2002b; Kuncel, et al., 2010; Ones, et al., 2010; Schmidt, Shaffer, & Oh, 2008), with researchers only recently paying attention to different dimensions of job performance (e.g. Borman & Motowidlo, 1993; Campbell, 1990; LePine, Hollenbeck, Ilgen, & Hedlund, 1997) or specific cognitive abilities (Lang, Kersting, Hulsheger, & Lang, 2010). Given that cognitive ability tests are used increasingly in New Zealand (Taylor, et al., 2002; cf Taylor, et al., 1993), more detailed investigation of this relationship, both theoretically and empirically, may be of practical benefit.

Theoretical issues that will be addressed here are those that refer to the nature and measurement of cognitive ability and job performance: what is cognitive ability? What

[1] 25% explained variance corresponds to the average correlation of .5 between scores (Schmidt & Hunter, 1998). Using multiple selection methods and creating a composite score can provide incremental validity; that is, the composite score has a greater total correlation with job performance ratings. For instance, taking into account both cognitive ability and integrity test scores can predict job performance better than by using cognitive ability scores alone, increasing the total correlation with job performance ratings to .65 (Ones & Viswesvaran, 2007; Schmidt & Hunter, 1998). Whichever combination of methods is used, however, cognitive ability tends to be the largest contributor to the resulting correlation (Robertson & Smith, 2001; Schmidt & Hunter, 1998).

is job performance? How do we quantify individual differences in cognitive ability and job performance? The present analysis begins with a review of intelligence literature, because cognitive ability cannot be fully comprehended as distinct from intelligence. Research relating to job performance and its measurement will be similarly outlined and synthesised. Subsequently, theories explaining the observed relationships between cognitive abilities and job performance dimensions will be analysed.

This study then aims to empirically examine which specific cognitive abilities relate to which dimensions of job performance, by analysing relationships between cognitive ability test scores (obtained during a selection process) and job performance ratings (provided in annual performance appraisal) of employees of a New Zealand government organisation.

One objective is to inspect the pattern of associations between different cognitive abilities to see how they compare with those found in previous research. Another objective is to outline the relationships between different dimensions of performance, to see whether they are related at all and to analyse the strength of those relationships. The overarching objective is to scrutinise the relationships between the specific cognitive abilities and job performance dimensions. Evidence of relationships may contribute to theories of the process via which cognitive ability influences job performance. The theoretical and empirical analyses are intended to contribute to a greater understanding of the finer details of intelligence-job performance association.

1.2 Intelligence

1.2.1 Conceptualisations of intelligence

Intelligence as a psychological construct (Cronbach & Meehl, 1955; MacCorquodale & Meehl, 1948) has stimulated over 100 years of scholarly research and debate worldwide. While people may imagine that the everyday understanding of

intelligence is basically consistent, individuals and cultural groups[2] have been shown to have subtle or vast differences in their conceptualisations (implicit theories) of intelligence (Berg & Klaczynski, 2002; Irvine & Berry, 1988; Ruzgis & Grigorenko, 1994). Memory tends to play an important role in Chinese theories of intelligence, more so than in those of other nationalities such as Australians (Chen, 1994; Chen, Braithwaite, & Jong Tsun, 1982; Chen & Chen, 1988). U.S. adults' implicit theories variously include practical problem solving ability and verbal ability (Sternberg, Conway, Ketron, & Bernstein, 1981). In Africa, by contrast, national and tribal groups tend to describe intelligent behaviour more in terms of interpersonal or social abilities than cognitive abilities (Ruzgis & Grigorenko, 1994). Different cultures can describe intelligent and foolish behaviour as exact opposites (Berg & Klaczynski, 2002). Individuals within cultures differ in their conceptualisations of intelligence too, one studied dichotomy being whether intelligence is believed to be a fixed trait or a trainable skill (Dweck & Leggett, 1988; Elliott & Dweck, 1988; Mangels, Butterfield, Lamb, Good, & Dweck, 2006). These few examples demonstrate just how diverse theories about intelligence can be.

Scholars specialising in intelligence also differ in their conceptualisations. Their different opinions can be summarised by examining the results of two symposia to answer the question: "what is intelligence?" in 1921 and 1986 (Journal of Educational Psychology, 1921; Sternberg & Detterman, 1986). After the 1921 symposium, Spearman, an influential intelligence researcher, was disconcerted that there seemed to be no consensus among the 14 scholars: "... a word with so many meanings that finally it had none" (1927, p. 14). Jensen, a contemporary adherent of Spearman's, agreed that

[2] Though culture ordinarily connotes nationality or ethnicity (or a combination, such as African-American), cultures or cultural groups in this discussion are considered to be any group of individuals with some common membership that is desirable: it could include, for example, a work team or organisation where the individual identifies with the team (Haslam, 2001; van Knippenberg, 2000).

there seemed to be 25 similarly discordant conceptualisations after intelligence experts offered their views in the 1986 symposium: "psychologists are incapable of reaching a consensus on its definition" (1998, p. 48). Sternberg and Detterman (1986) examined common themes across the symposia, however, and concluded that intelligence is generally agreed to relate to: adaptability to the environment, basic mental processes, and higher-order thinking (such as logical reasoning and problem solving ability). Scholars contributing to the more recent symposium placed greater emphasis on the acquisition and processing of knowledge, and also tended to recognise the influence of experience and context on intelligence. These emphases were less evident in 1921 (Sternberg & Detterman, 1986).

Though there is some consensus that intelligence is related to cognitive ability, researchers still debate whether intelligence and cognitive ability are, for practical purposes, the same thing (Cowan, 2005; Eysenck, 1988b; Jensen, 1998; Neisser, et al., 1996; Wilhelm & Engle, 2005). Some argue intelligence is a general cognitive ability, others argue that it involves many possibly independent abilities. Another disagreement is the importance of environmental influences compared to heredity in intellectual development. Theories of multiple types of intelligence and the variety of models explaining intellectual development demonstrate researchers' differing perspectives, and some of the prominent theories are outlined in section 1.2.2 and 1.2.3.

Cognitive ability is intricately related to intelligence because you cannot study one without encountering the other. Terminology that recurred in the reviewed research, for example, included 'intelligence', 'competence', 'cognitive (or mental) ability', 'potential', 'skill', 'capacity' and 'aptitude'. Sometimes it appeared as though authors used pairs of these terms indiscriminately or even interchangeably, though they were

usually elucidated. For the purposes of this discussion, the terms can be assumed to imply different things.

One important implication is the stability or malleability of intelligence, which may be deliberately invoked by the use of different terms (Cowan, 2005). Typically, 'ability', 'potential', 'capacity' and 'aptitude' imply intelligence is a fixed or enduring trait (e.g. Gottfredson, 2002a; Jensen, 1998). 'Competence' and 'skill' imply that intelligence can be developed through learning and practice (e.g. Sternberg, 2005)[3]. The former terms also imply that intelligence is a latent (hidden) characteristic of an individual, while the latter terms tend to refer more specifically to the demonstration of 'intelligent' behaviour.

1.2.2 Theories of multiple intelligences

Gardner (1993) argued that musical ability such as Mozart's should be considered the equivalent (in terms of 'intelligence') with Einstein's logical-mathematical ability. According to Gardner, these abilities can just as adequately be described as 'intelligences', and the multiple intelligences making up his model are linguistic, musical, logical-mathematical, spatial, bodily-kinaesthetic, intrapersonal (knowledge of awareness of self) and interpersonal (knowledge and awareness of others)[4]. Though Gardner theorised that each person may be born with strengths in some intelligences and weaknesses in others, each intelligence is seen as dynamic and can be developed. The intelligences are considered largely independent of each other, though they are typically used in concert.

Sternberg's (1985, 1997, 2000) 'triarchic' theory is a meta-theory of intelligence comprised of three subtheories relating to internal, external and experiential aspects of

[3] See also *Abilities and Aptitudes*, Snow (1994); and *Competence versus Performance*, Gelman (1994).
[4] Gardner later suggested there may be other intelligences including naturalistic, existential and spiritual (1999).

intelligence (Sternberg, 1985). He also theorised that the type of intelligence applied in solving problems can be analytic, practical or creative.

The internal subtheory, also called the componential subtheory, states that the cognitive processes of solving problems and accomplishing tasks involves meta-components, performance components, and knowledge acquisition components (Sternberg & Gardner, 1982). Meta-components recognise a problem exists, identify important features of the problem, allocate cognitive resources such as attention, select strategies for solving the problem, monitor sub-goals in reaching the solution, and evaluate the solution and related outcomes. Performance components enact meta-component directions. Knowledge acquisition components selectively encode information contained in a problem, which is combined with and compared to previously acquired information.

The external subtheory argues that individuals require intelligence in order to successfully adapt to, select and shape situations or environments.

The experiential subtheory hypothesises that intelligence involves the duel tasks of coping with novelty and automatising processes. For example, when learning to ride a bicycle, the actions required are initially completely unfamiliar. With practice, these actions will become automatic, requiring little or no deliberate cognition.

Sternberg suggested that analytic, creative and practical intelligence are used to solve different kinds of problems, or to solve problems in different ways. Analytic intelligence is applied in deconstructing aspects of a problem, sorting through relevant information and successfully applying performance components to find the solution. Practical intelligence is used in everyday problem solving; that is, the problems and tasks that do not require deep analysis. Practical intelligence involves the use of tacit knowledge, which comprises knowledge about how to accomplish everyday tasks, often

gleaned by individuals through experience rather than formally explained or taught (Sternberg, 2000). Creativity will be required when unique ideas are generated to solve the problem. Like Gardner's multiple intelligences, Sternberg's analytic, practical and creative intelligence are viewed as relatively independent of each other, though they function jointly in accomplishing goals.

It should be noted that 'creativity' is a concept as contentious as intelligence (Kaufman & Sternberg, 2010). In more recent articles, Sternberg tends to describe creativity as a complementary attribute for successful intelligence rather than a type of intelligence (e.g. Sternberg, 2003a, 2003b). Practical intelligence has more support as a distinct facility compared to analytic intelligence, largely because of its face validity. Specifically, it is difficult to accept that intelligence may be adequately indexed by responses to test items that appear to bear little relation to real-world problems (see section 1.2.6), whereas tests of practical intelligence relate directly to problems in context (Neisser, et al., 1996).

Social intelligence was first suggested by E. L. Thorndike as important for understanding and behaving appropriately toward other people; 'acting wisely in human relations' (Thorndike, 1920, p. 227). Emotional intelligence is the contemporary form of social intelligence (Mayer, Salovey, & Caruso, 2000), which combines individuals' sensitivity to emotion in other people, their ability to elicit desirable responses from others, and their ability to recognise and regulate their own emotions. Emotional intelligence is comparable to Gardner's intrapersonal and interpersonal intelligences (Matthews, Zeidner, & Roberts, 2005), and is also conceptually related to Sternberg's practical intelligence, because everyday problem solving often requires interaction with people (Lievens & Chan, 2010).

1.2.3 Models of intellectual development

Theories relating to intellectual development can be categorised roughly by the phenomena considered important: biological models , contextual models and systems models (Davidson & Downing, 2000). Sternberg (1990) argued that the different types of models reflect the researcher's metaphor for intelligence: biology, anthropology or computation. Biological models tend to frame intelligence as a function of genetic, physiological or neurological characteristics; contextual (anthropological) models suggest intelligence is socially acquired and expressed; computational or systems models see intelligence as complex interactions of various input resulting in highly diverse output.

E. L. Thorndike and Yerkes studied the biological bases of intelligence via comparative psychology (Thorndike, 1898; Yerkes, 1912, 1943a), arguing that differences in intelligence between humans and animals (especially non-human primates such as apes and chimps) were quantitative rather than qualitative given the shared genetic make up of species (Thorndike, 1901; Yerkes, 1943b; Yerkes & Learned, 1925). Comparative studies are not widely supported in contemporary intelligence research, and these researchers tend to be remembered mostly for their substantial contributions to human ability testing (Thorndike, 1903, 1904; R. M. Thorndike, 1994; von Mayrhauser, 1994; Yerkes, Bridges, & Hardwick, 1915).

Eysenck and Jensen also supported the idea that intelligence was biologically determined (Eysenck, 1971; Jensen, 1987). Eysenck studied the relationship of neurological functioning and intelligence, for example by correlating simple or choice reaction times with speed of cognitive functioning in tests of intelligence (P. Barrett,

Eysenck, & Lucking, 1986; Bates & Eysenck, 1993). With some evidence suggesting a link, he theorised that intelligence was related to neural efficiency (P. Barrett & Eysenck, 1992; Eysenck, 1982, 1986). Though empirical results in this area have been mixed, the field of study continues to demonstrate links between psychometric intelligence (see section 1.2.4) and neurophysiologic anatomy and function (e.g. Neubauer & Fink, 2009; Waiter, et al., 2009).

Jensen was influential in developing research relating to intelligence and information processing (Jensen, 1989, 1992; P. A. Vernon & Jensen, 1984). He is noted also for his controversial theories about the genetic bases of intelligence and ethnic group differences in test scores (Jensen, 1978; Kamin & Grant-Henry, 1987).

Vygotsky had an opposing viewpoint, and suggested that higher mental functioning originated not in the genes of humans but in their social environment (Vygotsky, 1962; Vygotsky, Rieber, & Hall, 1997). Contextual models of intellectual development like Vygotsky's suggest that intelligence is socially constructed. That is, individuals understand the meaning of intelligence and what constitutes intelligent behaviour through their socio-historical and ecological context (Ogbu & Stern, 2001). The idea that intelligence presents itself consistently and predictably across individuals and populations, measurable via testing, is a 'Western' viewpoint (Berry, 2001; Ortiz & Dynda, 2005)[5]. Extreme relativistic contextual models hypothesise that cognitive development is a product of one's daily experiences, meaning intelligence can not be defined in such a way as to apply universally to humans (Pervin, 2001). More commonly, however, contextual models suggest the environment is a major source of

[5] 'Western' is an adjective used to refer collectively to North America, Europe, Australia, New Zealand and (arguably) South Africa. Many cultural traditions and languages are alike across these nations, and they may be more similar to each other than to other nations due to common socio-historical influences. However, it is generally considered a misnomer to combine all the cultures making up these nations into one group called 'westerners'. The word is used in this discussion primarily as a reflection of terminology that appears in the reviewed literature. It should be thought of as a convenient descriptor of mostly common practice across a group of countries, though used most often to refer to North America, and not wholly representative of specific nations.

variability in intellectual development, as opposed heredity being the major source of variability, as in biological models (Berry, 2001).

One systems model taking into account the interaction of biology and environment is Piaget's theory regarding the growth of knowledge. Piaget (1947, 1951, 1961, 1980) argued that knowledge is developed as a product of a child interacting with its environment, in the same way that genes develop via adaptation to the environment. Potential for intelligence exists biologically, but knowledge can only be gained through interaction with the environment. Piaget's ideas were especially influential to the field of artificial intelligence, because computers need to be programmed not just with the process required to solve a problem, but with the information (or 'knowledge') of what aspects of the problem mean in order for an appropriate process to be selected (Bruchez-Hall & Gruber, 1994; Inhelder & Sinclair, 1994).

Influenced by Piaget, and similarly focusing on the growth of knowledge and interactions between internal and external factors, Ackerman and Beier (2005) argued that the actual volume and content of knowledge obtained by individuals is a function of their intelligence-as-process (learning ability), personality, interests and intelligence-as-knowledge (information they have learned; PPIK theory[6]). PPIK theory postulates that these individual differences interact to form trait 'complexes', motivating people to acquire different amounts of knowledge related to diverse topics (see also Ackerman, 1997; Kanfer & Ackerman, 2005).

Systems models of intellectual development can be broadly interpreted as the environment contributing different opportunities for learning, with individuals taking variable advantage of the opportunities (Lohman, 2001; Snow, 1996). Because of the complexities of biological and environmental input, output such as cognitive ability, personality, motivation and behaviour varies greatly between people. For example,

[6] Intelligence-as-**P**rocess, **P**ersonality, **I**nterests, intelligence-as-**K**nowledge.

personality and temperament may influence whether individuals take advantage of opportunities or not, which may influence the development of cognitive ability (see section 1.2.4). There is evidence that cognitive abilities are related to personality factors and temperament, but correlations are usually small, and inconsistent in both size and direction between studies (e.g. Ackerman & Heggestad, 1997; Strelau, Zawadzki, & Piotrowska, 2001; cf Wood & Englert, 2009). Cognitive ability, personality and motivation are therefore generally treated as independent (e.g. Kyllonen, 2002; Schmidt & Hunter, 1998). This conclusion may be incorrect, if the instruments are unreliable or interpretations of results are invalid. Scores of personality factors and temperament in particular are subject to alternative interpretations (Hofstee, 2001; Kline, 2001). In addition, relationships between personality and motivational factors and their influences on intellectual development may be idiosyncratic, which does not necessarily mean these characteristics are independent.

The precise process of intellectual development may remain elusive (Grigorenko, 2002), but there is evidence for the biological bases of cognitive function (Cattell, 1980; Jensen, 1987), as well as evidence that environmental factors affect intellectual development (Bellinger & Adams, 2001; Ogbu & Stern, 2001; Schaie & Zuo, 2001). Main effects and interactions between heredity and environment contribute to individual differences in cognitive ability (Dickens & Flynn, 2001; Petrill, 2002; Ramey, Ramey, & Lanzi, 2001)

1.2.4 The psychometric approach to intelligence

Up till now, this review has outlined theories of intelligence and models of intellectual development while referring to test results only as required. Many layperson and scholarly theories of intelligence, however, are derived from the history and current

practice of intelligence measurement (Embretson & McCollam, 2000; Sternberg & Kaufman, 1998). This psychometric approach to understanding intelligence is prominent in any research relating to intelligence or intellectual development, whether it is acceptance of the paradigm and related conclusions (e.g. Eysenck, 1988b; Jensen, 1987; Kline, 1991), or a model that has been developed as a challenge to the paradigm (e.g. Gardner, 1993; Sternberg, 1985). Boring provoked discussion with his oft-quoted statement, "Intelligence is simply what the tests of intelligence test" (Boring, 1923, p. 35). The history of 'psychometric intelligence' and the debate around acceptance of the testing paradigm for understanding intelligence discussed henceforth leads to the explanation for the contemporary use and meaning of the term 'cognitive ability'.

Galton is generally considered the pioneer of intelligence testing (Brody, 2000; Jensen, 2002; Wasserman & Tulsky, 2005). Galton's tests included mainly sensory acuity and discrimination. In the domain of auditory ability, for example, he might examine whether people could hear a musical note above a certain frequency, or whether they could distinguish different musical pitches. Galton used the terms 'mental ability', 'intelligence', 'eminence' and 'genius', though he never offered a formal definition of his terminology (Jensen, 1998). This may explain the apparent incongruity between his choice of physical abilities to test and the latent construct he was trying to measure. Galton's results were disappointing to him, as they did not reveal a relationship between abilities and achievements, nor evidence for the heritability of intelligence that he was expecting. This may have been due in part to lack of appropriately sophisticated statistical procedures at the time, as well as a lack of reliability in his measurement instruments (Jensen, 1998). Despite problems with both his theoretical and practical developments, Galton's work provided a basis for two of the most renowned intelligence researchers: Binet and Spearman.

Binet was one of the architects of perhaps the most influential test of intelligence to date, the Binet-Simon scale (Binet, Simon, & Kite, 1916a; Binet, Simon, & Town, 1912). Binet rejected Galton's notion that intelligence could be measured using only elementary cognitive tasks (such as sensory discrimination), and argued that tests of intelligence should require more complex problems to be solved. When commissioned to investigate the educational needs of children who appeared intellectually below average, Binet and Simon developed their scale to allow identification of diminished intellectual capability in children. The test contained items requiring different abilities to solve them (for example, vocabulary and arithmetic) and varying in difficulty. The basis of the difficulty levels was that more difficult items should be able to be answered by children of increasing ages. Therefore, if a child answered questions that were at a difficulty rating corresponding to a younger age, this child would be considered 'retarded' or 'feeble-minded', while a child correctly answering question at a level greater than their own age might be considered 'gifted' (Binet, Simon, & Drummond, 1914; Binet, Simon, & Kite, 1916b). The scores were based on the sum of correct responses, so two children could receive the same score even if they got different types of items correct (e.g. verbal compared to numeric).

It is ironic that an earlier attempt of Binet's to develop a test of intelligence was unsuccessful, because he believed that intelligence involved such complexity of operations that the number of test items required to develop a sufficiently comprehensive view of a person's intellectual functioning would be impracticable. In Binet's own view, his test may have indicated at best a small sample of a child's intellectual capabilities.

Binet and Simon's tests were the original paper-and-pencil intelligence tests, though the Binet-Simon scale was not converted to an intelligence quotient (IQ).

Terman translated the Binet-Simon scale at Stanford University, creating the now famous Stanford-Binet intelligence test (Terman, 1911, 1916; Terman, et al., 1917). Stern (1949; Stern & Spoerl, 1938; Stern & Whipple, 1914) created the IQ, which was the test-taker's mental age (score on the test) divided by his or her chronological age. Later, IQ scores were standardised so that for any given reference group (e.g. 6 year olds, university students), the mean score would be 100, with a standard deviation of 10 or 15 (Wasserman & Tulsky, 2005). IQ appeared to become a synonym for intelligence in some quarters (e.g. Jensen, 1969).

Spearman (1904, 1927) remains hugely influential in the world of psychology; not only for his contributions to the field of intelligence measurement, but also because of the data analysis techniques he developed in order to study intelligence. Specifically, he refined correlation techniques and developed factor analysis, statistical techniques that are hugely important in contemporary psychometrics (Carroll, 1954, 1978; Cattell, 1958; Spicer, 2005).

Spearman (1904) conducted studies on sound, light and weight discrimination, and re-analysed previous literature on 'mental tests', including Galton's and Binet's. The types of tasks he was comparing included measures of sensory acuity and discrimination, motor ability, reaction time, attention and memory. Having attenuated correlations by removing the influence of unreliable measurement, he discovered that there were positive correlations between all types of mental tasks. Spearman (1927) then applied his factor analytic technique and discovered that the variance in different mental test scores could largely be accounted for by a single factor, which he called g, short for general mental ability. He theorised that test scores were accounted for mostly by g, and any additional unexplained variance would be attributable to s: a specific ability relevant only to that test item, which was initially considered independent of s

for any other test item. This conceptualisation of intelligence is known as Spearman's two-factor theory[7].

Spearman's 1904 article referred specifically to "general intelligence", but he did not come to equate *g* with intelligence (1904; c.f. 1927). He suggested that they may be related in some way, but acknowledged that no unified theory of intelligence or its components existed. He said of the *g* factor: "that which this magnitude measures has not been defined by declaring what it is like, but only by pointing out where it can be found" (p. 75). Using potential energy in physics as a metaphor for *g*, Spearman conceptualised *g* as mental energy, the same way that a physicist may think of kinetic energy. After further developments in factor analysis, and much debate with Thurstone (1923, 1933) in particular, Spearman acknowledged that there were also group factors to be found among similar problem types (Spearman & Jones, 1950). While *g* consistently accounted for some proportion of the variance in almost all tasks, group factors were found to contribute to a reduction in score variance beyond *g* on similar tasks. Through these discoveries, a hierarchical model of cognitive abilities was developed (also known as a factor-analytic model), with *g* at the apex, grouped abilities below, and more specific abilities making up each group.

The *g* factor theory was not unilaterally accepted. One alternative model was the theory of fluid and crystallised intelligence (*Gf-Gc* theory), suggested by Cattell (1935, 1940, 1958, 1961, 1963)[8]. Though Cattell did not reject the existence of *g,* he suggested that the apex of the hierarchy would be better described as broadly comprising two categories (*Gf* and *Gc*), so intimately entwined that they were difficult to distinguish from each other, resulting in the appearance of a single factor (Cattell, 1963, p. 2). Fluid

[7] This label could be interpreted as a misnomer, given that *s* could be any multitude of factors. Carroll suggested it was better described as Spearman's "one-general-factor theory" (Carroll, 1993, p. 53).
[8] Another alternative model was Guilford's structure of intellect theory. This theory was influential, but lacked empirical support. For parsimony it has therefore been excluded from the main discussion, but a summary appears as appendix one.

intelligence (*Gf*) was conceptualised as the innate portion of intelligence, the ability to learn and acquire new knowledge. Crystallised intelligence (*Gc*) was the manifestation and application of learned material, bound by the experiences of individuals and strongly influenced by their cultural environment and education.

The *Gf-Gc* model was updated over time as Cattell and Horn (1966) discovered evidence for more group factors, adopting Thurstone's term "primary mental abilities" (Thurstone, 1938). Later descriptions of the model (summarised in Horn & Blankson, 2005) had memory factors including short-term apprehension and retrieval also called short-term or working memory, and fluency of retrieval from long-term storage also called long-term memory. Information perception and processing factors included visual processing, auditory processing and processing speed. There was also evidence of a special numeric factor representing a unique blend of *Gf* and *Gc*, called quantitative knowledge. Cattell and Horn's *Gf-Gc* theory was not necessarily intended to completely describe all possible forms of cognitive functioning, but rather was an effort to describe in particular the structure of interrelated abilities as typically applied in tests.

Given evidence of the heritability of *g* and *Gf* factors (Cattell, 1980; Eysenck, 1988a; Spearman, 1927) and also the evidence of environmental influences on intellectual development and functioning (Flynn, 1987; Irvine & Berry, 1988; P. E. Vernon, 1969); Hebb (1942) suggested distinguishing between intelligence A and intelligence B. Intelligence A denotes the intelligence genotype: the innate, biological basis of intelligence. Intelligence B is the intelligence phenotype: individuals' 'ultimate' intelligence level, influenced not only by their genes but also by their upbringing. Intelligence B refers not only to intellectual development, but also the cultural factors affecting whether individuals are considered intelligent within their milieux. Given that intelligence A is unobservable and intelligence B requires culture-specific observations

and interpretations, Vernon (1969) extended this taxonomy to include intelligence C, 'psychometric intelligence', which is cognitive ability as it is manifested in test performance. Individuals' test scores are unlikely to comprehensively represent their intelligence in the broadest conceptualisations of the term, yet can still reveal something about their cognitive processing (Carroll, 1943, 1987).

Jensen (1998) argued that the word intelligence was interpreted in so many alternative ways that it was best to eliminate the word 'intelligence' entirely from scientific repertoire and concentrate on the results of tests, especially *g*. He suggested 'mental ability' or 'cognitive ability' instead, with *g* denoting general mental ability.

Though compromises such as Vernon's 'intelligence C' or Jensen's 'mental ability' do not eliminate ambiguity or difficulties of definition, the use of the term 'cognitive ability' is adequate for describing a characteristic of individuals that influences their test scores. Practicable tests will only be capable of sampling limited domains of cognitive functioning, and are entrenched in a particular socio-historic context. Though there is hope that there may one day be a common definition of intelligence, it may not be discoverable via existing psychometric means. For the remainder of this discussion, therefore, cognitive abilities are assumed to be implicated in the intra-individual aspect of intelligence, but are not assumed to comprehensively describe a person's mental facilities, capabilities or achievement potential.

1.2.5 Carroll's three-stratum theory

Carroll (1993) offered the clearest definition of cognitive abilities, publishing it along with perhaps the most comprehensive meta-analysis ever undertaken to examine the structure of human cognitive abilities (Alfonso, Flanagan, & Radwan, 2005; Jensen, 2004; Lubinski, 2004; McGrew, 2005). Carroll was an active scholar in the field of

cognitive psychology, as well as contributing much to the understanding and development of the field of psycholinguistics (Carroll, 1944, 1953, 1964; Carroll & Burke, 1965). Combining his vast research on cognition, language and factor analysis, he developed tests of language proficiency and examined how cognitive processes were implicated in the acquisition of language (Carroll, 1943, 1954, 1973, 1987; Carroll & Maxwell, 1979). Carroll was therefore highly influential in expanding the understanding of cognitive processes, though in the field of psychometric intelligence his 1993 opus is considered his crowning achievement.

Carroll defined having an 'ability' as: having a certain probability of successful performance (psychometrically, above 50%), given favourable conditions, on a defined class of tasks (1993, p. 8)[9]. To be defined as 'cognitive' the tasks must require primarily "processing of mental information" (p. 10) to be performed successfully. 'Task' is defined as an activity with some set objective or end result, the criteria for which are known to the individual performing the activity. Carroll also makes an important distinction between level and speed. Level relates to the difficulty of the tasks to be performed: a greater level of ability is required in order to successfully perform these tasks. Speed has long been implicated as a facet of intelligence (Berger, 1982; Danthiir, Roberts, Schulze, & Wilhelm, 2005), but its actual importance is undefined: is one person 'less able' than a second person if they both have the same probability of performing successfully, but the first requires more time to complete the task? The answer would appear to be no because the definition does not include speed of performance, and yet a faster and equally as correct response would appear to suggest superior information processing. Carroll hypothesised, therefore, that abilities may have

[9] Carroll's (1993) book opens with 13 pages dedicated to defining cognitive abilities, so the précis here may include some semantic ambiguity that Carroll specifically sought to avoid.

distinct speed and level factors, and in his analyses he noted whether tests were time-limited or not.

As a result of his reanalysis of 477 datasets assessing performance on a range of cognitive ability tests, Carroll developed the three-stratum theory of cognitive abilities. The strata related specifically to the generality of the ability factor across tasks. Carroll suggested thinking of the three strata as representing narrow, broad and general cognitive abilities (stratum I, II and III, respectively).

Like Spearman, Carroll discovered a general ability factor accounting for some proportion of the variance in a broad range of tasks. He designated this *g* factor as stratum III, the apex of the hierarchy. At stratum II were 7-9 factors, ordered roughly by their strength of association with the general factor, including (in order) fluid intelligence, crystallised intelligence, general memory ability, broad visual perception, broad auditory perception, broad retrieval ability and broad cognitive speediness. The other 2 factors were special cases of other second-stratum abilities: one was a combination of fluid and crystallised intelligence, distinct nonetheless from the general factor; and the other factor was a reaction-time speed factor related to rapidity of both decision-making and psychomotor response, which tended to be subsumed under broad cognitive speediness[10]. Stratum I was made up of specific abilities, yet each ability type was still applicable to a range of tasks. Some stratum I factors and their corresponding stratum II factors will now be described in more detail.

Fluid intelligence, named after Cattell and Horn's theory, denoted reasoning abilities that were not indicative of learning, and included stratum I abilities such as general sequential reasoning, induction and deduction (Spearman's 1927 'eduction' of relations). Crystallised intelligence contained those abilities that were more dependent

[10] Because of the similarity in factors with Cattell and Horn's model, the three-stratum theory and *Gf-Gc* models are collectively known as the Carroll-Horn-Cattell (CHC) theory of cognitive abilities (Alfonso, et al., 2005; McGrew, 1997, 2005).

on education and upbringing such as language development, reading comprehension, communication ability, written and oral production and fluency. General memory and learning included memory span and free recall ability. Stratum I abilities such as spatial reasoning, perceptual speed, visualisation and illusion perception were subsumed under stratum II's broad visual processing, while hearing threshold, pitch discrimination, and rhythm maintenance were subsumed under broad auditory perception. Broad retrieval ability could also be called idea production, and included creativity factors as well as fluency of expression, words and associations. Cognitive speediness included test-taking speed and numerical facility, with processing speed including reaction time (see Carroll, 1993, p. 626, for a diagram and additional stratum I abilities).

Carroll's three-stratum theory offered a general outline (rather than a rigid structure) for the way in which cognitive abilities are organised and function. This means that stratum I abilities were often applicable to more than one group factor at stratum II. Additionally, the strength of relations between the factors was not expected to be constant across individuals and tests. Finally, the factors were not rigidly defined within their respective strata: Carroll stated that the strata were an indication of the generality of the ability across tasks, and there may be intermediate strata between them (pp. 635-636). Carroll also noted that the entire analysis was limited to those abilities tested in the datasets that were analysed. For instance, there is no particular reason why olfactory (smell), gustatory (taste) or somatosensory (touch or 'feeling') abilities should not appear at the level II stratum with broad visual and auditory abilities, except that these senses do not tend to be tested.

To summarise, *g* is a factor of cognitive ability that explains a large proportion of variance in test scores. Other closely related factors include a capacity for learning (fluid intelligence, *Gf*), and the storage and use of learned material or 'acculturated

knowledge' (crystallised intelligence, Gc, plus broad memory and retrieval factors; Carroll, 1993; Cattell, 1986; Horn & Noll, 1997). *G*, *Gf* and *Gc* factors are arguably the best indicators of broad cognitive function available at the present time (Gottfredson, 2002a). While *g* is not generally equated with intelligence in contemporary literature, there is some debate around continued focus on *g* as a predictor for a multitude of life accomplishments.

1.2.6 Debate around psychometric g

The debate relates to the generality of the *g* factor (Sternberg & Grigorenko, 2002). So-called *g*-theorists argue that, in applied settings, *g* is all that is required for a highly predictive index of a multitude of life outcomes including education attainment, job success, health and overall life expectancy, with other individual differences adding little meaningful prediction (Gottfredson, 2002b; Jensen, 1998; Kuncel, Hezlett, & Ones, 2004; Kuncel, et al., 2010). Though there is some empirical evidence for this point of view (see section 1.4.1) *g*-theorists tend to argue that individual differences in *g* are the cause of the variability in outcomes. This aspect of the theory is open to interpretation. Critics of *g*-theory do not deny the existence of *g*, but rather question whether it is the cause of various life outcomes, and also question whether other individual differences might be at least as important as *g* (e.g. Sternberg & Wagner, 1993).

One *g*-theory criticism argues that test problems are contrived, and so different from real-world problems that the abilities and other characteristics required to solve them are likely to be qualitatively different. Test problems "tend to (a) have been formulated by other people, (b) be clearly defined, (c) come with all the information needed to solve them, (d) have only a single right answer, which can be reached only by

a single method, (e) be disembedded from ordinary experience, and (f) have little or no intrinsic interest. Practical problems, in contrast, tend to (a) require problem recognition and formulation, (b) be poorly defined, (c) require information seeking, (d) have various acceptable solutions, (e) be embedded in and require prior everyday experience, and (f) require motivation and personal involvement" (Neisser, et al., 1996, p. 79).

Another arguments is that the cognitive abilities typically tested, such as logical reasoning, are only valued in western contexts and that the very ideas of 'testing' and 'test environments' are also western (Berg & Klaczynski, 2002). This 'testing bias' in society leads to self-fulfilling prophecies in predicting broad life outcomes: that is, tested cognitive abilities are valued so are emphasised in educational assessment, and are therefore predictive of educational attainment. Furthermore, educational attainment and cognitive abilities are selected for and rewarded by organisations, so become predictive of occupational attainment. Educational and occupational attainment are predictive of other outcomes like nutrition, wealth and access to healthcare, which are all predictive of overall life expectancy (Sternberg, 2001). Therefore, the influence of *g* is compounded by western society's emphasis on it, and *g* is not as predictive of broad life outcomes in other societies (Irvine & Berry, 1988).

Another criticism rejects the idea that more specific abilities add little useful contribution. One reason that there are only small increments is because the 'specific abilities' are limited to those tested for (Gardner, 1993; Stankov, 2002), and test problems are very different from real-world problems, so abilities other than those underlying test scores might contribute a great deal to life accomplishments.

Another explanation for small increments is that they are statistically underestimated. In other words, interpretation and reification of the *g* factor may have been influenced by the language and methodology of statistics. *G* is often interpreted as

'dominating' the lower-order abilities[11]. However, the variance 'accounted for' by *g* is actually shared with the group factors that contribute to it, and isolating *g* attributes all the shared variance to *g* alone (Gustafsson, 2001; van der Maas, et al., 2006). This practice does result in lower-order factors providing only small increments, because *g* is assumed to be 'controlling' the shared portion of variance, when stratum II factors may actually contribute a non-trivial portion (Lang, et al., 2010). Lang and colleagues (2010) noted that conventional statistical procedures did not provide the means of dividing the shared variance between factors in predictive analyses. Utilising an alternative procedure to allow such division, they found that *g* was not always the best predictor. Van der Maas and colleagues (2006) also demonstrated how complex relationships and interactions between cognitive abilities might result in a *g* factor even if it was not a distinct ability.

Though *g* may be a useful predictor, its substance not only appears to be variable across tests, individuals and generations (Flynn, 1987), but its nature also remains, to some extent, unspecified (R. L. Thorndike, 1994). The focus on unspecified 'general' cognitive ability may limit theoretical and practical developments for explaining which cognitive abilities influence which behaviours and how. Therefore, the practice advocated here is to be specific about which abilities are tested, which firstly recognises that a test's content and context is inevitably limited with regard to the construct domain being sampled (and therefore conclusions about individuals are also limited to the range of abilities tested). Secondly, specificity may allow more meaningful theoretical explanation of how tested abilities relate to various achievements.

[11] 'Lower-order' implies 'of lesser importance', though the phrase only relates to the order in which the factors were extracted.

1.2.7 Measuring cognitive abilities

In order to understand measurement models of cognitive ability, first a brief explanation of psychometric theory may be helpful. Psychometric theory hypothesises that individuals may have different levels of latent (hidden) traits such as cognitive ability, and levels can be described by assigning sequential, monotonically increasing values. An individual's trait level can be estimated by analysing observable behaviour (Borsboom, 2005; Lord & Novick, 1968). For example, cognitive abilities are latent traits that are hypothesised to be responsible for correct responses to test items (Lord & Novick, 1968). Therefore, properties of test items, item responses and overall test scores can assist in the estimation of an individual's cognitive ability.

There are currently two predominant measurement models used for estimating cognitive abilities (Embretson & Reise, 2000)[12]. The technique used in one model, classical test theory (CTT), is to analyse item properties during test construction, and estimate cognitive ability level by comparing an individual's final test score with scores of a representative reference group (Cronbach, 1990; Lord & Novick, 1968)[13]. The other model, item response theory (IRT), estimates individuals' most likely cognitive ability level from their responses to single items (Hambleton, Swaminathan, & Rogers, 1991; van der Linden & Hambleton, 1997). Each item has a certain probability of being answered correctly by individuals with different cognitive ability levels. The probability of a correct response for a given ability level can be calculated based on item properties.

---

[12] The empirical investigation undertaken here uses multiple choice tests with only one correct answer, the test form most often utilised in measuring cognitive abilities due to ease of standardising them (Cronbach, 1990). There are other test forms, such as essay questions, that may include a 'level of correctness' in the response e.g. A, B or C grades are all 'correct' answers because they are pass grades, but an A-grade response is a better answer than a C-grade response. Including these and other types of tests into this discussion requires a level of complexity and detail that is not relevant to the empirical study so has been omitted. Further detail on characteristics and applications of multiple types of psychological and educational tests can be found in Lord and Novick (1968), Cronbach (1990), van der Linden and Hambleton (1997), and Brennan (2006),

[13] Representative meaning the individual and the group have similar, comparable characteristics.

One of the problems of CTT is that estimation of item properties in test construction (item analysis) is heavily dependent on pre-testing groups of individuals from the population (Hambleton, et al., 1991). Item difficulty is the proportion of the group that correctly answers the item, and an adequate test should have a range of item difficulties in order to test a range of ability levels (Lord & Novick, 1968). Item discrimination is how well individuals' responses to each item correlate with their overall scores. If an item-total correlation is very low, this item does not influence the test score, thus does not discriminate well between individuals receiving different scores (Lord & Novick, 1968). This can happen if an item is very easy or difficult and is answered correctly or incorrectly by many examinees, because the item response will not significantly relate to the final test score. The same can happen for items that are poorly written, for example if they have more than one correct answer, because individuals might respond correctly but be scored as incorrect, and this scoring error would reduce the correlation between the item response and the test score. Items are included and excluded from tests on the basis of these analyses.

Overall test scores are usually unweighted and summative, meaning the difficulty and discrimination of each item does not contribute to the test score (Embretson, 2010; Lord & Novick, 1968). This means that it is possible for an individual to get more difficult items correct compared to another individual, but if they get the same total number of items correct they receive the same score.

The other weakness of CTT is the comparison of an individual's test score to a reference group. Similarly to the item analyses, the result is highly dependent on the reference group, and the representativeness of a group may be highly subjective (Shultz & Whitney, 2005).

Item response theory (IRT) does not have CTT's dependence on representative reference groups. IRT focuses on the test items and making strong assumptions about the relationship of individuals' ability levels to correct or incorrect responses (Drasgow & Hulin, 1990; Embretson & Reise, 2000; Hambleton & Bejar, 1983; Hambleton, et al., 1991; Lord & Novick, 1968; van der Linden & Hambleton, 1997; Yen & Fitzpatrick, 2006)[14]. IRT's main assumption is that the probability of a correct response to an item is dependent on the combined function of the item's difficulty and the individual's ability (expressed as theta, θ): if the individual's ability is greater than the item difficulty, the probability of a correct response will be greater than .5 (Rasch, 1960)[15]. The response function (item characteristic curve; ICC) appears graphically as a normal ogive (a cumulative normal probability curve) with the probability of a correct response given an ability level [P(θ)] on the Y axis, and ability level [θ] on the x axis (see figure 1.2).

Taking into account just the item difficulty parameter is known as the Rasch model (Rasch, 1960), but there are also two- and three-parameter logistic models, that allow the shape of the curve to change (Yen & Fitzpatrick, 2006). The second parameter taken into account is how well the item discriminates between differing abilities. This is most easily described as how steep the ICC is: the steeper the curve, the better the item is discriminating between individuals with different ability levels. The third parameter is required in multiple choice tests and is called the lower asymptote or pseudo-guessing parameter. In multiple choice tests it is possible to respond correctly by selecting an answer at random, so the minimum probability of a correct response at any ability level will be above zero.

[14] IRT is generally preferred as a measurement model, but is not without criticism (e.g. Hutchinson, 1991). For both IRT and CTT, more detailed analyses of problems with each model are too numerous and complex to be addressed.

[15] This accords with Carroll's definition of ability (section 1.2.4)

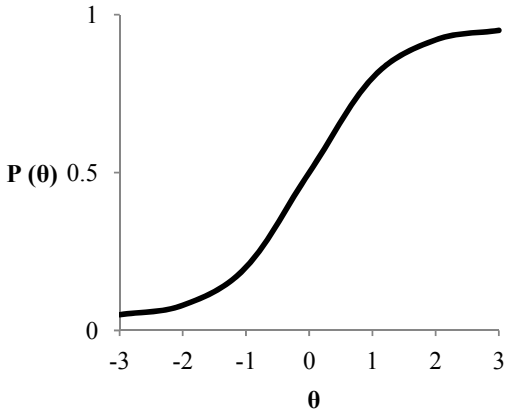


Figure 1.2: Item characteristic curve with item difficulty set at 0: When θ = 0, probability of a correct response is 50%

As stated, IRT assumes the only variable affecting responses is the examinee's ability level. Two specific assumptions are therefore that item parameters are invariant, and items are locally independent. Item parameter invariance means that the parameters are the same for every individual on every testing occasion. Local independence means a response to one item is not dependent on a response to any other item. As a result, the logistic function can estimate an individual's most likely ability level over any group of items, given the known item parameters and the person's correct/incorrect response set[16].

IRT allows adaptive testing, tailoring item selection for each examinee (Drasgow & Hulin, 1990). When test item parameters are known, a computer can select the next most appropriate test item using an algorithm based on the examinee's responses to previous items and known information about other possible items. Put very simply, if an individual answers a question incorrectly, the computer will select an easier item (an item with a lower difficulty parameter). If the individual then guesses correctly, a harder item will be presented. Using the known item parameters, the ICC and the examinee's responses, each item is selected so that the response will provide the greatest amount of

[16]'Most likely ability level' is written here for simplicity. When an individual gets all items correct or incorrect, the standard error for the most likely ability level is plus or minus infinity. One method of correcting for this is taking into account the expected distribution of ability scores in the population, a normal distribution (Embretson & Reise, 2000).

additional information about the examinee's ability (Drasgow, Leucht, & Bennett, 2006).

Arguably, IRT-based tests may have better utility as cognitive ability level estimates can be made with greater precision using fewer test items as a result of adaptive testing, and examinees may prefer adaptive tests, because the items will be neither too simple nor too difficult (see Schleicher, Venkataramani, Morgeson, & Campion, 2006). Fundamentally, however, whether a test is CTT- or IRT-based, the construct being measured is still cognitive ability. Research thus far has not shown significant differences in ability level estimates when comparing IRT- and CTT-based test scores (Mead & Drasgow, 1993), supporting the concurrent validity of the measurement models.

Key summary points from section 1.2: (1) intelligence is not defined consistently, (2) there is evidence for both genetic and environmental influences on intellectual development and knowledge, (3) there is a general cognitive ability factor, g, that accounts for some proportion of variance in 'intelligence' tests, meaning all cognitive abilities are likely to correlate positively, (4) the generality of g is restricted to tested abilities, and its precise nature and substance are unspecified, so research into specific abilities may be of theoretical and practical benefit, and (5) there are two primary measurement models for testing cognitive abilities, CTT and IRT, and though they differ in principles of test construction, cognitive ability estimates are the same regardless of test type.

1.3 Job performance

Having outlined key research relating to cognitive ability and its measurement, job performance is the next topic to examine. First, components that constitute a range of performance definitions are outlined, and then job performance is defined with reference to these components. Each component relates to different research themes, providing a framework for the review. Like cognitive ability, much of the understanding of job performance is a result of measuring it, and this review analyses how aspects of measurements relate to different components of the performance definition. Ultimately it will be clear what is meant by job performance, and how the components can be conceptualised and operationalised for diverse jobs.

1.3.1 Defining job performance

Similar to the term 'ability', the precise definition of 'performance' is often overlooked (Campbell, 1990). In the words of Lebas and Euske (2002, p. 67), "performance is one of those 'suitcase words' in which everyone places the concepts that suit them, letting the context take care of the definition". Campbell and colleagues were less circumspect in their criticism of the lack of definition: "the word *performance* is misused and exploited to the extreme in society at large, and is frequently butchered beyond recognition in psychology" (Campbell, McCloy, Oppler, & Sager, 1993, p. 35, italics in original).

'Performance' can describe diverse behaviours and outcomes. It is used to describe ability test scores, it is associated with artistic exhibitions such as plays and concerts (E. Bell, 2008), athletic abilities are displayed in sporting performance, and 'job performance' is some quality relating to employees at work (Campbell, 1990). With such a vast range of applications, how can performance be precisely defined? Analysis

of arts, sports, organisational and job performance definitions resulted in three (often implicit) components for any definition of performance: a judgement, a purpose and comprising processes, products or both (E. Bell, 2008; Hill, 2001; Motowidlo, 2003; Neely, 2002).

The definition of job performance is employees' actions and/or the attendant outcomes that impact upon an employing organisation's goals (cf Campbell, et al., 1993; Murphy & Cleveland, 1995; Roe, 1999; Sonnentag & Frese, 2005; Taylor, 2003; Viswesvaran & Ones, 2000)[17]. Examining the definition's components, the judgement belongs to an observer who evaluates the impact or relevance of a process or product to the organisation's goals. The purpose is the organisation's goals. The processes and products are employees' actions and the attendant outcomes, respectively.

The first component of performance is an inherent judgement. This means that there must be a potential judge such as an audience or supervisor (E. Bell, 2008; Lebas & Euske, 2002; Motowidlo, Borman, & Schmit, 1997; D. Smith & Bar-Eli, 2007; Sonnentag & Frese, 2005). The judge of job performance could be employees' supervisors, customers, direct reports, or employees themselves (Dalessio, 1998; Farr & Newman, 2001; Fletcher & Baldry, 1999). The rise of empirical work investigating multi-source feedback acknowledges that different judges have different views on how well employees do their jobs (Hedge, Borman, & Birkeland, 2001). One implication of the inherent judgement is that performance may not only be based on the actions of the performer, but is also a function of characteristics of the judge. This issue will be discussed in more detail in section 1.3.3, as it relates in particular to ratings (quantitative judgements) of the value of performance.

[17] Employee 'action' is used in preference to employee 'behaviour' to reflect purpose and direction, but the terms are semantically interchangeable.

The second component of performance is that agency and purpose are implied. That is, individuals know they are performing (Sackett, Zedeck, & Fogli, 1988), and their actions occur in the context of known criteria (E. Bell, 2008; Campbell, 1990; Lebas & Euske, 2002). 'Known criteria' does not necessarily mean well-specified or appropriate criteria, but there is some sense of a goal. A job has a purpose relevant to the organisation's goals, which employees should be aware of and direct actions towards. Criterion relevance is the area of job performance research most closely reflecting the purpose component, and will be discussed further in section 1.3.3 as it relates to measurement of job performance.

The third component is that performance can be defined both as a process and a product. In a rugby game, for example, the processes of game play constitute performance (ball handling skills, speed of running, accuracy of kicks), but performance can also be defined as the product: how many points the team scored and whether the team won or lost. There is some debate about whether job performance is best defined in terms of processes or products. In work contexts there are frequently environmental factors that limit generalisability from employee actions to outcomes. For example, total sales in a week (product) can reflect a salesperson's selling activity (process), how many customers were in the store that week (environmental factor), though the product is likely to result from an interaction of the processes and the environment (Murphy & Cleveland, 1995). As a result, it is unclear which aspect is most appropriate for judging job performance.

Researchers supporting the 'performance-as-product' approach state that performance should be defined in terms of desirable outcomes when environmental factors are favourable or the employee can control them (Bernardin & Beatty, 1984; Roe, 1999; Viswesvaran & Ones, 2000). Researchers with a 'performance-as-process'

point of view argue that job performance should be based on employees' actions, and results or outcomes should be treated distinctly as effectiveness rather than performance (Campbell, 1990; Campbell, et al., 1993; Motowidlo, et al., 1997; Murphy & Cleveland, 1995; Sonnentag & Frese, 2005).

A pragmatic approach combining performance-as-processes and performance-as-products may be best with regard to job performance. As demonstrated above, there are problems focusing solely on outcomes (performance products) if factors outside of an employee's control affect the number or desirability of outcomes (Murphy & Cleveland, 1995). In such instances, outcomes are not reflective of employees' actions, which means the outcomes cannot accurately be defined as performance products. Also, some employee actions do not directly influence performance products, but are still beneficial for achieving organisational goals (for example, contextual performance, see section 1.3.2). Conversely, organisations are typically more interested in the performance products than the processes used to achieve them (Pulakos & O'Leary, 2010), because specifiable job outcomes are usually the subgoals necessary for the organisation to achieve its main goals. Focusing solely on processes also diminishes recognition of the outcomes that are within the control of employees (Viswesvaran & Ones, 2005). Binning and Barrett (1989) argued that performance products should be described with specific reference to the processes that are necessary. While this approach merges process and product aspects of performance, some outcomes can be achieved utilising different processes that are equally advantageous[18]. In addition, it is unrealistic for supervisors to observe all employees' performance processes, meaning performance products may be required as indicators of performance. This practical issue relates to measuring job performance, so will be revisited in section 1.3.3. In theory and practice,

[18] For example, a rugby team can score 21 points by converting 3 tries, kicking 7 penalties, or scoring 3 non-converted tries and kicking 2 penalties.

job performance definitions can include both products and processes while heeding the attribution of these to a particular employee. This reconciliation recognises that both performance processes and products can influence the accomplishment of organisational goals and subgoals.

Theoretical developments in job performance research tend to relate to processes and products, while judgement and purpose components are more reminiscent of practical issues. Because this section so far has analysed theory of job performance components, it is more appropriate to review research of processes and products first. Discussion of more practical issues, specifically measuring job performance, is reserved for section 1.3.3.

A growing body of literature is developing knowledge about and specificity of performance products and processes (Borman & Motowidlo, 1993; Campbell, 1990; Viswesvaran, Ones, & Schmidt, 1996). Such analyses are not easily conducted because organisations and jobs are hugely diverse systems, comprised of a multitude of various tasks, people and environments. Their diversity limits the ability of a researcher or practitioner to generalise job performance characteristics to many jobs. In particular, performance products of different jobs do not tend to resemble each other, limiting meaningful comparison; for example, there is little meaning in comparing real estate agents' sales volumes and doctors' success rates for curing patients. Performance products therefore cannot realistically be generalised except as job-related outcomes. Researchers therefore focus on broad dimensions of job performance, the broadest of which can generalise to all jobs. The more specific performance dimensions, usually groupings of similar behaviours, are applicable to a greater or lesser extent between jobs. When applied in practice, researchers and practitioners must always keep in mind

how relevant each dimension is to the jobs under investigation. The next section discusses these dimensions of job performance.

1.3.2 Dimensions of job performance

Job performance for all jobs can broadly divided into either typical and maximum performance (Sackett, et al., 1988), or into task and contextual performance (Borman & Motowidlo, 1993). Typical and maximum performance explain fluctuations in individual job performance, while task and contextual performance relate to characteristics of job performance itself. The latter dimensions are therefore of particular interest, though the distinction of typical and maximum performance is pertinent.

Typical performance is categorised as *will-do* performance (Deadrick & Gardner, 2008; DuBois, Sackett, Zedeck, & Fogli, 1993). In daily work activities, individuals are likely to exert fluctuating levels of effort, and job tasks will not always require their full range of abilities (Sackett, et al., 1988). Maximum performance is categorised as *can-do* performance (DuBois, et al., 1993) and is the upper limit of how well employees are capable of performing when they exert optimal effort at the highest levels of their abilities (Sackett, et al., 1988). Distinguishing between typical and maximum performance can help when choosing predictors, because of individual differences expected to influence each dimension (Cronbach, 1990). For example, individuals' cognitive ability and integrity test scores may be good predictors of what an employee can do (Ones & Viswesvaran, 2007), while personality may indicate what an employee will do (Cronbach, 1990; Marcus, Goffin, Johnston, & Rothstein, 2007). The typical-maximum distinction is also important when considering ways in which performance is dynamic (Deadrick & Gardner, 2008), which will be discussed in section 1.4.3.

Contextual performance was a concept developed by Borman and Motowidlo (1993) to complement task performance. They argued that theories and measurements of job performance tended to focus on processes and products that contributed to the technical core goals of an organisation (task performance), while diminishing the potentially large effects of other more discretionary behaviours on the working environment, such as being optimistic and persevering in the face of adversity (contextual performance). Contextual performance elements include organisational citizenship behaviours (OCBs), such as loyalty and promoting organisational goals (Organ & Paine, 1999); and prosocial behaviours, including cooperating with or helping others (Borman & Motowidlo, 1997).

As alluded to in section 1.3.1, contextual performance does not necessarily have a straightforward effect on job-related or organisational goals and cannot typically be specified in relation to individual performance products. Task performance by definition is expected to contribute directly to organisational goals, and some jobs could be described entirely in terms of performance products without reference to processes (for example, ‘sales performance’).

This demarcation has been an important contribution to job performance theory and research, because task and contextual performance have been shown to contribute to overall performance judgements, rewards and overall organisational effectiveness in relatively equal amounts and (to some extent) independently of each other (Motowidlo & Van Scotter, 1994; van Scotter, Motowidlo, & Cross, 2000; Whiting, Podsakoff, & Pierce, 2008). Though these broad dimensions of job performance can be assumed to apply across all jobs, it is likely that the value of each may differ depending on the job. For example, contextual performance may be of little import compared to task performance for a commercial pilot, except as far as rapport with the co-pilot or crew

assists with safely and successfully flying a plane; whereas the cabin attendants' job performance may benefit greatly from having cheerful and helpful colleagues.

There are a number of performance taxonomies that take into account more specific dimensions of job performance (e.g. Campbell, 1990; Hunt, 1996). These are groupings of behaviours, thus specifying some performance processes. Though too numerous to specify here, Borman, Bryant and Dorio (2010) summarised the most commonly cited taxonomies into the following eight dimensions: (1) productivity and proficiency; (2) problem solving; (3) organising and planning; (4) leadership and supervision; (5) information processing; (6) communicating and interacting; (7) useful personal qualities (e.g. adaptability, persistence); and (8) counterproductive behaviour. The first six dimensions could relate to either task or contextual performance depending on the demands of the job (Motowidlo, 2003). The sixth dimension, communicating and interacting, is likely to be implicated in both task and contextual performance, though again, how and to what extent will depend on the task demands of the job. In practice, therefore, task and contextual performance are likely to relate to each other, to the extent that particular classes of behaviours apply to both.

The final two dimensions, counterproductive behaviour and adaptability as a useful personal quality, may influence job performance rather than constituting it. Perhaps because they do not quite fit, these dimensions are increasingly the foci of distinct branches of research. A summary of this literature examines how they relate to job performance.

Counterproductive behaviour is deliberate harm to an organisation or its members including stealing or abusive treatment of colleagues or customers, and can include withdrawal behaviours such as lateness or absence (Rotundo & Spector, 2010; Sackett & DeVore, 2002). Campbell's (1990) version of this dimension was "maintaining

personal discipline" (p. 709), and was defined as the avoidance of rule-breaking. Withdrawal behaviours could be described as a failure to perform at all (for some period of time). The definition of job performance refers to 'employees' actions and/or the attendant outcomes', so avoiding counterproductive behaviours and failing to perform do not strictly constitute performance. Engaging in counterproductive behaviour or withdrawal relates to discipline rather than performance.

Adaptability in a work context is defined as how well employees deal with changes in work tasks or contexts, and includes versatility and flexibility (Dorsey, Cortina, & Luchman, 2010; Pulakos, Arad, Donovan, & Plamondon, 2000)[19]. Adaptability is becoming a more important individual quality as work tasks and contexts become less clearly defined and more likely to change from a pre-specified job description (Schmitt, Cortina, Ingerick, & Wiechmann, 2003). Some researchers expand the adaptive performance dimension to include how proactive employees are (Fay & Sonnentag, 2010; Sonnentag & Frese, 2005). Proactivity includes taking initiative in shaping work tasks and design (Fay & Frese, 2001; Frese, Garst, & Fay, 2007). Adaptability is likely to influence job performance, but it still may be inaccurate to say that it equates with performance. Proactivity and other 'useful personal qualities' may constitute contextual performance, as the useful qualities include persistence, industriousness, and facilitating peer and team performance (Borman, et al., 2010). The distinction is between a quality of an individual and a quality of an individual's actions.

Up to this point, only individual job performance has been scrutinised. Organisations are social systems, however, and employees may perform their jobs in teams. Team performance is not likely to be a simple average of the performance of individuals, because of the complex social dynamics that occur in groups, and the variations in how work is organised between team members (De Dreu, Harinck, & van

[19] C.f. adaptability and intelligence, section 1.2.1.

Vianen, 1999; Mohammed, Cannon-Bowers, & Foo, 2010; van Knippenberg, 2000). The individual focus can be attributed to the fact that organisations tend to hire individuals, but team performance may differ qualitatively from individual performance, making a comparison appropriate (Morgeson, Reider, & Campion, 2005; Sonnentag & Volmer, 2010).

Individual contextual performance such as helping and cooperating may facilitate team performance (Borman & Motowidlo, 1993)[20]. However, the relationship may not be straightforward. Team performance literature usually adopts a performance-as-product approach, often specified as team effectiveness or team outcomes (e.g. Somech, Desivilya, & Lidogoster, 2009). S. T. Bell (2007) defined team performance as "the extent to which a team accomplishes its goals or missions" (p. 595). Goal accomplishment may be more appropriate for defining team performance, because team processes may reflect group dynamics influencing performance rather than actions and interactions that constitute performance (De Dreu, et al., 1999). The complexities of group dynamics cannot be afforded deep analysis, though diversity of individual members as well as conflict and cooperation processes have been shown to affect team performance in irregular ways (Somech, et al., 2009; van Knippenberg & Schippers, 2007). Individual contextual performance is defined in terms of processes that do not necessarily have direct outcomes, so contextual performance may be more closely related to cooperation processes in teams. However, high cooperation and low conflict in teams do not necessarily lead to goal accomplishment (De Dreu, et al., 1999; Schulz-Hardt, Mojzisch, & Vogelgesang, 2008). As a result, contextual performance cannot be assumed to have a direct positive relationship with team performance.

[20] Contextual performance may be better categorised as task performance in teams, because technical core tasks may require team members to work harmoniously and interdependently (Morgeson, Reider, et al., 2005).

To summarise, task performance and contextual performance are different aspects of individual job performance. Task performance relates specifically to employees' actions towards accomplishment of technical job tasks that relate to core organisational goals, while contextual performance relates to how employees support the environment in which work takes place. Task performance may be described in terms of processes or products, but contextual performance is typically describable only as processes. Task and contextual performance are expected to relate to each other when more specific performance processes affect both, for example communication and interaction.

A team's performance is likely to relate to the individual performance of its members in irregular ways because of the complexities of group dynamics. Individual contextual performance may relate to team dynamics rather than team performance.

Having outlined some key performance products and processes, this discussion can now examine the other components of performance: purpose and judgement. The purpose of job performance is the accomplishment of organisational goals and subgoals. It may not always be clear how particular performance processes and products relate to these goals, and a key concern is whether measurements of job performance are adequately representing the purpose of performance. Similarly, different judges may have diverse perceptions of both the purpose of performance and the actions and outcomes constituting it. Therefore, different judgements of performance are likely to reflect the judge, not just the performer. These issues relate particularly to measurement of job performance.

1.3.4 Measuring job performance

Briefly, it is important to be aware of the history of job performance measurement. In industrial and organisational (I/O) psychology research and practice, job performance

is the major criterion of interest, being the ultimate factor to try and predict. If job performance could be predicted to a high degree with particular selection methods, organisations would be reassured that they can hire the best candidate for any job vacancy using particular methods. Job performance as the criterion has had a profound influence on its treatment as a research variable (Cleveland & Colella, 2010; Wallace, 1965). That influence is primarily the necessity of measuring it (Austin & Villanova, 1992), focusing researchers' attention on the scope and accuracy of job performance measurements, rather than a clear understanding of its components (e.g. Borman, 1979; Hedge & Kavanagh, 1988; Landy, Barnes, & Murphy, 1978). As a result, measurements of job performance may not represent job and organisational goals as well as researchers and practitioners require.

Decisions about performance measurement are complex: which processes and products are relevant for performance (Astin, 1964)? How should processes and products be quantified (Bernardin, 1977)? Should processes and products be considered discrete, or aggregated into a composite performance score (Schmidt & Kaplan, 1971)? The first question relates to relevance, which is how well the processes and products being recorded represent the purpose of performance. The second two questions relate to the quantitative operationalisation of job performance; in other words, job performance scores depend on adequate and justified assignment of numbers to performance processes and products. A job performance score that represents the precise contribution of an employee to achieving organisational goals and nothing else has perfect relevance. In practice, however, the employee is likely to have engaged in some action or produced something relevant to organisational goals that have not been included in the performance score, meaning the score is 'deficient'. There is also likely to be some factor influencing the score which is related neither to the employees'

actions nor their contribution to organisational goals, hence the score is ‘contaminated’. In order to assess the validity of selection procedures for predicting job performance (the criterion), scores need to have maximum criterion relevance, and minimum criterion deficiency and criterion contamination (see figure 1.3; Astin, 1964; Cleveland & Colella, 2010; Viswesvaran, 2002).

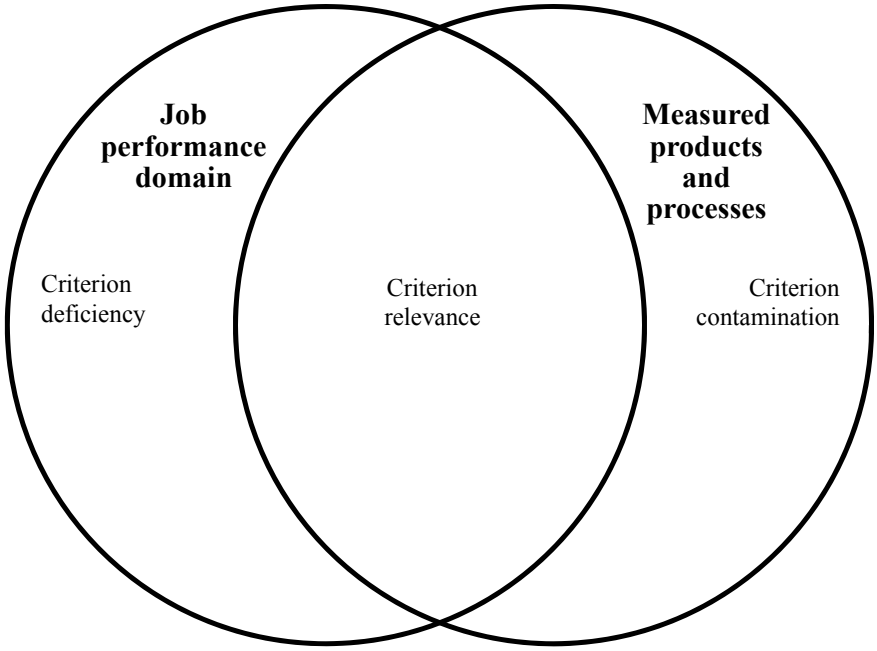


Figure 1.3 Relationships between criterion relevance, deficiency and contamination

It is impracticable for employees to be constantly observed at work, so some relevant performance processes are likely to remain unidentified in the measurement of job performance. Though theoretical arguments already support the notion that outcomes can constitute performance, the practical argument also supports recording performance products. Measures of outcomes are typically classified as objective criteria, and tend to be gathered from organisational records (Viswesvaran, 2002).

Examples include the number of students passing a course and number of research publications reflecting a university lecturer's performance; the dollar value of sales in a specified time period representing the performance of a salesperson; the time it takes to build a house representing a builder's performance.

One problem with focusing on outcomes is that records tend to conceal details that affect scores, often making it difficult to judge the level of relevance, deficiency or contamination in the score. For example, a mobile phone salesperson's dollar value of sales in a given week indicates the overall contribution to the organisation's 'bottom line' for that week (task performance, relevant)[21], but does not indicate whether the employee represented the organisation favourably to customers making them more likely to return (contextual performance, deficiency), or how many mobile phones were available to sell to how many customers (factors outside the employee's control, and lack of baseline for performance, contamination)[22]. One way to reduce contamination is for job performance scores to take into account limitations on performance. For example, seasonal differences in numbers of customers mean anticipated fluctuations in expected sales for different weeks (baseline expectations). Positive or negative deviation from the baseline could represent the salesperson's job performance. Note this may still be contaminated, as the expected value may be incorrect, but potential contamination is reduced. Deficiency, on the other hand, tends to require a score for a different aspect of performance.

[21] 'Bottom line' is the ultimate goal of an organisation.

[22] Another problem is that limited focus on performance products can influence subsequent actions. For example, if speed of house building was the only outcome of interest for builders' performance, they may be inclined to build houses faster by compromising on safety standards. If safety standards are a requirement for good performance, the outcome would be better framed as the time it took to build a house while strictly adhering to safety standards. Effects on employees' subsequent job behaviours are important considerations in performance measurement, but examination of possible effects is a topic relating to performance management rather than the nature and measurement of job performance, and therefore is beyond the scope of this discussion.

Subjective ratings of job performance may have the potential to minimise deficiency by providing estimates for relevant performance processes, excluding irrelevant factors, and taking into account situational factors (Landy & Farr, 1980; Pulakos & O'Leary, 2010). However, there are two main issues to be considered here: (1) the difficulty of transforming employee actions to ratings on scales; and (2) if performance is constituted in part by its judgement (argued in section 1.3.1), then the judge needs to be scrutinized almost as closely as the employee. The first point relates to aspects of rating instruments, but before considering the rating instruments it may be more appropriate to consider the raters, the performance 'judges'.

Ideally, it should not matter who the actual judge of performance is, as long as the judgement precisely reflects the relevance of particular processes and products to organisational goals. In practice, this may be an unrealistic expectation. For example, Borman (1978) found almost perfect agreement between raters in controlled experimental conditions, (inter-rater correlation of .9), but this high a correlation does not tend to be found in field settings. The following paragraphs examine how and why an unbiased performance judgement may be an impossible goal.

Agreement between different judges on real-world ratings of job performance is typically quite low (Viswesvaran, et al., 1996). It appears to makes little difference whether the judgements being compared are from supervisors, peers or direct reports, as agreement tends to be relatively low whether it is calculated between individual judges or between different groups (Murphy, Cleveland, & Mohler, 2001).

One explanation for lack of agreement is that the judges have different opportunities to observe employees, meaning ratings are deficient (DeNisi, Cafferty, & Meglino, 1984; Wherry & Bartlett, 1982). More commonly, however, rater effects or rater errors are interpreted as the cause of low correlations between judges' ratings,

suggesting ratings are contaminated (Landy & Farr, 1980). Halo error is an example of contamination, which is the tendency for a judge to develop a general impression of an employee, and then rate the employee on all performance dimensions according to that general impression, rather than the employee's actual behaviour on each distinct dimension (Nathan & Lord, 1983; Viswesvaran, et al., 1996; Viswesvaran, Schmidt, & Ones, 2005). However, when ratings appear to show halo, it is possible that the employees are actually consistent across all dimensions of their jobs (Bernardin & Pence, 1980). Also, if job performance is constituted in part by the judge, differences between judges would not necessarily be interpreted as errors.

Landy and Farr (1980) suggested that judges may have implicit theories about job performance; that is, which employee actions are job-related and therefore relevant to organisational goals and outcomes. Different implicit theories would lead to different behaviours being attended to in observation, ultimately resulting in different ratings of job performance. This suggestion supports the perspective that performance is a function of beliefs and motives of the judge. DeNisi (1996) argues, similarly, that ratings of performance should be considered with regard to the rater's cognitive processes. The cognitive process model outlines how, in producing an appraisal rating, the rater must (1) observe an employee's job performance, (2) remember it, (3) assume that (or decide whether) the observed behaviour was representative of the employee, (4) translate that information into a rating on a scale, and (5) be mindful of the employee's reaction to the rating (DeNisi, 1996; DeNisi, et al., 1984). This simplified explanation of DeNisi's model outlines a number of areas where supervisor beliefs, biases and errors could influence ratings. Training programmes for raters and formal diary-keeping methods can reduce some of these issues, but they may never be able to be completely

resolved (Bernardin & Buckley, 1981; Borman, 1975; Pulakos, 1984; Saal, Downey, & Lahey, 1980).

Rating instruments have been the focus of a considerable amount of research in order to improve the transformation of observed behaviours to quantitative estimations of the organisational or job-related value of those behaviours (Bernardin, 1977; Borman, 1974; Borman, et al., 2001; Latham & Wexley, 1977; P. C. Smith & Kendall, 1963). Much of the work on rating scales focused on improving the accuracy of raters, however, meaning the accuracy of a rating as an index for the value of job-related behaviour was entangled with the accuracy of selecting a rating (Latham & Wexley, 1977; Murphy, Garcia, Kerkar, Martin, & Balzer, 1982; Wherry & Bartlett, 1982). In order to accurately index the value of observed behaviours, the best approach may be in specifying objectives that are directly related to job and organisational goals, and indicating after some pre-arranged time frame how well employees achieved those objectives (London, Mone, & Scott, 2004).

Pulakos and O'Leary (2010) suggested setting three to five broad objectives for each employee. In line with Locke and Latham's (2002) goal setting theory, each objective must be: (1) job relevant; (2) specific; (3) difficult but achievable; and (4) measurable in terms of timeliness, quality, quantity and/or cost effectiveness. Ideally, employees' objectives should be within their control to accomplish, and objectives requiring employees to work together will be specified as team objectives. There should be some flexibility to alter objectives as unforeseen events occur. When deciding whether objectives have been met and how well (that is, rating job performance), each rating could include three broad levels indicating inadequate, satisfactory and exceptional performance, allowing employees to perform beyond expectations. This technique provides information of direct importance to organisations about employee

and team task performance, allowing for some flexibility around uncontrollable factors[23].

Setting objectives satisfies all three components constituting job performance: the purpose is specified, as the objective must relate directly to job and organisational goals. The objective is by definition a performance product, and employees have some flexibility in the processes they use to accomplish it. The judge still defines performance in terms of the content and difficulty of the objectives, and in the judgement of how well objectives were achieved, but there should be less influence of biases and errors. One deficiency of objective setting is that contextual performance is diminished, so including a rating of observed contextual performance may complete the picture of employees' job performance.

1.4 Relationship of cognitive ability to job performance

Thus far, this review has examined the nature and measurement of cognitive ability and job performance in detail. Key conclusions so far are that cognitive abilities underlie correct responses on intelligence tests, and most of the variance in test scores can be explained by *g*, *Gf* and *Gc* factors. Job performance is comprised of judgement, purpose, and process and product components, important distinctions existing between individual task and contextual performance and team performance. This section examines how cognitive abilities and job performance are related. It is first appropriate to briefly examine antecedents of job performance.

1.4.1 Antecedents of job performance

Individual job performance is hypothesised to be a function of employees' declarative knowledge, procedural knowledge and skill, and motivation (Campbell, et

[23] It also may be an effective performance management tool (Smither & London, 2009).

al., 1993; Schmitt, et al., 2003). Declarative and procedural knowledge are described as 'knowing that' and 'knowing how', respectively (Sternberg, 2005). Individuals' declarative knowledge tends to relate in particular to their education. Procedural knowledge is not sufficient for 'doing', so procedural skill is also necessary. Procedural knowledge and skill are associated with practical intelligence, emphasising the importance of relevant experience (Schmitt, et al., 2003; Sternberg, 2000; Sternberg, Wagner, Williams, & Horvath, 1995; Wagner & Sternberg, 1985). Motivation is the impetus to exert effort, including how much and how long for (Campbell, 1990). Motivation is important because any amount or type of knowledge or skill may be of little practical use to an organisation if an employee chooses not to (or has no reason to) exert effort in applying them (Elliott & Dweck, 1988; Kanfer, 1990; Locke & Latham, 2002).

Opportunity to perform (or perceived opportunity) occasionally appears as an additional antecedent to account for environmental factors on performance such as availability of tools, number of customers, or opportunities for supervisors to observe (Blumberg & Pringle, 1982; Frese, et al., 2007; Wherry & Bartlett, 1982).

These antecedents are assumed to operate via relationships and interactions between cognitive, physical and psychomotor abilities, personality, integrity, interests and experience (Campbell, et al., 1993; Schmitt, et al., 2003). The focus here is narrowed to cognitive abilities.

1.4.2 Cognitive ability and individual job performance

The *g* factor of cognitive ability has a stronger relationship with supervisor ratings of job performance than other individual difference variable including integrity and 'big

five' personality factors[24]: (Barrick & Mount, 1991; Hunter & Hunter, 1984; Ones & Viswesvaran, 2007; Ree, Earles, & Teachout, 1994). Specifically, after correction for unreliability and range restriction, *g* correlates at about .5 with ratings of overall job performance, explaining 25% of the variance in ratings.

One theory explaining this consistent finding is that cognitive abilities are important for acquiring declarative knowledge, and to some degree procedural knowledge (Schmitt, et al., 2003). There is support for this theory because the relationship between cognitive abilities and performance has been shown to be mediated to a large degree by job-relevant knowledge (Bergman, Donovan, Drasgow, Overton, & Henning, 2008; Ones & Viswesvaran, 2007; Schmidt & Hunter, 2004). Because of the importance of job-relevant knowledge, it is surprising that biodata had low criterion-related validity in Schmidt and Hunter's (1998) meta-analysis, because qualifications and job-relevant experience would be assumed to contribute to job knowledge and therefore job performance. It is possible that biodata were not weighted adequately in favour of these personal history factors (Arthur & Villado, 2008). In addition, the relationship between experience and job performance varies over time spent in a job (see section 1.4.3), and time in the job was not examined as a potential moderator.

In interaction with physical abilities and personality variables, cognitive ability may also influence acquisition of procedural skill through learning and experience (Motowidlo, et al., 1997). This relationship is logically weaker than the relationship between cognitive ability and declarative or procedural knowledge, however (Schmitt, et al., 2003; Sternberg & Wagner, 1993).

Cognitive ability and motivation are not typically associated (e.g. Campbell, et al., 1993; Motowidlo, et al., 1997; Ree, et al., 1994), though Perkins and Tishman (2001)

[24] Openness to experience, conscientiousness, extraversion, agreeableness and emotional stability (McCrae & Costa, 1985).

suggest that to some degree, cognitive ability may be a result of motivational processes (see also section 1.2.3).

Cognitive ability may link with some dimensions of job performance to a greater extent than others. For instance, Motowidlo, Borman and Schmit (1997) theorised that cognitive ability may be more closely related to task performance because of tasks' technical requirements, while personality would be more closely associated with contextual performance given the more interpersonal nature of the dimension (see also J. Hogan & Holland, 2003). In some studies cognitive ability was found to be strongly related to both task and contextual performance, though personality added significant incremental validity to both dimensions as well (Bergman, et al., 2008; Motowidlo, Brownlee, & Schmit, 2008). In contrast to these findings, other research showed that when overall job performance was predominantly non-technical, cognitive abilities contributed negligible amounts to the reduction of variance in performance ratings, with ratings being predicted better by trait complexes, comprised of interactions between personality, motivational styles, interests and self-concepts (Kanfer, Wolf, Kantrowitz, & Ackerman, 2010).

There is little research examining specific cognitive abilities on overall job performance task performance or contextual performance. Ree, Earles and Teachout (1994) found that beyond the criterion-related validity of *g*, consideration of specific abilities added little incremental validity of practical value when matched with job demands. However, Lang and colleagues (2010) analysed the relative weights of specific abilities and *g* and found that verbal comprehension was a better predictor of job performance in one study. The utility of g as a predictor is not in dispute, but there may be theoretical and practical value in examining how specific abilities are related to

different dimensions of job performance without partialling out the variance attributable to *g*.

In summary, individual differences in the general factor of cognitive ability account for the greatest proportion of variance in job performance ratings when compared with other individual differences. There is evidence that cognitive ability relates to job performance through the acquisition and application of declarative and procedural knowledge related to the job. Cognitive ability may be related more closely to task performance than contextual performance because declarative and procedural knowledge are expected to relate to task performance, while personality factors and motivation are more likely to influence contextual performance. Research into the relationships of specific abilities and job performance has demonstrated there is limited practical contribution of specific abilities when variance attributed to *g* is excluded, but there may be relationships between specific abilities and job performance when variance usually attributed to *g* is shared. There may be value in not excluding variance attributable to *g* at all, and examining these abilities in comparison to different dimensions of individual job performance.

1.4.2 Cognitive ability and team performance

Though sparse compared to individual job performance literature, there is some research investigating the relationship of individual cognitive ability to team performance. Group dynamics may influence team performance to such an extent that it becomes difficult to discern systematic relationships between individual differences variables and team performance (Mohammed, et al., 2010). However, individual differences more closely associated with contextual performance may influence team performance. A single individual's interpersonal skills and aspects of their personality,

for example, are hypothesised to have a direct influence on team performance as they can hinder or facilitate group functioning (LePine & Van Dyne, 2001; Morgeson, Reider, et al., 2005; though see also section 1.3.2).

The relationship of cognitive ability to team performance is not direct. LePine and colleagues (1997) found that when teams have comparable average cognitive ability, a team containing a member with low cognitive ability performs as well as teams where all the members have roughly the same ability level. However, teams who have a single member with low conscientiousness do not do as well as teams with comparable mean conscientiousness. The interpretation of these results was that team members may be more willing to compensate for a member with low cognitive ability compared to a member with low conscientiousness. This demonstrates how the effect of cognitive ability is moderated in relation to team performance, while personality has a direct effect, mediated perhaps by its effect on other team member's motivation (LePine, et al., 1997; van Knippenberg, 2000). Therefore, while personality factors may have direct relationships to team performance, it is generally not anticipated that a given individual's cognitive ability will be strongly related to team performance.

The effect of cognitive ability on team performance may be contingent on multiple factors (Mohammed, et al., 2010). It has been found, for example, that an individual team member's cognitive ability level may influence team performance depending on how work is organised in the team. LePine and colleagues (1997) found that the lowest cognitive ability level of a team member related most closely to performance when the tasks required each member to contribute something unique. However, if there is a team leader coordinating and directing the other members, then the leader's ability may be more predictive of team performance. Or if tasks are solved by the group collectively, then the average cognitive ability level of the group might be more predictive of

performance (LePine, et al., 1997). Task complexity also modifies the relationship between cognitive ability and team performance, as teams with higher average cognitive abilities adapted to sudden changes in tasks better than teams with lower mean cognitive ability (LePine, 2003). In summary, an individual's cognitive ability is unlikely to predict team performance without taking other factors into account.

1.4.3 Limitations of field research

This final chapter outlines some of the limitations involved in estimating the relationship between cognitive ability and job performance when data are collected from employees of organisations. Adequate consideration of the following points is important for the empirical study.

When collecting real-world job performance ratings, researchers occasionally fail to recognise or acknowledge the context in which performance appraisal occurs in organisations. There are likely to be purposes for performance appraisals other than accurately scoring job performance (Murphy & Cleveland, 1995). For example, ratings provided for research alone tend to be substantially different to those provided as part of a performance management system (Borman, et al., 2010; Reb & Greguras, 2010). It appears that in some organisations, or for some raters, accurately rating performance is not the primary goal. The appraisal may be a political exercise where ratings are deliberately manipulated; for example, to ensure an employee gets a pay rise (Longenecker, Gioia, & Sims, 1987). Or appraisal may be a social exercise where the appraisal provides one-to-one time between a supervisor and an employee which might not otherwise be available (Murphy & Cleveland, 1995). It is therefore important to consider the context in which ratings were produced and how the context might affect the ratings.

Another problem in researching the relationship of cognitive ability to job performance is that job performance is dynamic (Austin, Humphreys, & Hulin, 1989; Deadrick & Madigan, 1990; Ghiselli, 1956; Murphy, 1989). An employee's behaviour may fluctuate between typical and maximum performance (Deadrick & Gardner, 2008), job-relevant training is assumed to lead to permanent enhancement of job performance (Tannenbaum & Yukl, 1992), and new team members may cause short- and long-term changes in team performance (S. T. Bell, 2007). Therefore, when attempting to measure performance, the implication is that the best that can be achieved is an aggregate score of job performance over time (Ghiselli, 1956; Reb & Greguras, 2010)[25]. The greater the length of time between a given job performance rating and the rating from time *x*, the more dissimilar the ratings will be (Deadrick & Madigan, 1990). Patterns of individual job performance follow a learning curve, though individuals tend to have different performance trajectories, which may be a function of individual differences (Deadrick, Bennett, & Russell, 1997; Ployhart & Hakel, 1998; Zyphur, Chaturvedi, & Arvey, 2008).

Investigations into cognitive ability and dynamic performance showed that the predictive validity of cognitive ability increased over time, while the validity of prior experience decreased (Deadrick & Madigan, 1990). One explanation for this result is that prior experience correlates well initially with job performance due to greater job-relevant knowledge. Once in the job, cognitive ability predicts acquisition of job- and organisation-relevant knowledge, which will differ in some ways to the knowledge acquired in previous jobs. Therefore, the higher the cognitive ability, the more relevant (and greater) the job knowledge over time, enhancing performance and increasing the

[25] Some researchers argued that job performance was stable over time and variations were due to unreliability in the measures (G. V. Barrett, Caldwell, & Alexander, 1985), but subsequent research showed that job performance does change over time (Deadrick & Madigan, 1990).

predictive validity of cognitive ability. Therefore, it is important to note how long employees have been in a job when examining predictive validity.

Job diversity can also affect the generalisability of findings from one organisation or one job to another. One example of how job diversity affects the relationship between cognitive ability and job performance is that how the relationship is largely mediated by job-relevant knowledge (section 1.4.1). Another example is that job complexity moderates the relationship between cognitive ability and job performance. That is, cognitive ability job performance are more highly correlated in complex jobs compared to less complex jobs (Schmidt & Hunter, 1998). The predominant explanation for this moderation is that jobs that are higher in complexity require greater cognitive ability in order to be performed well. Jobs with less complexity can be performed by individuals with high or low ability levels, meaning that personality and motivation may be more important factors contributing to job performance in lower complexity positions. Job complexity may be caused by high cognitive ability in some situations, not only because individuals with high cognitive ability scores might apply for more complex positions, but also because employees capable of taking on more tasks may voluntarily increase the complexity of their jobs, which in turn is associated with higher ratings of job performance (Morgeson, Delaney-Klinger, & Hemingway, 2005). It may also be the case that task performance is emphasised over contextual performance in more complex jobs, leading to greater associations between cognitive ability and overall job performance for more complex jobs (Morgeson, Delaney-Klinger, et al., 2005; Parker & Wall, 2002)[26]. The applicability of different performance processes (outlined in section 1.3.2) to task and contextual performance are likely to vary between jobs also (Borman,

[26] For example, the importance of task and contextual performance for a pilot compared to cabin attendants (section 1.3.2).

et al., 2010; Campbell, 1990). Therefore researchers need to be cautious in generalising performance results without considering diversity of jobs.

In summary, issues that researchers need to recognise in generalising results are: (1) the context from which job performance ratings were produced and the likely effect that context had on those ratings, (2) that job performance is dynamic, and the relationship of cognitive ability to performance becomes stronger over time, (3) jobs and organisations are diverse, and performance ratings from one job or organisation may not be adequately generalised to ratings from another. Specific job- or organisation-related mediators and moderators of cognitive ability and job performance relationships should be considered when possible.

1.5 Research objectives

This aims of this research are to examine cognitive ability scores and job performance ratings of employees from a New Zealand government organisation. These data will be compared with previous research to see if they support similar conclusions, and some analyses are intended to address some perceived gaps in existing research. First, because there is overwhelming evidence for positive manifold resulting in a general factor of cognitive ability (section 1.2.5), when analysing three specific cognitive abilities – verbal, numeric and abstract reasoning – it is expected that they will be positively correlated. These cognitive abilities represent fluid and crystallised intelligence in particular, though are all likely to load highly on the *g* factor.

Individual task performance and contextual performance are expected to be related to each other because particular performance processes may be applicable to both (section 1.3.2). Contextual performance might relate to team performance because individual differences that are hypothesised to relate to contextual performance may

also have a direct influence on team performance (section 1.4.2). However, because contextual performance is comprised predominantly of processes, it may influence team dynamics rather than team performance, because team performance is described almost exclusively in terms of products (section 1.4.2). In this case, the influence of contextual performance would be mediated by team dynamics, and may not result in a relationship with team performance.

One of the primary contributions of this research may be in the analysis of relationships between specific cognitive abilities to task, contextual and team performance (sections 1.4.1 and 1.4.2).

Abstract reasoning ability (educing logical relationships in shapes and patterns, Spearman, 1927) tends to load highly on fluid intelligence and *g* factors, indicating that it may be pertinent especially when employees are new to a job, or when they undergo training. Because the employees in this study were all relatively new to their jobs, it is hypothesised that abstract reasoning will relate to task and contextual performance.

Numeric reasoning ability (performing quantitative operations) loads on fluid intelligence, crystallised intelligence and *g* factors, and is likely to influence task performance in particular, given the technical nature of numeric operations and tasks. However, numeric reasoning may relate to task performance when jobs require this ability specifically, for example in accounting, physical or social sciences, engineering or architecture. Because the participating organisation is a service organisation, numeric reasoning may not relate strongly to task or contextual performance.

Verbal reasoning ability (knowledge of word meanings and relations between words) loads predominantly on crystallised intelligence and *g* factors. It is likely to be implicated across virtually all occupations, given the ubiquity of using language and the necessity of communication, although customer service, teaching and managing may

require particular strengths in this area for task performance. In as much as verbal ability relates to communication skills, however, there may be a strong relationship between this specific ability and interpersonal aspects of contextual performance. Given the service focus of the organisation, it is expected that verbal reasoning will be related to task performance and contextual performance.

Abstract, numeric and verbal reasoning abilities are not expected to show a relationship with team performance, because complex interpersonal dynamics are likely to moderate the influence of individual-level cognitive abilities.

The data will be interpreted bearing in mind the considerations outlined in section 1.4.3, including the context of performance ratings, the length of time participants have been employed by the organisation, and recognising limits to generalisability due to job and organisation diversity, specifically examining how characteristics of the organisation's service goals might influence the results.

2 Method

The data were collected from records held by a New Zealand government organisation[27]. The organisation regularly uses cognitive ability tests as a selection tool, and uses performance appraisal ratings as annual measures of job performance. The data were analysed quantitatively, relying on correlation analysis to examine the relationships between variables. This chapter describes in detail: the known characteristics of the participant group; the instruments used, including their development and psychometric properties; details of each variable; the procedure for data collection; and analyses conducted, including those relating to data adequacy and substantive investigations.

The substantive hypotheses, summarised from section 1.5, were:

1. Verbal, numeric and abstract reasoning will be related to each other.
2. Task performance and contextual performance will relate to each other.
3. Task and contextual performance will not relate to team performance.
4. Team performance will not relate to cognitive abilities.
5. Abstract reasoning will be related to task and contextual performance.
6. Numeric reasoning will not be related to any job performance ratings.
7. Verbal reasoning will be related to task and contextual performance.

2.1 Participant characteristics

Participants were 43 employees of the organisation hired within a 5 month period between 1st December 2009 and 28th April 2010. Twenty-six (60.5%) were female. The minimum age of a participant was 20, maximum 52, mean age 33.37 (standard deviation 7.7 years).

[27] The organisation will remain anonymous in accordance with a confidentiality agreement.

Twenty-six participants were of New Zealand European ethnic origin, 4 were Samoan, 3 Māori, 3 British, 1 Indian, 1 Malaysian, 1 Australian and 1 North American. Three participants did not state their ethnicity.

Although there are a higher proportion of women in the group compared to the population of New Zealand, the ethnic group proportions are roughly representative of the New Zealand population (EEO Trust, 2007).

The participants held different job roles in the organisation. A list of participants' specific job titles may identify the organisation, but job roles included administrator, manager, coordinator, analyst and advisor.

2.2 Instruments

2.2.1 Internet reasoning test

The internet reasoning test[28] (IRT2) is a short-form, unsupervised version of the General Reasoning Test battery (GRT2). The GRT2 was developed by Psychometrics Ltd in 1993, and was used as the basis for the IRT2 in 2008. The IRT2 was developed in response to a suggestion in an independent review regarding the use of the GRT2 for New Zealand police selection (Hattie, 2007). The IRT2 is distributed by Psytech International Ltd and is used in this study courtesy of the New Zealand distributor, OPRA Consulting Ltd.

The IRT2 was designed to assess cognitive ability in adults for the purposes of personnel selection, for job candidates from the general population[29]. Because of its

[28] 'Internet' reasoning refers to the mode of administration, not the nature of the reasoning abilities being tested.

[29] The same as the GRT2. Cognitive ability tests tend to produce ceiling or floor effects (all candidates obtaining very high or low scores) if the test is not pitched at an appropriate level for the group of candidates. The GRT1 (graduate reasoning test) was designed for jobs requiring candidates with higher than average ability levels such as university graduate positions, managerial and executive-level jobs. There is not an equivalent IRT1, with the IRT2 being variously referred to as simply the IRT. This has been avoided in order to reduce confusion between the acronyms for the internet reasoning test (IRT2) and item response theory (IRT).

close relation to the GRT2, there is no dedicated technical manual for the IRT2, so this section will specify when GRT2 characteristics are supplementing missing information about the IRT2.

The IRT2 test battery is made up of three subtests: a verbal reasoning subtest, a numeric reasoning subtest and an abstract reasoning subtest. These tests collectively estimate general cognitive ability, fluid intelligence and crystallized intelligence factors that are designed not to be heavily dependent on quality and quantity of education, but are job-relevant. Subtests are administered one at a time, and each subtest is timed. Items were designed to have simple content and be quite short, to minimise and most efficiently use testing time. All items have six possible responses, in order to reduce the effects of guessing.

The verbal reasoning subtest has 17 items and candidates have 4 minutes in which to answer them. Verbal reasoning requires knowledge of general vocabulary and relations between words, and is broadly related to crystallized intelligence as well as general cognitive ability. An example of a verbal reasoning item is:

Sick means the same as:

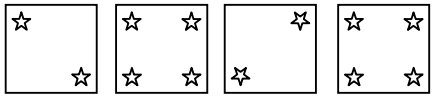
1. Labour
2. Ill
3. Healthy
4. Disease
5. Evil
6. Rest

The numeric reasoning subtest has 15 items and candidates have 4 minutes in which to answer them. Numeric reasoning requires knowledge and application of mathematical operations, and is related to both fluid and crystallized intelligence, and general cognitive ability. An example of a numeric reasoning item is:

Which of the following is the odd one out?

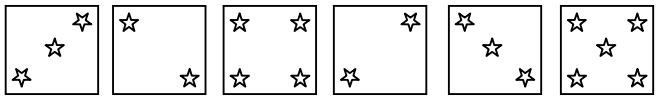
1. 2/4
2. 6/8
3. 3/6
4. 4/9
5. 4/8
6. 2/6

The abstract reasoning subtest includes 15 items and candidates have 6 minutes in which to answer them. Abstract reasoning requires induction and deduction of logical relations in shapes and patterns, and relates to fluid intelligence and general cognitive ability. An example of an abstract reasoning item is:



What comes next?

1. 2. 3. 4. 5. 6.



Testing procedure: The organisation received applications for vacant positions and initially screened applicants based on personal history information. The participants in this research were job candidates who successfully proceeded through this initial screening and were invited to participate in cognitive ability testing as part of the ongoing selection process. Participants were able to complete the test any time they chose, in any location. Access to a computer and the internet was required[30] and a time frame for completing the test would have been set by the hiring organisation. The computer programme led the participants through the test battery, with each subtest being administered separately with dedicated instructions and examples. See appendix two for specifics of how candidates proceeded through the computer programme to

[30]If participants did not have access, they could sit the tests at the organisation's premises.

complete the subtests. The hiring organisation received a score report when each participant completed a test.

Anxiety can negatively influence test performance, so should be reduced as much as possible (Reeve, Heggestad, & Lievens, 2009; Zeidner & Matthews, 2005). One criticism of the testing procedure is that the examples offered to candidates as a practice prior to test administration are quite easy. Although this may initially reduce participant anxiety associated with testing in general, anxiety may return or be even worse when the actual test begins and test items are harder than the examples might have implied. One mitigating factor is that practice tests can also be taken online via OPRA Consulting Ltd's website, and candidates are encouraged to sit these practice tests first. Though the practice tests do not use the same website format as the GeneSys testing platform, candidates can practice answering test items of similar difficulty to items in the IRT2 under time pressure, which should help reduce anxiety in the testing situation.

It is good that participants are encouraged to complete the tests when they will be free from distractions, and also that the web browser is specially designed to cover the toolbar that runs across the bottom of a computer desktop, actively reducing distraction. The programme is well designed to cope with disruption as well, and when there are problems, candidates can access the administrator via the original email.

Test administration: The IRT2 randomly draws the items for each subtest from a pool of equivalent items, in order to reduce the risk of public disclosure of items, which may occur with unsupervised tests.

The main advantage of the IRT2 is that it is so easily administered because the hiring organisation does not have to arrange for candidates to attend a supervised session and the candidates can sit the test at a time and in a place convenient for them. However, because the hiring organisation cannot be sure that the test was actually

completed by the named candidate, Psytech recommends that the hiring organisation set up a supervised test as well. The recommendation appears in every score report, along with a reminder that IRT2 scores are an indicator only, and that scores must be interpreted within the context of other information about the candidate.

Scoring: The IRT2 is initially scored by calculating the proportion of items each candidate got correct out of the number of items they attempted. Raw scores are translated into 'normed' sten scores and percentile rank scores by comparing the candidate's raw score to raw scores from a reference group chosen by the hiring organisation. In the present study the participants for the verbal and numeric reasoning subtests were 1,523 New Zealanders, with 1,551 New Zealand respondents in the reference group for the abstract subtest. See appendix three for demographics of each subtest reference group. Information in appendix three shows that while there were many missing details overall, the known demographics suggested the reference group was comprised of a diverse group of participants.

Sten scores have a mean of 5.5 and a standard deviation of 2, providing scores for participants between 1 and 10. This means most candidates score between 4 and 7, and scores within this range can be considered relatively indistinguishable (a point emphasised on the training course for interpreting the test reports). The percentile rank indicates the proportion of the reference group who scored less than the candidate. For example, a percentile rank 94 means the candidate's score was the same as or higher than 94% of the reference group. Because norms must be regularly updated to ensure they continue to represent the population, and any changes in scores that may have occurred over time (see Flynn, 1987), GeneSys system users can regularly download updates from the OPRA Consulting Ltd website.

Score report: The report includes (1) a computer-generated interpretation of candidates' sten score, (2) each candidates' raw score (total number of items correct) on each subtest; (3) the number of items they attempted out of the total number of items on each subtest; (4) the sten score between 1 and 10 on each subtest; and (5) the percentile rank on each subtest. Sample reports are publicly available at http://www.opragroup.com/index.php/assess/ability-assessments.

Psychometric properties: Because the IRT2 was adapted from the GRT2, psychometric properties of the GRT2 are assumed to generalise to the IRT2.

Internal consistency reliability estimates for the GRT2, based on 5,183 sets of responses from New Zealand adults, showed Cronbach's alphas of .83, .88 and .82 for the verbal, numeric and abstract reasoning tests, respectively (Budd, 1993). These coefficients suggest excellent internal consistency reliability, indicating that the items in each subtest are measuring the same ability. Test-retest correlation coefficients were .81, .84 and .78 for the verbal, numeric and abstract subtests, respectively, based on a group of 54 students who took the test twice with a two-week gap between testing sessions. These coefficients also indicate an acceptable level of temporal consistency in scores. However, Kline (1993) recommended at least a 3 month gap between testing sessions, because participants may have remembered their answers and attempted to repeat them rather than trying to answer correctly, which would lead to inflated test-retest coefficients. In addition, Kline suggests that participant groups should have at least 100 respondents and be representative of the population of interest. Given that the test-retest group had 54 students (whose nationality and other demographics are unknown), these test-retest coefficients may not adequately index the stability of the GRT2 in the general population.

GRT2 subtest scores correlated between .59 and .63 (N = 5,183) with each other showing evidence of both convergent and discriminant validity. Convergent validity was demonstrated because the reasoning abilities correlated quite well (suggesting a g factor), and discriminant validity was demonstrated because the correlations are not so high as to suggest that the subtest items are so similar that they measure the same thing.

Concurrent validity was demonstrated by correlating 46 sets of GRT2 scores with the AH3 subscale[31]. The subtest scores correlated at .58, .76 and .65 with the AH3 subscales, which may be considered further evidence of convergent validity (Netemeyer, Bearden, & Sharma, 2003), though the number of sets was disappointing, and no details of participant characteristics or methods were provided. Significance levels were not reported for any of the validity correlation coefficients, though it is assumed that they are statistically significant at $p < .05$. Further discriminant validity was demonstrated by correlating 94 sets of GRT2 subtest scores with scores on the technical test battery (TTB) subtest scores (another test battery distributed by Psytech International Ltd). Correlations ranged from .34 to .47 and were significant at $p < .01$, indicating that the abilities measured are distinct from each other to some degree.

GRT2 subtest scores were shown to correlate to varying degrees with job performance in a range of professions, demonstrating some degree of criterion-related validity. Correlations appeared to be specific to each group examined, however, perhaps because of differences in job demands and the dimensions of the job being evaluated; for example, verbal reasoning scores correlated .32 with overall job performance of 68 managers in the livestock industry in New Zealand, but did not correlate at all with numerical skills, IT skills or overall job performance in a group of 118 bankers (nationality not specified).

[31] The AH3 subscale was from a commonly used series of tests developed by Alice Heim at Cambridge University (Budd, 1993)

Though the technical manual did not provide detailed results of item analysis, it stated that item-total correlations were at least .3, indicating a good minimum relationship between the response to a single item and the total subtest score (good item discrimination), and that the items were "of graded difficulty" (Budd, 1993, p. 3), presumably indicating an adequate range of p-values[32]. The GRT2 items also did not show any measurement bias for any particular ethnic groups in New Zealand (Chernyshenko, 2005). Men scored higher than women in the numeric reasoning subtest, which is a common finding (Ones, Viswesvaran, & Dilchert, 2005). The difference is not large, however, and in the context of the other subtests and selection procedures, selection decisions should result in adverse impact. There was no evidence of measurement bias for gender in British or international participant groups (Budd, 1993).

Three independent reviews have judged the GRT2 as having adequate psychometric properties for its purpose (British Psychological Society, 2003; Chernyshenko, 2005; Hattie, 2007). As stated, it is assumed that these conclusions can largely be generalised to the IRT2. If there are differences in psychometric properties between the IRT2 and GRT2, they are likely to relate primarily to reliability: because there are fewer items in IRT2 subtests, Cronbach's alpha is likely to be lower, as this coefficient is sensitive to number of items in a test (Lord & Novick, 1968). Similarly, given the random presentation of similar items (as opposed to a repeat of exactly the same items) test-retest reliability would be lower also. Given that there is some margin for all the reliability coefficients to decrease before they fall below the critical level of .7 (Kline, 1993), a small reduction in reliability from the GRT2 to the IRT2 is not expected to be problematic. Though a reduction in reliability may cause slightly smaller

[32] P-values are the percentage of participants (from pre-testing) correctly answering items, with smaller percentages representing more difficult items (De Champlain, 2010).

correlations due to greater variability in scores, validity evidence is not expected to differ to a great extent because the constructs being measured are the same.

2.2.2 Adaptive general reasoning test

The Adaptive general reasoning test (AdaptGRT) was developed by Psychometrics Ltd for Psytech International Ltd in 2007, to provide an adaptive test utilising the principles of item response theory (IRT). The advantages of such a test are (1) that cognitive ability can be estimated with more precision and greater reliability using fewer items; (2) item calibration and parameters are less dependent on characteristics of pre-tested participants; (3) that the risk of test exposure is very small, because there is only a small probability that an item will be repeated on different occasions; and (4) that items that are selected as a result of candidates' answers, meaning that the test will be neither too difficult nor too easy but pitched at an appropriate level around each candidate's ability (Chernyshenko, Stark, & Drasgow, 2008).

The nature of IRT means that it is possible to continuously check the model assumptions, refine the test items, and add more items to the pools as required. The technical manual for the AdaptGRT describes the test's development and analyses of the original item pools, but does not specify details of alterations that have been made since the test's inception. Therefore the description below relates to the initial version of the AdaptGRT, which may have some subtle differences to the version the participants' received during their selection procedure. Possible differences will be outlined as appropriate.

The AdaptGRT assesses verbal, numeric and abstract reasoning abilities for selection purposes like the IRT2. Explanation of these cognitive abilities, item design and examples of item types are the same as those outlined in section 2.2.1. As with the

IRT2, all items have six possible response options, to reduce the effects of guessing. Many aspects of the testing procedure are identical to the IRT2 (see appendix two for details).

Test procedure: The difference in AdaptGRT test procedure compared to IRT2 is the number of items in each subtest and the time allocation for answering them. There are 15 items in all AdaptGRT subtests. The verbal reasoning subtest allows 6 minutes to respond to the items, while the numeric and abstract reasoning subtests allow 8 minutes each.

Test development: The main difference between the IRT2 and the AdaptGRT relates to how items are calibrated and selected for inclusion in a subtest, and how cognitive ability is estimated (see section 1.2.7).

The first step in developing the AdaptGRT was item generation. For the verbal, numeric and abstract reasoning subtests of the AdaptGRT, initial pools of 266, 177 and 126 items, respectively, were generated (Chernyshenko, et al., 2008). Because it was impracticable to pretest participants using all 571 items, between 20 and 30 items from the pools were presented for each subtest. The number of participants pretested varied between 205 and 6739 for each subset of items, with an average of 1586 participants per subtest. The pretests were presented as practice tests for job candidates who would be sitting other Psytech International tests (such as the GRT2) as part of a selection process. This meant that the population of participants sitting the practice tests was likely to be similar to those who would participate in the AdaptGRT during selection. For confidentiality, demographics of pretested participants were not recorded, but according to the AdaptGRT technical manual, most were applicants for skilled jobs in the United Kingdom, with about equal proportions of men and women, from various national and ethnic groups (Chernyshenko, et al., 2008, p. 18). Though the

generalisability of this group internationally may be questionable, IRT assumes that item parameters are invariant: that is, the logistic function of an item does not change no matter who is responding, when they respond, or in which context they respond. It is a important when using IRT models that such assumptions must be checked (Hambleton & Bejar, 1983; Hutchinson, 1991). Psytech International Ltd regularly checks these assumptions, as shown in the score reports stating that the reference groups are 526, 564 and 546 international respondents who were examined in 2010 on the verbal, numeric and abstract reasoning subtests, respectively.

Item analysis for the AdaptGRT was initially run using classical test theory practices: that is, p-values representing item difficulty and item-total correlations representing item discrimination were analysed. Items with p-values below .15 (that is, below chance level of a correct response) were re-examined for content problems and eliminated if necessary, along with items with p-values nearing 1, as these would not discriminate at all between candidates. Items with item-total correlations that were either negative or near zero were also re-examined and either rewritten or eliminated. As a result of these criteria, 26 (10%) verbal reasoning items were eliminated, 11 (6%) numeric items were eliminated, and 6 (5%) abstract reasoning items were eliminated, leaving 240 verbal reasoning items, 166 numeric reasoning items and 120 abstract reasoning items.

The AdaptGRT is a forced-choice, multiple choice test, meaning there are six responses for each items and examinees cannot skip items. The pseudo-guessing parameter was therefore included. In addition, there was evidence of variability in the item-total correlations, meaning items did not necessarily discriminate equally well, meaning the discrimination parameter was also required. This resulted in the use of the three-parameter logistic model (3PLM-IRT). The parameters included were item

difficulty, discrimination and the pseudo-guessing parameter. Item parameters were estimated using BILOG-MG for Windows (Chernyshenko, et al., 2008). Participants' responses modelled against their cognitive ability estimates fit the 3PLM-IRT model very well, suggesting that the model assumptions were supported by the data.

With the item parameters calculated, their suitability for adaptive testing was examined. Items with discrimination parameters below 0.4 were deleted because of the low probability of these items being selected by the computer programme. That is, an item that does not discriminate between differing ability levels very well is unlikely to provide a large additional amount of information about a candidate's ability, so is unlikely to be selected with the algorithm designed to maximise information. Items with discrimination parameters between 0.4 and 0.5 were retained only if they had difficulty parameters in the medium to high ranges, as there were fewer items available in these ranges. As a result of these additional criteria, some items were eliminated and the final item pool had 200 verbal reasoning items, 155 numeric reasoning items and 110 abstract reasoning items. The AdaptGRT technical manual provides the results of both IRT and CTT item analyses for every item in the pool. To summarise item parameters: discrimination parameters for all subtests were between 0.42 and 2.41 with a mean of 0.87 and standard deviation 0.31, which indicates generally good levels of discrimination. Item difficulty parameters were between -4.80 and 4.07, mean -0.78, SD 1.25, demonstrating a good range of cognitive ability levels being tested, though with slightly more in the lower ability ranges. Pseudo-guessing parameters were between .03 and .50, mean .16, SD .06, showing generally low probabilities of correct responses due to guessing.

Scoring: Cognitive ability is estimated by examining the candidate's likeliest range of ability levels given the responses to the items. The authors of the AdaptGRT

technical manual use the analogy of diagnosing an illness: the physician examines symptoms (analogous to item responses) and works out the most likely diagnosis (cognitive ability estimate). The AdaptGRT uses the expected a posterior (EAP) method of estimation, which takes into account the expected distribution of the population (a normal distribution). This allows cognitive ability estimates to be calculated even in the case of all-correct or all-incorrect responses, which would have an infinite standard error if simple maximum likelihood estimation were used.

Psychometric properties: In IRT, reliability is judged by examining the test information functions. As shown in figure 1.2, each item has an information function (IIF) which is a plot of the candidate's cognitive ability against the probability of a correct response given the candidate's cognitive ability. The test information function (TIF) is the examinee's cognitive ability against the sum of IIFs. As stated in the AdaptGRT technical manual, TIFs show which cognitive ability levels are being measured well and which levels are not measured as well by the range of items (Chernyshenko, et al., 2008, p. 25). The TIFs for the AdaptGRT subtests show high information and low standard errors of measurement around the middle ranges of cognitive abilities, that being the range in which the largest proportion of a given population is likely to score. The TIF is lower and standard error rates higher further away from the mean, with above average cognitive ability levels in particular being more susceptible to error. This is most likely due to there being fewer items available to test the higher ability ranges, especially in the numeric and abstract reasoning subtests. Despite this weakness around higher ability levels, standard error rates did not exceed .7 (theta [θ] ranging from -3 to +3, mean 0 and SD 1), which is not too high. This is likely to be an area where refinements have been made since the test's inception and publication of the technical manual, with more difficult items being included in order to

increase information around higher cognitive ability estimates while decreasing the standard error.

Score reports: The score reports for the AdaptGRT have similar information to the IRT2 reports, giving the candidate's percentile rank on each subtest, but AdaptGRT reports use stanines rather than stens, meaning cognitive ability is ranked from 1 to 9 rather than 1 to 10. The AdaptGRT technical manual offers a table with a detailed list of cognitive ability estimates found during pretesting, and how these converted into stanine scores and percentile ranks. The frequency distributions of scores closely approximated the standard normal distribution.

Because the AdaptGRT is a relatively new test with regular revisions and refinements, it has not yet been independently reviewed. Its psychometric properties appear sound, however, with good data-model fit and low standard errors.

Though the IRT2 and AdaptGRT test procedures differ, the domains of cognitive ability are reliably and validly estimated by both tests.

2.2.3 Performance appraisal ratings

The performance appraisal rating instrument was developed in-house by the participating organisation, and there was no formal analysis of the psychometric properties of the instrument available.

Context of the performance appraisal ratings: The organisation sets specific task objectives quarterly for each individual employee, each work team, and the organisation as a whole. During each quarterly meeting to set objectives, managers and employees discuss how well the set objectives are being met. Once a year, managers rate the employees on how well they have achieved their individual objectives and how well their teams achieved their objectives during the past year, as well as giving individuals a

ranking for how well they demonstrated the values of the organisation. The values of the organisation are (1) honouring people as people, (2) valuing the freedom to succeed, and (3) taking pride in the organisation. Individual performance is interpreted as task performance and 'values demonstration' is interpreted as contextual performance (Borman & Motowidlo, 1993).

Rating categories: Possible ratings of individual performance were: 'Good Performance', meaning the employee met all their objectives to a satisfactorily high level; 'Room for Improvement', meaning some objectives either were not met or fell below ideal standards; and 'Unacceptable', meaning that either too few (if any) objectives were met, or that none met acceptable standards.

Team performance level was rated as either: 'Exceeded', meaning all set objectives were met to a very high standard, exceeding expectations; 'Met', meaning all objectives were met to a satisfactory standard; or 'Partially Met', meaning that some objectives were either not met or were not of a satisfactory standard.

Values demonstration was rated as: 'High', meaning the employee consistently 'championed' the values of the organisation; 'Medium', meaning that values were consistently demonstrated to a required level; or 'Low', meaning that demonstration of values was below that expected of an organisation representative (employee). The three levels for each type of performance rating were coded as 1, 2 and 3 for the lowest, middle and highest rating, respectively[33].

Each employee's overall performance score consisted of the following proportions: 10% of the score is based on the organisation achieving its objectives; 40% on whether employees have met their individual objectives; 20% on the employee's team meeting its objectives, and 30% on the employee demonstrating organisational

[33] Transforming codes 1, 2, and 3 to 0, 1 and 2, respectively, did not lead to any difference in substantive results because of the linear transformation.

values. In this study, whether organisational objectives had been achieved was kept confidential, so was scored as 10/10 for all participants. Job performance ratings were expressed as proportions; that is, a rating of three would be 3 out of 3, so individual performance rating of 1, combined with values demonstration 2 and team performance level 2 would equal an overall performance score of 57% (rounded from 56.667).

2.2.4 Organisation's use of instruments

The organisation uses the IRT2 and the AdaptGRT regularly as part of their selection process. It is important to note that there is no cut-off score for selection; that is, no minimum score which would lead directly to a candidate's elimination. The organisation uses cognitive ability scores as a piece of information about the candidate in much the same way that personality test results, academic achievements, previous work experience, and previous employer references provide information about a candidate. All information is cross-referenced and examined for consistency. For example, cognitive ability scores should relate to academic achievements, strengths or weaknesses in practical problem solving ability may be described by former supervisors, and situational judgement tests during interviews provide an opportunity for cross-referencing cognitive ability and personality test results (McDaniel, Hartman, Whetzel, & Grubb, 2007; McDaniel & Whetzel, 2007). Cross-referencing information is important for supporting the validity of inferences made about candidates based on their cognitive ability scores (Binning & Barrett, 1989; Putka & Sackett, 2010).

Given the nature of the appraisal procedure, job performance ratings are likely to be relatively reliable. Setting specific objectives and rating how well these objectives are met combines objective and subjective evaluation, which is assumed to correct some

rater biases, while allowing situational factors beyond an employee's control to be taken into account (see section 1.3.3; Pulakos & O'Leary, 2010).

The ratings are an integral part of the organisation's performance management system, so the aim is for each employee and team to accomplish as much as they are capable of on behalf of the organisation (Pulakos, 2007). The regular discussions to set and evaluate objectives means employees should not be in any doubt about their likely annual rating, or what changes they may need to make in their work behaviour to achieve higher ratings. Bonus pay and reward systems (such as promotion opportunities) are based to a great extent on these job performance ratings, which is likely to motivate employees to seek higher ratings (Locke & Latham, 2002). Because of the organisational context, performance ratings are expected to cluster around higher levels.

The group of employees participating in this study were hired by the organisation a maximum of 9 months prior to receiving their ratings, meaning that job performance ratings do not reflect a whole year's employment. Those in the job for longer may have had a better idea of what behaviours would lead to higher ratings because of the regular meetings.

2.3 Variables

Table 2.1 provides a list of the original variables, as well as variables that were derived from the originals for the purposes of further analysis. They are organised in categories relating to participant characteristics, cognitive ability percentile ranks and job performance ratings. Variable labels described here are used for the remainder of this section and throughout section 3. Possible values for each variable are included.

Levels of measurement accord with descriptions by Nunnally and Bernstein (1994), and Netemeyer, Bearden and Sharma (2003).

Table 2.1: Research variables

Category	**Variable name and description**	**Label**	**Possible values**	**Level of measurement**
Participant characteristics	Gender of participant	GEN	0, 1 (M, F respectively)	Ordinal
	Age of participant (in years)	AGE	Any (numerical)	Ratio
	Length of service with organisation (given as the proportion of a year)	LOS	0.01-1.0	Interval
	Ethnicity of participant	ETH	Any	Nominal
Cognitive ability scores	AdaptGRT Verbal reasoning percentile rank	ADVR	0-100%	Interval
	AdaptGRT Numeric reasoning percentile rank	ADNR	0-100%	Interval
	AdaptGRT Abstract reasoning percentile rank	ADAR	0-100%	Interval
	IRT2 Verbal percentile score rank	IRVR	0-100%	Interval
	IRT2 Numeric reasoning percentile rank	IRNR	0-100%	Interval
	IRT2 Abstract reasoning percentile rank	IRAB	0-100%	Interval
Job performance ratings	Individual performance rating	IPR	1, 2, 3	Ordinal/Interval*
	Team performance level	TPL	1, 2, 3	Ordinal/Interval*
	Values demonstration	VAD	1, 2, 3	Ordinal/Interval*
Derived variables	Verbal reasoning^	VR	0-100%	Interval
	Numeric reasoning^	NR	0-100%	Interval
	Abstract reasoning^	AR	0-100%	Interval
	Overall performance[#]	OP	40-100%	Interval

* Job performance ratings are assumed to be interval, meaning the difference between each consecutive rating was equivalent. Conceptually this assumption may be incorrect, so the variables were analysed as both ordinal and interval levels of measurement in order to compare results and check substantive conclusions.

^ Test labels were removed because of the assumption that abilities were measured consistently despite the different measurement models.

[#] Overall performance was calculated from job performance variables, as expressed in section 2.2.3: OP = 10 + 40IPR/3 + 20TPL/3 + 30VAD/3.

2.4 Method of data collection

This research was conducted on data from a field setting. The researcher sought permission from OPRA Consulting Ltd and Psytech International Ltd for access to the testing platform for the cognitive ability tests, technical manuals (which are publicly available), testing procedures and reviews.

A director from OPRA approached client organisations that used the tests, inviting them to participate in the research. Contact details for Human Resources (HR) managers of potentially interested organisations were then passed to the researcher. The researcher contacted the HR managers with specific details of the research including what was under investigation and why, what data were required, and ethical arrangements. Only one organisation chose to participate. The HR manager of the organisation compiled the variables listed in section 2.3 in a Microsoft Excel spreadsheet, which was then emailed to the researcher. The HR manager also provided details of their selection and performance appraisal procedures as required by the researcher. Data were transferred to SPSS for analysis.

2.5 Research design

Data were from the organisation's records, so there was no specified sampling procedure. The organisations who were asked to participate were not randomly selected, and there was only one organisation out of those approached who deemed it appropriate to release cognitive ability and performance appraisal data. Similarly, the group of participants was not a random sample of employees from the participating organisation, but rather only those employees who agreed to allow their cognitive ability scores and performance appraisal ratings to be collated for the purposes of this research. While an ideal minimum number of participants were sought, the actual data collected represents

the maximum amount of data available to the researcher within the time limit for the research. Implications of the lack of sampling strategy will be outlined in section 4.

2.6 Data analysis and hypotheses

Initial analyses were conducted to examine features of variables such as means and distributions, and also to evaluate whether it was appropriate to include variables in subsequent analyses.

Preliminary analyses included (1) number of cases for each variable, to assess missing data; (2) frequency of nominal variables, to examine group sizes; (3) frequency of job performance ratings, to examine distributions; and (4) minimums, maximums, means and standard deviations for age, length of service, cognitive ability percentile ranks and overall performance, in order to investigate variable distributions.

Substantive hypotheses required Pearson correlation analyses to investigate systematic linear relationships between variables[34]. Job performance ratings could be considered ordinal rather than interval level of measurement, so mean differences in cognitive abilities at each level of job performance rating were tested with ANOVA.

[34] A contributor suggested analysing relationships using an index of absolute similarity (Gower, 1971). There was limited benefit in examining the results this way because Pearson correlations were the best procedure for answering the research questions, so the method, results and a brief discussion of this analysis have been relegated to appendix four.

3 Results

The data for 43 participants, provided by the HR manager of a New Zealand government organisation, were received in an Excel spreadsheet email attachment and were transferred to SPSS v17.0 for analyses.

3.1 Univariate variable analyses and evaluation of adequacy

Descriptive statistics were calculated and missing data analysed to evaluate the adequacy of each variable for use in subsequent analyses, and to take note of data features that would require caution in interpretation.

Frequencies were examined to compare group sizes on gender, ethnicity, and job performance ratings values. Means and distribution statistics for age, length of service, cognitive ability scores and job performance ratings were also examined.

Participant characteristics were analysed for relationships with cognitive ability scores and job performance ratings, in case they needed to be taken into account in relation to the hypotheses.

Power was estimated prior to correlation analysis. Based on values given in Cohen (1988), with 43 participants and alpha for significance set at .05, only correlations greater than roughly .40 could be detected with at least 80% power. Lack of power was problematic during analysis, as correlations below .25 were not significant, correlations between .25 and .3 were approaching significance ($.05 < p < .1$), and correlations above .3 were significant at $p < .05$ (see table 3.7). Non-significance was therefore interpreted with caution, especially for correlations between .2 and .3. Correlations smaller than around .2 were unlikely to indicate a relationship of practical importance (Cohen, 1988; Murphy, Myors, & Wolach, 2009).

3.1.1 Participant characteristics

There were 26 women and 17 men. The unequal group sizes meant that subsequent analyses involving gender differences needed to be interpreted with caution.

Participants were from diverse ethnic groups and nationalities: the majority (26 participants) were New Zealand European (Pākehā), 4 were Samoan, 3 Māori, 3 British, 1 Indian, 1 Malaysian, 1 Australian, 1 North American, and 3 participants who did not state their ethnicity. Group sizes were too different to allow further analysis. Combining groups was inappropriate as group sizes would remain unequal, and group labels could be inaccurate or misleading[35].

Participant ages ranged from 20 to 52 years, the mean age of the group being 33.72, standard deviation (SD) 7.7 years. The minimum amount of time a participant had been employed by the organisation (length of service) was roughly 4 months, maximum time with the organisation 9 months, mean 6 months and SD around 2 months. See table 3.1.

Table 3.1 Participant's ages and length of service

Variable [a]	N	Min	Max	Mean	SD
AGE	43	20	52	33.37	7.70
LOS	43	0.34	0.75	0.49	0.12

[a] See section 2.3 for variable descriptions, labels, and possible values.

[35] Combining groups means deciding which groups are more similar, which is likely to reflect a judgement rather than a fact. For example, participants who have one Māori grandparent and three Pākehā grandparents could meaningfully select 'Māori', 'New Zealand European' or both as their ethnicity. An English immigrant who has obtained New Zealand citizenship could meaningfully select 'New Zealand European' or 'British'. Any combination could be flawed, leading to potentially invalid interpretations and generalisations.

3.1.2 Cognitive ability test scores

Participants took different cognitive ability tests during their selection processes. There were also some missing scores, indicating that some participants did not sit all of the subtests.

Twenty-six participants had taken all three IRT2 subtests, 2 had taken only the verbal and abstract reasoning IRT2 subtests. Thirteen participants had taken all three AdaptGRT subtests, 2 had taken only the verbal reasoning AdaptGRT subtest. Cognitive ability variables were relabelled by the reasoning ability tested, because the IRT2 and AdaptGRT tests provided reliable and valid scores of the same constructs so scores were considered comparable (to be discussed further in section 4.3). Relabeling resulted in verbal reasoning percentile rank scores[36] for all 43 participants, numeric reasoning scores for 39 participants and abstract reasoning scores for 41 participants. Missing data were excluded pairwise during correlation analysis.

Participant's verbal reasoning scores showed a negatively skewed distribution (see table 3.2), mostly because the minimum score on the AdaptGRT was 53%. This could indicate range restriction, which may have occurred if the organisation's selection procedure favoured candidates with higher verbal reasoning ability. Verbal ability scores were found at minimal levels, so data were not fully range restricted, and there is no evidence of a minimum verbal reasoning score.

If verbal reasoning ability is normally distributed in the population as is assumed, these data may not be representative of the population. Caution is required in generalising results. Skewness can also reduce the size of Pearson correlations.

[36] Henceforth 'percentile rank scores' will be referred to as 'scores' to reduce wordiness. Percentile rank scores will also be symbolised with % for simplicity.

Table 3.2: Cognitive ability score distributions

Variable	N	Min	Max	Mean	SD	Skewness	Skew SE	Kurtosis	Kurt SE
ADVR	15	53%	99%	84.67%	12.57%	-1.01	.58	1.45	1.12
ADNR	13	9%	99%	57.38%	31.21%	-.43	.62	-1.16	1.19
ADAR	13	14%	99%	55.46%	32.19%	.01	.62	-1.70	1.19
IRVR	28	2%	99%	44.36%	28.35%	.15	.44	-1.08	.86
IRNR	26	2%	91%	39.23%	24.98%	.44	.46	-.52	.89
IRAR	28	14%	99%	56.79%	22.0%	-.19	.44	-.20	.86
VR	43	2%	99%	58.42%	30.78%	-0.42	0.36	-1.12	0.71
NR	39	2%	99%	45.28%	28.17%	0.23	0.38	-1.06	0.74
AR	41	14%	99%	56.37%	25.30%	-0.10	0.37	-0.86	0.72

3.1.3 Job performance ratings

There were only three possible ratings for individual performance rating, team performance level and values demonstration, and ratings could be interpreted as ordinal or interval level of measurement (see table 2.1). Therefore frequencies of each rating and combination of ratings were examined (tables 3.3 and 3.4), as well as means, standard deviations, skewness and kurtosis for each type of job performance (table 3.5).

Table 3.3: Number and percent of participants receiving different levels of performance ratings ($N = 43$)

Rating	IPR		TPL		VAD	
	N	**%**	**N**	**%**	**N**	**%**
1	1	2.3	1	2.3	2	4.7
2	8	18.6	14	32.6	39	90.7
3	34	79.1	28	65.1	2	4.7

Individual performance ratings and team performance levels were clustered around higher ratings as expected. Values demonstration is not related to specific objectives, and scores showed a different pattern, with 90% of participants receiving a rating of 2.

Out of 27 possible score combinations[37], only 8 actually appeared in the data set, and 5 of those had only one participant receiving that combination. The combinations of values are shown in table 3.4.

Table 3.4: Number and percent of participants receiving each combination of job performance ratings, and corresponding overall performance percent*

IPR	VAD	TPL	OP	Number of participants	% of participants
1	1	3	53%	1	2%
2	2	1	63%	1	2%
		3	77%	7	16%
3	1	3	80%	1	2%
	2	2	83%	13	30%
		3	90%	18	42%
	3	2	93%	1	2%
		3	100%	1	2%

*For clarity and brevity, only the 8 combinations of ratings that appeared in the data set are included in this table.

Two combinations of scores were particularly common: 3 for individual performance rating, 2 for values demonstration and either 2 or 3 for team performance level, with 31 participants (72%) receiving one of these two combinations of ratings. Though expected, given the context of the performance appraisal, these frequencies will need to be considered in subsequent interpretations.

Descriptive statistics in Table 3.5 echo the frequency results. Means for individual performance rating and team performance level are quite high. The negative skewness

[37] Three possible values for IPR, TPL and VAD, meaning 3 x 3 x 3 = 27 combinations

in individual performance ratings and team performance levels were explained by the high frequency of high ratings, while the negative skew for overall performance ratings was explained by the high frequency of two combinations of ratings. Values demonstration ratings showed an extreme central tendency (leptokurtic distribution around the mean).

Table 3.5: Distribution statistics for job performance ratings

Variable	N	Min	Max	Mean	SD	Skewness	Skew. SE	Kurtosis	Kurt. SE
IPR	43	1	3	2.77	.48	-1.97	.36	3.31	.71
TPL	43	1	3	2.63	.54	-1.03	.36	.02	.71
VAD	43	1	3	2.00	.31	.00	.36	8.89	.71
OP	43	53.33%	100%	84.42%	8.09%	-1.66	.36	4.72	.71

Individual and team performance ratings would not be expected to be normally distributed given the context of performance appraisals. The organisation is unlikely to tolerate continued low scoring of individual performance and team performance in particular, meaning that individuals scoring at this level may receive warnings that performance must improve for employment with the organisation to continue. Though statistically the distribution could be interpreted as range restricted, in practice a normal distribution of scores is unlikely.

The high frequency of ‘medium’ ratings on values demonstration suggests that this may be a default score, with ‘high’ or ‘low’ ratings being the result of specific, observed deviation from expected performance. Because there were only two data points each at the higher and lower levels, these groups could not be adequately compared statistically.

3.1.4 Participant characteristics and cognitive ability scores

Men scored higher than women on both numeric and abstract reasoning tests (see table 3; $F = 6.872$, $p = .012$ for numeric reasoning subtest scores; $F = 6.518$, $p = .015$ for abstract reasoning subtest scores).

Gender differences in cognitive abilities are common (Ones, et al., 2005), and these results should not be problematic for further analyses, but will be taken into account.

Table 3.6: Means and standard deviations of cognitive ability scores for men ($N = 17$) and women ($N = 26$).

	Men		**Women**	
	Mean	**SD**	**Mean**	**SD**
VR	67.06	27.781	52.77	31.831
NR	59.20	29.747	36.58	23.805
AR	68.19	21.467	48.80	25.025

There were no other significant relationships between cognitive abilities and age or length of service, thus no other results that might be problematic.

3.1.5 Participant characteristics and job performance variables

Mean scores on performance dimensions were not examined because men's and women's variances were significantly different on these ratings, and there were 1.5 times more women than men. The difference between overall performance ratings approached significance (Men's mean 80.59, SD 9.17; women's mean 74.04, SD 13.27; $F = 3.147$, $p = .083$), suggesting that men might score higher than women on average, but even if low power contributed to the lack of significance, the difference was small enough to be negligible.

There was a possibility that length of service would correlate significantly with job performance ratings, because the participants who had been in their jobs longer may have had greater opportunity to learn what aspects of job performance were valued most highly (section 2.2.3). The findings showed that there was no significant relationship between length of service and job performance ratings, however (table 3.7).

Table 3.7: Pearson correlations for non-nominal participant characteristics, cognitive ability scores and job performance ratings

	AGE	LOS	VR	NR	AR	IPR	TPL	VAD
VR	-.15	.01	-					
NR	.11	.01	.38*	-				
AR	-.08	-.01	.24	.36*	-			
IPR	.01	.05	.44**	.24	.20	-		
TPL	-.15	.09	-.11	-.15	-.10	-.16	-	
VAD	.07	-.27^	.04	.09	-.05	.32*	-.14	-
OP	-.03	-.02	.31*	.16	.10	.84**	.26^	.57**

^ p<.1 * p<.05 ** p<.01 (two-tailed)

The only correlation approaching significance showed the opposite effect, with the negative correlation between length of service and values demonstration suggesting that more recently hired participants tended to receive higher ratings of values demonstration (r = -.27; p = .076).

From an inspection of the scatter plot, this correlation appears to be influenced by the length of service of four participants: the 2 participants who received a rating of 'low' on values demonstration (length of service = 0.69 and 0.71) compared to the length of service of the 2 participants who received a rating of 'high' (length of service = 0.44 and 0.54). The 39 remaining participants, who received a 'medium' rating on values demonstration, had a mean length of service of 0.48, values ranging from 0.34 to

0.75. Because only a small proportion of cases appeared to strongly influence this correlation, it was not considered trustworthy. Length of service was therefore considered not to have a strong relationship with either cognitive ability scores or job performance ratings.

3.2 Hypotheses

Hypotheses predicted systematic relationships (or lack of) between cognitive ability scores and job performance ratings, so Pearson product-moment correlation analyses were used to examine the strength and direction of linear relationships. See table 3.7 for correlations.

3.2.1 Relationships among cognitive ability subtest scores

The first hypothesis stated that cognitive abilities would be related to each other, so Pearson correlations between verbal, numeric and abstract reasoning scores were analysed (table 3.7). Significant positive correlations were found between verbal and numeric reasoning ability scores (r = .38, p = .018) and between numeric and abstract reasoning scores (r = .36, p = .023) but not between verbal and abstract reasoning scores (r = .24, p = .128). The lack of a significant result for this latter pair of abilities was most likely due to lack of power (Cohen, 1988). The hypothesis was partially supported.

3.2.2 Relationships among job performance dimensions

The second hypothesis stated that task and contextual performance would be related to each other. Correlation analysis supported this hypothesis, with a significant positive correlation between individual performance ratings and values demonstration (r = .32, p = .036).

The third hypothesis stated that team performance would not be related to either individual task or contextual performance. This hypothesis was also supported as the correlation between team performance level and individual performance rating was not significant (r = -.16, p = .307), neither was the correlation between team performance level and values demonstration (r = -.14, p = .357). Caution needs to be exercised in accepting the null hypothesis due to low power (1 - β = .17), though these correlations are quite small, and not likely to be of practical significance.

It was conceptually possible that job performance ratings were ordinal level of measurement as opposed to interval, which might indicate that Pearson correlations are not appropriate. However, because there are three possible levels of rating for all three job performance types, systematic relationships among rating levels are adequately indexed with a correlation coefficient. ANOVA is not appropriate for examining associations among ratings because of the lack of variance in rating levels (for example, every participant who scored a 2 for individual performance rating also scored a 2 for values demonstration).

3.2.3 Relationships between cognitive abilities and job performance ratings

Team performance did not correlate significantly with any of the ability scores, as expected. As with task and contextual performance, any systematic relationship between cognitive abilities and team performance is likely to be moderated by many other factors.

Abstract reasoning scores did not show significant correlations with individual performance ratings or values demonstration, which was contrary to expectations. There was a positive correlation between abstract reasoning scores and individual performance

ratings, but this was not significant. Low power may have contributed to the lack of significance in this case.

Numeric reasoning was not expected to relate to job performance ratings and there were no significant correlations, supporting this expectation. Like abstract reasoning scores, however, numeric reasoning scores showed a non-significant positive correlation with individual performance rating, which may have become significant had there been greater power.

It is unlikely that the gender differences found in mean numeric and abstract reasoning scores influenced these correlations, because there were no large or significant associations between gender and job performance ratings.

Verbal reasoning was hypothesised to relate to task and contextual performance. This hypothesis was partially supported, with verbal reasoning scores significantly positively correlating with individual performance ratings ($r = .44$, $p = .003$). Verbal reasoning scores did not correlate with values demonstration.

Verbal reasoning showed a significant positive correlation with overall performance ($r = .31$, $p = .04$), though this may have been mediated by individual performance rating, as there was a large correlation between individual performance rating and overall performance ($r = .84$, $p < .001$). There were not enough data to test this proposition statistically (Baron & Kenny, 1986).

There was a possibility that job performance ratings could be interpreted as ordinal level of measurement rather than interval, which may lead to non-linear relationships between cognitive abilities and job performance ratings. Non-linearity would reduce the size of Pearson correlations. Thus, the hypotheses about the associations of cognitive abilities to different types of job performance were also tested using ANOVA.

Mean differences in cognitive ability scores at different levels of performance rating (1, 2 or 3) are shown in table 3.8. Individual performance rating and team performance level each had only one participant receiving a rating of one, so these data could not be included in the ANOVA.

Table 3.8: Mean reasoning ability scores by job performance ratings

		N	VR		NR		AR	
			Mean	SD	Mean	SD	Mean	SD
IPR	1	1	53%	*	18%	*	26%	*
	2	8	26% [a]	26%	36%	28%	52%	18%
	3	34	66% [a]	27%	49%	28%	59%	27%
TPL	1	1	27%	*	72%	*	56%	*
	2	14	68%	32%	48%	27%	61%	29%
	3	28	54%	30%	43%	29%	54%	24%
VAD	1	2	67%	19%	42%	33%	63%	52%
	2	39	57%	31%	45%	29%	56%	24%
	3	2	74%	36%	65%	*	54%	52%

[*] Only one participant meant there was no standard deviation in these cells

[a] Indicates values where difference was found to be significant

Mean differences in numeric and abstract reasoning scores at different levels of performance ratings were generally small, and the differences were not statistically significant. This supports a conclusion that higher numeric and abstract reasoning scores are not strongly associated with job performance ratings. There was a trend of increasing mean numeric and abstract reasoning scores as levels of individual performance ratings increased, but the mean differences do not indicate an association of practical significance. Thus, abstract reasoning did not have the hypothesised positive

relationship with job performance ratings, but numeric reasoning did demonstrate the expected lack of association with job performance ratings.

Significant differences were found between the mean verbal reasoning scores for those scoring a 2 compared to a 3 on individual performance rating (mean scores 26% and 66%, respectively, $F = 12.88$, $p < .001$). Given the large differences in group sizes, Levene's test for equality of variances was examined, showing no significant difference in variances. The hypothesis that greater verbal reasoning is associated with higher ratings of task performance was supported.

3.3 Summary of key results

3.3.1 Results indicating caution is required in interpretation

The small number of participants contributed to low power to detect significant relationships, meaning that non-significance in particular needed to be interpreted with caution.

Participant's scores on cognitive ability tests were treated as equivalent despite some participants taking the IRT2 and others taking the AdaptGRT. If the assumption that scores are equivalent is incorrect, interpretations may also be flawed. Missing numeric and abstract subtest scores were also regarded as missing at random, but if they were missing because they were very low, for example, this may have influenced results, and thus the validity of inferences.

Verbal reasoning ability scores showed large negative skewness, with scores clustering around higher percentiles. Scores in the population are expected to be normally distributed, and IRT2 and AdaptGRT pretesting during test development indicated that scores were normally distributed (sections 2.2.1 and 2.2.2). The negative skewness therefore indicates that participants may not be representative of the

population, and generalisations should be tempered accordingly. In addition, though the statistical procedures are robust to some departure from normality, skewness may have reduced the sizes of correlations and mean differences.

Individual and team job performance ratings showed expected clustering around higher levels. Statistically this could be interpreted as partial range restriction, though practically it is unlikely that a normal distribution of scores can exist, as employees would be expected to either enhance their performance or leave the organisation. Results nevertheless require cautious interpretation.

The high proportion of participants receiving a rating of 'medium' on values demonstration could be a result of a default rating of medium for all employees unless their supervisors observe or hear about them acting in some way warranting a higher or lower rating. This would make a comparison of employees receiving 'high' and 'low' ratings interesting, but too few participants receiving these ratings in this data set precluded statistical comparison. There was an indication that more recently hired participants received a higher rating of values demonstration, for example, but conclusions cannot be responsibly drawn on such a small number of data.

There were mean gender differences in numeric and abstract reasoning abilities, but these did not appear to influence relationships among these abilities and job performance ratings. The small number of participants ruled out a regression analysis to test this assumption.

3.3.2 Results relating to hypotheses

The first hypothesis stated that cognitive abilities would be positively related to each other. This was partially supported, with significant correlations between verbal reasoning and numeric reasoning scores, as well as numeric reasoning and abstract

reasoning scores. The lack of significance of the positive correlation between verbal reasoning and abstract reasoning scores may have been contributed to by low power as a result of small participant group size and skewness in the verbal reasoning scores.

The second hypothesis stated that task performance and contextual performance would be related to each other. This was supported, with evidence of a positive correlation between individual performance ratings and values demonstration.

Team performance was not expected to relate to individual task or contextual performance, and this hypothesis was supported with small, non-significant negative correlations between team performance level and individual performance rating, and between team performance level and values demonstration.

Team performance was also not expected to relate to cognitive ability. This was the case, with only small, non-significant negative correlations between team performance and verbal, numeric and abstract reasoning scores.

The fifth hypothesis anticipated positive correlations between abstract reasoning and task and contextual performance. No significant relationships were found between abstract reasoning and individual performance rating or values demonstration, however. The positive correlation between abstract reasoning scores and individual performance rating may have been non-significant due to lack of power, but mean differences in abstract reasoning scores at different levels of individual performance rating were also small and non-significant.

Numeric reasoning was not expected to correlate significantly with task or contextual performance, which was the case. As with abstract reasoning, numeric reasoning showed a non-significant positive correlation with individual performance rating, but this relationship was not strong enough to result in statistical significance or practically important mean differences.

The seventh hypothesis expected verbal reasoning to relate to task and contextual performance. This was partially supported, because verbal reasoning scores did not show a significant association with values demonstration ratings. However, verbal reasoning showed a strong positive correlation with individual task performance, corroborated by a large mean difference in verbal reasoning scores for participants scoring a 2 compared to a 3 on individual performance rating.

Discussion

This study investigated relationships between cognitive ability and job performance. The cognitive ability scores were percentile rankings of 43 job candidates, undergoing testing as part of a selection procedure, who were subsequently employed by the participating organisation. Job performance ratings were from the employees' first annual performance appraisal.

4.1 Key findings

4.1.1 Cognitive abilities

The first hypothesis stated that verbal, numeric and abstract reasoning scores would be positively correlated with each other. This hypothesis was partially supported, with numeric reasoning scores correlating positively with verbal and abstract reasoning scores. Verbal reasoning scores did not have a statistically significant positive correlation with abstract reasoning scores. The lack of significance was probably due to low power, as a result of the small number of participants and a skewed distribution of verbal reasoning scores. The overall pattern of relationships between cognitive abilities roughly accorded with previous research (see sections 1.2.4 and 1.2.5; Carroll, 1993; Horn & Cattell, 1966; Jensen, 1998; Spearman, 1904; van der Maas, et al., 2006).

Women scored lower on average compared to men on numeric and abstract reasoning, which accorded with previous research on cognitive ability in general, as well as research relating specifically to the cognitive ability instruments (Budd, 1993; Chernyshenko, 2005; Ones, et al., 2005; Ree, Carretta, & Steindl, 2002). There were not enough participants to examine whether the gender differences affected relationships between cognitive abilities and job performance ratings.

4.1.2 Job performance ratings

It was hypothesised that ratings of task and contextual performance would be related, because performance processes such as communicating are likely to influence both task and contextual performance (see section 1.3.2). This was the case, with individual performance ratings correlating positively with values demonstration.

It seems unlikely that task and contextual performance are mutually exclusive. The theory explaining the relationship here is that dimensions of performance processes may be common to task and contextual performance. There are three other explanations for the relationship. Firstly, some aspects of personality and motivation may influence task and contextual performance to a similar degree. For example, conscientious employees would probably work hard at all aspects of their jobs (Barrick & Mount, 1991; Borman & Motowidlo, 1993). In this study, accomplishing objectives and demonstrating the values of the organisation were task and contextual elements pertaining to all employees' jobs, and a conscientious employee would presumably wish to fulfil specified expectations to a high standard (J. Hogan & Holland, 2003). Secondly, task and contextual performance lend themselves to reciprocity (Kiker & Motowidlo, 1999). For example, one of the organisation's values was to take pride in the work of organisation. An employee who demonstrated pride would probably have a more positive perception of the work, which would motivate them to accomplish more objectives to a higher standard. The third theory is that good or bad task or contextual performance may affect managers' perceptions of employees, leading to halo effects in ratings (see section 1.3.3; Feldman, 1981; Landy & Farr, 1980). All four theories may explain some element of the observed relationship (Murphy & Cleveland, 1995), though because the correlation was not large (Cohen, 1992), there is still evidence that task and

contextual performance are distinct from one another (c.f. Motowidlo & Van Scotter, 1994).

Individual-level influences on team performance were expected to be obscured by team dynamics, meaning that systematic relationships between individual task and contextual performance and team performance would not be found (see section 1.3.2). This was the case, as team performance level did not show significant correlations with either individual performance ratings or values demonstration. Essentially, these results mean that it cannot be assumed that because an individual is accomplishing their own objectives that their team will be similarly productive. But it also cannot be assumed that unproductive members decrease the productivity of their teams. This result is important simply as an indication that more factors need to be taken into account for examining influences on team performance, which accords with previous research (Haslam, 2001; Mohammed, et al., 2010).

There is not a lot of published research that explicitly compares different levels of job performance in organisations like they were compared here, a state of research that has recently been criticised (Wildman, Bedwell, Salas, & Smith-Jentsch, 2011). Hierarchical analysis is a better way of studying the effects of individual performance on team performance compared to the simple correlation analyses in this research, as hierarchical analysis examines the performance ratings of individual team members to the team's collective performance rating (e.g. LePine, et al., 1997). The methodology of this research unfortunately did not allow exploration of hierarchical relationships. Suggesting that individual task and contextual performance be compared to team performance hierarchically is not an original idea (see Motowidlo, 2003), but it has not been taken up very often (Wildman, et al., 2011). Research has examined individual differences assumed to relate to performance (e.g. LePine, et al., 1997; Morgeson,

Reider, et al., 2005), team dynamics that influence team performance (e.g. De Dreu, et al., 1999; Somech, et al., 2009), and individual performance alone and within a team (Glew, 2009; Sonnentag & Volmer, 2010), but does not tend to specifically analyse team members' individual performances compared to their collective performance.

The majority of individual and team performance ratings were at the maximum level. This was expected, because ratings were the result of an ongoing performance management system (see section 1.4.4). The organisation manages performance by setting objectives for individuals and teams. Employees and managers meet quarterly to discuss how well previous objectives have been met, and to set new objectives. This system fits the criteria for an effective performance management system (see section 1.3.3; Pulakos & O'Leary, 2010; Smither & London, 2009), as employees and teams have the opportunity and incentive to increase their ratings by completing more objectives to a higher standard over time (Blumberg & Pringle, 1982; Locke & Latham, 2002). The clustering of higher ratings suggests that the system in the participating organisation is effective for motivating employees, ultimately resulting in high performance ratings.

It is unknown whether feedback on values demonstration was offered during the regular performance management meetings, though it appears unlikely. Over 90% of participants received a rating of 'medium' on this dimension, which may indicate that this is a default rating, while 'high' and 'low' ratings are the result of some observed deviation from perceived 'ordinary' behaviour. If this suggestion is correct then values demonstration may be a function of the raters' opportunities to observe, their beliefs and biases to a greater degree than individual or team performance (see section 1.3.3; DeNisi, 1996). Though rater-specific effects do not necessarily constitute error (because

performance is in part defined by the judgement of it) these are issues for raters to be aware of.

4.1.3 Cognitive abilities and job performance

The primary contribution of this research was to investigate the nature of relationships between specific cognitive abilities and specific dimensions of job performance.

As expected, verbal, numeric and abstract reasoning did not correlate significantly with team performance (see section 1.4.3). Individual-level traits and actions are conjectured to influence team dynamics, and because dynamics have complex relationships with team performance, there would be little or no evidence of direct systematic relationships between individual-level variables and team performance.

It was hypothesised that verbal and abstract abilities would relate to contextual performance, but neither scores for these abilities nor numeric reasoning scores correlated with contextual performance. Mean differences in ability scores at different levels of contextual performance ratings were also neither large nor statistically significant. This supports Motowidlo, Borman and Schmit's (1997) theory that cognitive abilities relate primarily to task performance (see section 1.4.2).

Abstract reasoning ability (or fluid intelligence) is associated with individuals' capacity for learning (Horn & Noll, 1997). Given that all of the participants were relatively new to the organisation, the longest length of service being 9 months, their abstract reasoning scores were expected to correlate with task performance because of the learning that would occur during the early stages of employment. This hypothesis was not supported, as abstract reasoning scores did not correlate significantly with task performance ratings.

One explanation for the lack of correlation is that the participants had experience in similar roles prior to being hired by the organisation (section 1.4.4). Deadrick and Madigan (1990) found that experience tended to be the best predictor of job performance in early months of employment, while general cognitive ability came to be a better predictor of job performance over longer-term employment. It is reasonable to assume that the participants had experience in similar jobs prior to employment with the organisation, which would moderate the effect of their fluid intelligence on job performance (Deadrick & Madigan, 1990). It is common for job advertisements to state that applicants require a number of years' experience (e.g. see www.seek.co.nz). If there are a number of applicants for a job, it is also common to screen out applicants with less previous experience (Taylor, et al., 2002). Thus, experience may have been a better predictor of individual performance during this stage of the participants' employment.

Another possible explanation for the lack of relationship between abstract reasoning scores and job performance regards fluid intelligence and ageing. As people age, the amount they need to learn to solve daily problems (need for fluid intelligence) decreases, while the amount they have learned (crystallised intelligence) about how to solve daily problems increases (Horn & Cattell, 1967). A similar theory is that the more an individual knows, the more connections can be made between new information and previously acquired knowledge (Piaget, 1961; Sternberg & Gardner, 1982), which means crystallised intelligence would increase as a function of itself, not just as a function of fluid intelligence. Both of these theories predict an increase in crystallised intelligence with age, and because the need for learning capacity decreases, fluid intelligence declines. Given the age range of the participants, and the test developer's assumption that test-takers would have acquired certain knowledge because of their anticipated age range and education levels (Psytech International Ltd, 2006), it is

possible that the abstract reasoning scores may have reflected the ability to utilise crystallised principles of logical reasoning rather than fluid learning capacity. Logical reasoning may not have been related to the daily demands of their jobs.

Numeric reasoning was not expected to show relationships with task performance because participants' jobs did not principally require proficiency in numeric operations. The lack of clear connection between numeric ability and contextual performance similarly meant there was no expected relationship. Both expectations were borne out in the results, as numeric reasoning scores did not significantly correlate with task or contextual performance ratings.

Abstract reasoning and numeric reasoning both had positive correlations with task performance ratings, but these correlations were not significant. It is possible that greater power would have resulted in significant effects. However, mean differences in numeric and abstract reasoning abilities at different levels of performance ratings also did not result in significant differences, and the differences that were evident were very small, so unlikely to be of practical significance.

Verbal reasoning scores, however, showed a significant positive correlation with individual performance ratings as hypothesised. Analysis of variance of mean verbal reasoning scores at different levels of individual performance rating corroborated this result, showing a higher mean verbal reasoning ability was associated with higher ratings of task performance. For this group of participants, therefore, it appears as though verbal reasoning was the key cognitive ability for task performance, explaining almost 20% of the variance in task performance ratings.

This result accords with one of Psytech's studies into the predictive validity of the general reasoning test (GRT2, see section 2.2.1).The technical manual for the GRT2 (Psytech International Ltd, 2006) describes a study of 39 training advisors working in

New Zealand, resulting in a positive correlation between verbal reasoning subtest scores and overall job performance, with no significant correlation for either abstract or numeric reasoning scores with job performance ratings. Psytech conducted a number of studies to assess the predictive validity of the test, but results generally showed inconsistent patterns of relationship between subtest scores and job performance ratings. Comparing this study's findings to Psytech's, it is possible that different abilities relate to job performance through task-relatedness.

Task-relatedness might explain the association of verbal reasoning and task performance in this study. The participating organisation in the current study was primarily a service organisation, meaning that there would be a high frequency of written and oral communication between employees and the public. Though caution must be exercised in interpreting a correlation as causation, it seems possible that employees with greater verbal reasoning ability may have had greater clarity and fluency of communication, thereby leading to a higher overall standard of task accomplishments.

The exact relationship between verbal reasoning ability and communication skills is somewhat elusive, though it is generally assumed that communication skills involve an interaction of verbal ability and social skills (e.g. Penley, Alexander, Jernigan, & Henwood, 1991; Sonnentag, 2000)[38]. Neither social skills nor communication skills are consistently defined constructs, so specific mechanisms for their relationship with verbal reasoning cannot be well elucidated. Despite the lack of consistent definitions or measurement, there is evidence suggesting that verbal reasoning ability, communication skills and social skills are related to each other and to job performance (S. T. Bell, 2007;

[38] The term used is 'social skills', to avoid using all the different terms in the cited literature (e.g. social competence, social skills, interpersonal skills, practical intelligence, emotional intelligence, social intelligence). This should not imply a consistent definition between research articles, but refers to a variable proficiency in interpersonal interaction, which is not necessarily a fixed trait.

Lievens & Chan, 2010; Morgeson, Reider, et al., 2005; Penley, et al., 1991; Schulte, Ree, & Carretta, 2004; Sonnentag, 2000; Sternberg & Hedlund, 2002).

If this theory is correct and verbal reasoning influenced task performance via communication skills, then it is unclear why verbal reasoning was not also related to contextual performance (as hypothesised), because communication and interaction processes were assumed to relate to task and contextual performance (section 1.3.2). This may suggest that either the theory linking verbal reasoning to performance via communication skills is incorrect, or the theory that communications skills are related to contextual performance is incorrect. Either way, a comparison of participants scoring 'high' or 'low' on values demonstration could have revealed more information about this performance dimension, but without more participants scoring at these levels this comparison was not viable.

4.2 Limitations

One clear limitation of this study was the small number of participants. It was unfortunate that this number could not have been at least doubled, but the data collected were the maximum available to the researcher. Consequences of the small number included reduced power to detect relationships, and some statistical procedures were also rendered unfeasible; for example, comparison of participants receiving 'high' and 'low' ratings of values demonstration, but also regression analysis and factor analysis.

Regression analysis of verbal reasoning subtest scores and individual performance ratings onto overall performance may have provided statistical evidence for the theorised mediated relationship (section 3.2.3). Factor analysis of cognitive ability scores could have allowed investigation of evidence for a general factor.

The small number of participants also necessitated combining IRT2 scores and AdaptGRT scores by subtest to allow meaningful statistical operations using cognitive ability scores. Though tests were judged to have adequate psychometric properties to validly combine scores in this way, it is a weakness of the methodology that this was necessary. If all participants had taken both tests then within-subjects comparisons of scores could have examined concurrent validity of the testing methods. Alternatively, a larger number of participants taking each test would allow between-subjects comparisons of cognitive ability and job performance coefficients. This latter investigation was a former objective of this research. It had originally been hypothesised that the greater precision of cognitive ability scores resulting from the AdaptGRT may lead to stronger correlations with job performance compared to those for IRT2 scores and job performance, but comparison of correlations would have required roughly 125 participants for each cognitive ability test for a result with adequate power to draw conclusions (Cohen, 1992).

Another limitation is that results cannot be assumed to generalise to other populations (Cook & Campbell, 1979). The group of participants cannot be classified as a scientific ‘sample’. The data represent available cases: a set of observations accessible due to the consent of one organisation and 43 employees who worked there. It is unknown how well this group of participants represents other employees in the organisation. It is also not realistic to assume that the participating organisation is necessarily similar in certain ways to other organisations. The results refer to these participants in this organisation, and it cannot be confidently concluded that the results are necessarily applicable to populations beyond this group.

In the current study, there were a number of potential reasons not to participate which would skew variable distributions in meaningful rather than random ways.

Firstly, the methodology called for participation of organisations using cognitive ability tests in their selection procedure. Distributions of cognitive ability scores could be expected to be different for employees of organisations that do not use cognitive ability tests in selection. In particular, verbal reasoning scores might be more normally distributed in a random sample than they were here. Secondly, while there were a number of organisations who used these instruments, there was only one that had job performance data they were willing to share. One explanation for this is that some organisations may not have had any job performance data. Anecdotal evidence suggests that some organisations intend to regularly appraise employees but do not. Because the performance ratings in this study were interpreted as reflecting a successful performance management system, different systems or a lack of a system would lead to different results. Finally, individuals receiving very low scores on either cognitive ability tests or performance appraisals may not be inclined to share those scores with more people than they are compelled to. This would result in higher average scores here than those that would be found in a random sample. Each of these potential influences means conclusions of this study need to avoid making generalisations without expressing the need for caution.

The inability of the researcher to collect a randomised sample of data does not indicate that the results are without value, however. Firstly, idealised goals of randomisation may not be realistically possible when collecting data in a field setting (Cook & Campbell, 1979; Locke, 1986). There is evidence that the findings are comparable to those of other studies, indicating a pattern of relationships that suggests commonalities among participant groups. Secondly, one of the main contributions of this research is the demonstrated benefit in examining specific cognitive abilities and specific dimensions of job performance. These are not expected to be congruent for all

groups because of the differences in job tasks, as well as the relative importance of task and contextual performance for overall performance. The research approach advocated is the analysis of small groups that are not necessarily representative of other groups.

4.3 Contributions and implications

One contribution of these findings is the practical benefit of examining relationships between different types of job performance. The participating organisation takes into account job performance components directly important to its goals: accomplishment of individual, team and organisational objectives, and demonstrating the values that lead to a positive work environment for everyone. The results here indicate that there is a relationship between the individual-level dimensions of the organisation's criterion score. These dimensions are task performance (accomplishment of individual objectives) and contextual performance (demonstrating values. Making up 70% of the criterion score, individual differences hypothesised to influence these dimensions may be of particular importance to the organisation.

This study demonstrated the value of examining specific cognitive abilities and specific types of job performance. Comparisons of the general factor of cognitive ability with the general factor of job performance (e.g. Hunter & Hunter, 1984; Ones, et al., 2010) have been helpful in establishing considerable evidence of a link between cognitive ability and job performance, but lack the power to explain how and why cognitive ability and job performance are related. Because verbal reasoning ability related to task performance in this study, and numeric and abstract reasoning did not, there is evidence that specific cognitive abilities may relate to task performance in different ways. Similarly, cognitive abilities related more to task performance than contextual performance. If contextual performance is particularly important for overall

performance, then cognitive abilities may not predict overall performance in these jobs very well.

Developing a theory of ability-task fit for job performance (c.f. person-job fit; Edwards, 1991) has the potential to enhance criterion-related validity estimates. That is, criterion-related validity coefficients may represent the task-relatedness of different cognitive abilities, so maximising ability-task fit also maximises the criterion-related validity coefficients of particular cognitive abilities for each job. Ability-task fit could explain why GRT2 predictive validity studies showed inconsistent results: the job tasks required some abilities more than others for successful performance, meaning that sometimes numeric reasoning was important, and other times verbal reasoning was more important. It may also explain why job complexity moderates the relationship between cognitive ability and job performance (section 1.4.4; Ree, et al., 2002; Schmidt & Hunter, 1998). More complex jobs may have a more diverse range of tasks, which may require greater levels of different types of cognitive abilities.

4.4 Future research

Future research should consider hierarchical analysis of multi-level performance, in order to understand how individuals contribute to each level. If individual contributions can be well understood, then organisations may have some idea of which individual differences may be particularly important for assessing when selecting new employees.

Future research could also narrow down which individual differences relate to which dimensions of performance, to develop theory about the relationships of traits to behaviour at work. The results here suggest verbal reasoning ability is related to task performance. This may be a unique relationship for this organisation or this group of

employees, representing task-relatedness of verbal reasoning. Or verbal reasoning could relate to job performance in all jobs. Closer examination also could have the practical benefit of enhancing criterion-relate validity of selection procedures.

4.5 Conclusion

This research found a link between a specific cognitive ability and a specific dimension of job performance. The investigation is unique, and thus the finding contributes to knowledge about how cognitive ability and job performance are related. Greater verbal reasoning ability was associated with higher ratings of task performance in this data set. Though this result may not generalise to other employees doing different jobs in various organisations, one implication could be that verbal reasoning has the highest criterion-related validity for task performance compared to other cognitive abilities. If this result is not ubiquitous, then job-relatedness of specific cognitive abilities may affect their criterion-related validity. Either way, the practical implication of this result is that organisations can enhance the criterion-related validity of their selection procedures by examining particular cognitive abilities and their relationships to the most valued dimensions of job performance.

6 Appendices

6.1 Appendix one: Guilford's Structure-of-Intellect model

Guilford's (1967) structure of intellect theory examined aspects of test items to determine the nature of the ability being tested. Guilford suggested that items had three features: the content of the problem, the operations required to solve the problem, and the product of the problem. Each of these features had a number of categories within it: content could be figural, symbolic, semantic and/or behavioural; operations could involve cognition, memory, divergent production, convergent production and/or evaluation; and products could be in the form of units, classes, relations, systems, transformations and/or implications (Guilford, 1967). Later, figural content was recategorised as visual and auditory, and the memory operation was divided into memory recording and memory retention (Guilford, 1988).

The structure of intellect, therefore, could be visualised as a cube made up of 5 x 6 x 6 = 180 separate cubes, each representing a problem-solving ability that was a combination of one content factor, one operations factor and one product factor. Individual differences in intelligence could thus be described as differences in the number of abilities an individual possessed, which could include strengths or weaknesses in a category or combination of categories. The theory was influential because it suggested a method of inferring abilities by examining problem characteristics, as well as highlighting the usefulness of analysing specific abilities and relations between them. Empirical support for the model, however, was insubstantial (Brody, 2000; Horn & Knapp, 1973; Suss & Beauducel, 2005; Undheim & Horn, 1977).

6.2 Appendix two: Cognitive ability testing procedures

6.2.1 Internet reasoning test

Note that the following procedure relates to any job candidate invited to sit the IRT2.

Administrators from hiring organisations must first be trained to Level B of the British Psychological Society before they are granted access to the GeneSys testing platform. This ensures all administrators understand the principles of standardisation, reliability and validity, and score interpretation, and the importance of following strict guidelines and procedures for testing.

Candidates are emailed an invitation to participate in the testing. Though the hiring organisation can add information as is deemed necessary, OPRA Consulting Ltd has downloadable default emails for each of the tests, which are to be kept as standard as possible. The default email for the IRT2 reads:

> Dear *candidate name*,
>
> Please find below the link for you to complete the IRT2 assessment (Internet Reasoning Test), distributed by OPRA Consulting Group.
>
> The IRT2 is made up of three components – Verbal reasoning (4 minutes), Numerical reasoning (4 minutes), and Abstract reasoning (diagrammatical; 6 minutes). Each component begins by displaying its own set of instructions, and then the timed assessment begins. The three assessment components will follow on from each other automatically.
>
> It should take you no longer than 20 minutes to complete this assessment, including reading time. Please make sure that you set aside enough time to complete this in one sitting and are free from any distractions (are alone in the room) before you begin. No calculators or any other adding tools are to be used.
>
> If you have any questions or comments, please do not hesitate to contact me.
>
> Kind Regards
> *Administrator name*
>
> *Organisation name*
> *Phone number*
> *Email address*
>
> Please do not reply directly to this e-mail. If you have any queries you can contact the session administrator at the following address: *email@address.com*

> Please note: When the questionnaire completes, you will be taken to a final page confirming that the results have been successfully submitted. You should not close the browser while the 'submitting results' (pencil moving back and forth) animation is visible. If you experience any problems during the questionnaire, such as a loss of internet connection or a system crash, you can restart your questionnaire session by clicking on the provided link again and continuing from where you left off.
>
> Please click on the following link to launch the questionnaire:
>
> http://www.genesysonline.net/gsonline/startupas3.asp?LANGID=EN*UNIQUECODE*
>
> Alternatively, please go to:
>
> http://www.genesysonline.net/gsonline/startupas3.asp?LANGID=EN
>
> and enter the following login details:
>
> Session ID: *xxxxxxxxx*
> Respondent ID: *YYYYYYYY*
>
> Please note that Adobe Flash Player 9 or higher is required to complete the questionnaire.

The email includes a web link to a unique webpage for each candidate within the GeneSys testing platform, where the candidate can participate in the tests chosen by the hiring organisation (e.g. the IRT2 and a personality test such as the 15 factor questionnaire, 15FQ+). As is clear from the email, participants are able to complete the test whenever they want, within a time frame site by the hiring organisation. If they do not have access to a computer or the internet, alternate arrangements can be made with the hiring organisation. If the candidate attends a testing session at the organisation's premises, the web link and testing procedure are exactly the same as if they take the test at home.

When the participants click on the web link in the email, a new browser window opens. This screen gives the candidate's first and last names, and asks if the details are correct. If not, candidates are asked to contact their administrator, with contact details available on the screen. If the name is correct, the candidate can click on Log in.

Another browser window opens once the candidate logs in, which is designed to cover the whole computer screen, including the toolbar that usually appears along the

bottom of the screen, which reduces potential distractions. On all screens there are left ← and right → arrows that the candidate can click on to go back to the previous screen or onward to the next screen. Similarly, clicking page up (PgUp) moves the candidate back one screen, while page down (PgDn) moves them forward one screen. If this browser window closes at any time during the testing, (e.g. if the candidate closed it or if the internet or computer crashes), clicking on the email link reopens the testing session at the point where it stopped.

After a welcome page, the first screen is the confidentiality agreement disclosing treatment of responses. It informs the candidates that the responses collected are confidential, and will be stored in a secure database maintained by a UK service internet provider. The information continues: responses will be forwarded to a trained person who commissioned the assessment (e.g. the human resources manager from the hiring organisation). Data will be stored securely by Psytech International Ltd for six months unless specific arrangements have been made, after which time it will be anonymised. Storing the data allows the developers to monitor the assessment quality. The final instruction tells participants to confirm they understand and accept the terms of the agreement by checking a box provided at the bottom of the page. The candidate will not be able to move to the next screen until the box has been checked.

The second screen requests biodata: the participant's name, age, sex (M or F), reference, ethnicity, education level attained, job area, industry and sector. The candidate's name is automatically inserted in this screen, and the default sex is Male. Candidates cannot progress if the name is deleted, but all other information is optional. The reference section is provided for the convenience of hiring organisations, but can be left blank. There are drop-down boxes with set options provided for ethnicity,

education, job area, industry and sector. Note that selecting “other” in any of these does not provide a text box for further specification.

The third screen welcomes the candidate to the verbal reasoning subtest. The candidate is informed that the test is designed to assess understanding of words and relationships between words. It states that the subtest has 17 questions, each with 6 possible answers, and that one and only one answer is correct.

The fourth screen contains the instructions for recording answers and moving through the test. F1 is a help screen, which when pressed at any time reiterates the instructions from this page. The candidate is instructed to use number keys to select an answer, and press enter to record the answer and move to the next question. The PgUp and PgDn keys move the candidate backwards or forwards in the test if they want to review previously answered items or skip items. Note: although it is not stated, responses can also be recorded by clicking on them with the mouse, and then moving to the next item by clicking on the right arrow.

The fifth screen introduces the candidates to example items for the subtest. The candidate can practice highlighting an answer by pressing a number from 1 to 6 or clicking on their preferred option, and pressing enter or PgDn. When candidates press enter, PgDn or click on the right arrow (even if they have not selected a response), the correct answer to the example item will be shown, with an explanation for why it is correct. In the verbal subtest, there are 4 examples, each being a different type, such as synonyms (as in the example above), antonyms (i.e. opposite meanings), which word is the odd one out (i.e. by not belonging to the same class or category as the other words) and word relations (e.g. mother is to daughter, as father is to....). The explanation for the verbal reasoning example given in section 2.2.1 is “The correct answer is 2, because sick means the same as ill”.

The sixth screen is entitled ‘Remember’ and informs candidates: that time is short, so they should work as quickly and accurately as possible; to change an answer, they need to highlight the new choice and press enter to record it; if they finish before time runs out, they will be given a chance to review and change answers; and if they need to review instructions during the test to press F1, though the clock will continue to count down.

The seventh and final screen informs candidates that the verbal subtest contains 17 questions, and they have 4 minutes to attempt them. It is stated that if candidates are unsure about what to do, they should ask the administrator (which would require returning to their email invitation and calling or emailing the contact listed). The final instruction is to press PgDn to begin the test.

During the test, a clock is in the top right hand corner of the page indicating how much time remains. In the centre at the top of the page, it is also indicated which number item the candidate is up to out of the total. In the middle at the bottom of the page, it says “Help F1”.

Candidates must provide an answer to each question. If a question is left black, a box appears stating, “You must answer all questions”, with an OK button which the candidate must click on to return to the test. The clock continues to count down.

When time is up, a box appears saying “You have run out of time”. The candidate clicks on an OK button in the box to progress to the next screen. If the candidate finishes all the questions and there is still time on the clock, a pop-up box will appear asking if the candidate wishes to review their answers. If yes, it returns to the test, where the candidate can use PgUp/PgDn or click the arrows to review answers. If the candidate says no, the test is completed. When the test is finished, a screen appears with an animated scribbling pencil graphic to indicate that answers are being recorded.

After the verbal subtest responses have been transmitted, the numerical reasoning welcome page appears. The candidate presses any key to continue. The candidate can take as much time for a break here as they choose.

Numerical reasoning subtest instructions appear. Candidates are informed that the numerical reasoning subtest is designed to assess their ability to use numbers; there are 14 questions, each with 6 possible answers; one and only one answer is correct.

The second screen has the examples, and again are of different types: odd one out (as in the example above), giving the next number in a series, arithmetic problems in numbers, and arithmetic problems in applied settings (e.g. if a car is travelling at 100kms per hour, how long will it take to travel 20kms?) Correct answers are given and explained. The description of the answer from the example given in section 2.2.1 is "the correct answer is 4, as all the other fractions can be reduced further".

The third screen is the same as the 'remember' screen for the verbal subtest.

The fourth screen also follows the format of the final verbal subtest screen, stating that the numerical reasoning test has 14 items and that the candidate has 4 minutes to attempt them.

Once the numerical test is completed and data transmitted, the candidates proceed to the abstract reasoning subtest, where they are informed that the test is designed to assess the ability to perceive and understand relationships between abstract shapes and patterns.

There are only three examples for the abstract reasoning subtest: which pattern comes next (as in the example), odd one out, and spatial relationships, e.g. shape *a* is shown one way, then shape *b* is the same as shape *a*, except rotated 180 degrees. Another shape (*c*) is shown, then the question asks "*a* is to *b* as *c* is to ...?" Answers are given and explained. The answer to the example item in section 2.2.1 is "the top four

boxes alternate 2, 4, 2, 4 stars respectively. The orientation of the boxes containing two stars alternates also. Therefore box 2 is the correct answer."

The 'remember' screen appears next.

The final screen follows the same format as that of the previous two subtests, but states that the test has 14 questions and the candidate has 6 minutes within which to attempt them.

When the abstract reasoning test is finished, the final screen says:

> Thank you. Your answers have been recorded successfully.
> The testing session is now at an end. Click the button below to close the browser window.

6.2.2 Adaptive general reasoning test

Much of the procedure is exactly the same for the AdaptGRT. The differences are specified below.

The email states:

> Please find below the link for you to complete the AdaptGRT assessment (Adaptive General Reasoning Test), distributed by OPRA Consulting Group.
>
> The AdaptGRT is made up of three components – Verbal reasoning (6 minutes), Numerical reasoning (8 minutes), and Abstract reasoning (diagrammatical; 8 minutes). Each component begins by displaying its own set of instructions, and then the timed assessment begins. The three assessment components will follow on from each other automatically.
>
> It should take you no longer than 25 minutes to complete this assessment, including reading time. Please make sure that you set aside enough time to complete this in one sitting and are free from any distractions (are alone in the room) before you begin. No calculators or any other adding tools are to be used.

The AdaptGRT verbal subtest has 15 questions and allows 6 minutes. The numeric subtest has 15 questions and allows 8 minutes. The abstract reasoning subtest has 15 questions and allows 8 minutes.

During the AdaptGRT testing sessions, candidates cannot go back to previous answers at any point during the test. When they finish a subtest, a box appears saying, "You have finished the test. Please click OK to continue".

6.3 Appendix three: IRT2 Reference group demographics

6.3.1 Verbal reasoning subtest reference group

This group had 1523 New Zealand participants. The mean age of participants was 33.62, minimum 16, maximum 64, with information for 9 people missing. There were 914 men, 606 women and 3 who did not state their gender. Six participants did not complete the test as part of a selection process. Out of the other 1517 participants, 65 definitely completed the tests as part of a selection process, and it is unknown whether the other 1452 were applicants or not. Given that the tests are used primarily for selection processes, however, it is likely that a large proportion of these were job applicants.

Table 6.1: Education levels of IRT2 verbal reasoning subtest reference group participants

Less than completed secondary	1
Completed Secondary School yr 12/13	3
Industry, Trade Training	154
Certificate, Diploma	1
Polytech, TAFE, Institute	0
University Degree	309
Post Graduate Qualification	64
Masters	0
MBA	0
PhD	7
Missing	984

Table 6.2: Industry sector worked in by participants of IRT2 verbal reasoning subtest reference group

Accounting	24
Advertising, Marketing	9
Agribusiness	91
Banking, Finance, Investment	53

Call Centre	154
Compliance, Law enforcement	97
Consulting	15
Construction, works, roads	6
Defence, Armed Services	0
Education, Training	31
Emergency Services	175
Engineering, Technical	22
Entertainment	1
Forestry	1
Health, Therapy, Care	13
Hospitality, Tourism	0
Insurance	14
Information Technology	56
Internet	17
Legal	71
Manufacturing, Assembly	27
Media, Broadcasting	5
Mining	7
Property, Real Estate	7
Public Relations	2
Research	14
Recruitment	19
Retail	81
Service	50
Telecommunications	104
Trades, Build, Auto, other	6
Transport, Shipping	35
Utilities, electricity, etc	5
Wholesale, Trading	10
Missing	301

Table 6.3: Ethnicity of participants of IRT2 verbal reasoning subtest reference group

Australian	29
Aboriginal/Torres Strait Islander	0
NZ European	211
Maori	15
African	2
Asian	7
European	392
Indian	11
Latin, Hispanic	0
Middle East	0
Pacific Islander	7
South African European	7
Other	31
Missing	811

Table 6.4: Occupation type of participants of the IRT2 verbal reasoning subtest reference group

Clerical, Administration	47
Company Director, Partner, Owner	0
Customer Service	36
Designer, Creative	2
Driver, Operator	2
Graduate (within past year)	6
Home executive	1
Hourly paid Worker	7
Human Resources	18
Managerial	81
Manager of $60M+ pa	0
Marketing	11
Process worker, Industrial	0
Professional	318
Salaried Staff	22

Sales	44
Student, School leaver	6
Supervisor, Team leader	26
Trades-person, Certified worker	10
Unemployed	2
Volunteer	0
Missing	884

Table 6.5: Type of organisation worked for by participants of IRT2 verbal reasoning subtest reference group

Public service, Fed/State Govt	472
Local Government/Authority	67
State Owned Corporation	60
Private – small/medium	4
Corporate, Multinational	9
Missing	911

6.3.2 Numeric reasoning subtest reference group

This group also had 1523 New Zealand participants. The mean age of the group was 33.82, minimum 16, maximum 64, with information for 21 participants missing. There were 918 men, 602 women and 3 who did not state their gender. Six participants did not complete the test as part of a selection process and 64 did. As with the verbal reasoning subtest group, it was not specified whether the other 1453 participants were job applicants or not, but it is likely that a large proportion of these were.

Table 6.6: Education levels of IRT2 numeric reasoning subtest reference group participants

Less than completed secondary	1
Completed Secondary School yr 12/13	3
Industry, Trade Training	155
Certificate, Diploma	1

Polytech, TAFE, Institute	0
University Degree	308
Post Graduate Qualification	64
Masters	0
MBA	0
PhD	7
Missing	984

Table 6.7: Industry sector worked in by participants of IRT2 numeric reasoning subtest reference group

Accounting	24
Advertising, Marketing	9
Agribusiness	91
Banking, Finance, Investment	53
Call Centre	154
Compliance, Law enforcement	97
Consulting	15
Construction, works, roads	6
Defence, Armed Services	0
Education, Training	31
Emergency Services	175
Engineering, Technical	22
Entertainment	1
Forestry	1
Health, Therapy, Care	13
Hospitality, Tourism	0
Insurance	14
Information Technology	56
Internet	17
Legal	71
Manufacturing, Assembly	27
Media, Broadcasting	5

Mining	7
Property, Real Estate	7
Public Relations	2
Research	14
Recruitment	19
Retail	82
Service	50
Telecommunications	104
Trades, Build, Auto, other	6
Transport, Shipping	35
Utilities, electricity, etc	5
Wholesale, Trading	10
Missing	300

Table 6.8: Ethnicity of participants of IRT2 numeric reasoning subtest reference group

Australian	29
Aboriginal/Torres Strait Islander	0
NZ European	209
Maori	15
African	2
Asian	7
European	392
Indian	11
Latin, Hispanic	0
Middle East	0
Pacific Islander	7
South African European	7
Other	32
Missing	812

Table 6.9: Occupation type of participants of the IRT2 numeric reasoning subtest reference group

Clerical, Administration	46
Company Director, Partner, Owner	0
Customer Service	36
Designer, Creative	2
Driver, Operator	2
Graduate (within past year)	5
Home executive	1
Hourly paid Worker	7
Human Resources	18
Managerial	82
Manager of $60M+ pa	0
Marketing	11
Process worker, Industrial	0
Professional	318
Salaried Staff	22
Sales	44
Student, School leaver	6
Supervisor, Team leader	26
Trades-person, Certified worker	10
Unemployed	2
Volunteer	0
Missing	885

Table 6.10: Type of organisation worked for by participants of IRT2 numeric reasoning subtest reference group

Public service, Fed/State Govt	472
Local Government/Authority	67
State Owned Corporation	60
Private – small/medium	3
Corporate, Multinational	9
Missing	912

6.3.3 Abstract reasoning subtest reference group

This group had 1551 New Zealand participants. The mean age of the group was 33.72, minimum 16, maximum 64, with information for 27 participants missing. There were 931 men, 611 women and 9 who did not state their gender. Seven participants did not complete the test as part of a selection process and 65 did. As with the verbal and numeric reasoning subtest groups, it was not specified whether the other 1479 participants were job applicants or not, but it is likely that a large proportion of these were.

Table 6.11: Education levels of IRT2 abstract reasoning subtest reference group participants

Less than completed secondary	1
Completed Secondary School yr 12/13	3
Industry, Trade Training	157
Certificate, Diploma	1
Polytech, TAFE, Institute	0
University Degree	330
Post Graduate Qualification	65
Masters	0
MBA	0
PhD	7
Missing	987

Table 6.12: Industry sector worked in by participants of IRT2 abstract reasoning subtest reference group

Accounting	24
Advertising, Marketing	9
Agribusiness	91
Banking, Finance, Investment	54
Call Centre	154
Compliance, Law enforcement	97
Consulting	17

Construction, works, roads	6
Defence, Armed Services	0
Education, Training	32
Emergency Services	175
Engineering, Technical	36
Entertainment	1
Forestry	1
Health, Therapy, Care	13
Hospitality, Tourism	0
Insurance	14
Information Technology	56
Internet	17
Legal	71
Manufacturing, Assembly	27
Media, Broadcasting	7
Mining	7
Property, Real Estate	7
Public Relations	2
Research	14
Recruitment	18
Retail	82
Service	50
Telecommunications	110
Trades, Build, Auto, other	6
Transport, Shipping	36
Utilities, electricity, etc	5
Wholesale, Trading	10
Missing	302

Table 6.13: Ethnicity of participants of IRT2 abstract reasoning subtest reference group

Australian	29
Aboriginal/Torres Strait Islander	0

NZ European	211
Maori	15
African	2
Asian	10
European	394
Indian	11
Latin, Hispanic	0
Middle East	0
Pacific Islander	7
South African European	8
Other	32
Missing	832

Table 6.14: Occupation type of participants of the IRT2 abstract reasoning subtest reference group

Clerical, Administration	46
Company Director, Partner, Owner	0
Customer Service	36
Designer, Creative	2
Driver, Operator	2
Graduate (within past year)	8
Home executive	1
Hourly paid Worker	7
Human Resources	20
Managerial	82
Manager of $60M+ pa	0
Marketing	11
Process worker, Industrial	0
Professional	318
Salaried Staff	22
Sales	44
Student, School leaver	6

Supervisor, Team leader	26
Trades-person, Certified worker	12
Unemployed	2
Volunteer	0
Missing	906

Table 6.15: Type of organisation worked for by participants of IRT2 abstract reasoning subtest reference group

Public service, Fed/State Govt	473
Local Government/Authority	69
State Owned Corporation	61
Private – small/medium	3
Corporate, Multinational	10
Missing	935

Appendix four: Gower index of similarity

6.4.1 Method

To measure the strength and direction of a relationship between two quantitative variables, researchers typically use the Pearson product-moment correlation. It has recently been suggested, however, that reliability and validity estimates would be better indexed by a coefficient of similarity (P. Barrett, 2010c). The Gower index is one such coefficient (Gower, 1971). The Gower index is calculated by averaging the discrepancies between all pairs of scores, dividing this by the maximum possible discrepancy value (which yields the average absolute discrepancy), then subtracting from 1 to express the integer as a proportion of agreement between scores (P. Barrett, 2010c, p. 14). Importantly, perhaps the biggest difference between the Pearson and Gower coefficients is that the Pearson assumes a linear (systematically monotonic) relationship between variables. The Gower index requires no such assumption – it simply indicates to what extent the scores agree.

The Gower index has been used mostly for investigating inter-rater agreement in clinical psychology measures (Zegers, 1991), though comparable similarity indices have been suggested for matching candidates to idealized employee profiles in selection contexts (P. Barrett, 2005). It is arguably better to measure inter-rater agreement with the Gower index than reliability with the Pearson coefficient (Green, 1981), because it describes how similar the ratings are without assuming that each raters' ratings fit a bivariate normal distribution or requiring any corrections if the range of the data is restricted (P. Barrett, 2010c). Barrett (2010c) suggested, however, that the Gower index could also be used to supplement the Pearson coefficient when investigating validity, in order to provide another perspective on the relationship of the data. The Gower index has been included as an analytic tool in this study for this precise purpose.

Excel spreadsheets for each pair of variables were created to calculate Gower indices using Gower v 1.0 (P. Barrett, 2010b). The programme included variable rescaling where required. Bootstrap v 1.0 (P. Barrett, 2010a) was used to create probability statistics relating to each Gower index produced. The Bootstrap programme randomly generated 10,000data sets with the same number of cases as each pair of variables and gave the proportion of the randomly generated data sets that resulted in a Gower index at least as high as that found for the pair of variables examined[39]. Though based on actual (randomly generated) data sets, the Gower bootstrap proportions are interpreted as probability estimates. Using 10,000 data sets ensured a large enough number that proportions would be relatively stable no matter how many times the random data sets were generated.

Note that each pair of variables be the result of values on the same measurement scale in order to be compared using the Gower index. If they are not, one set of

[39] The user can choose any number of data sets. The researcher chose 10,000 as a sufficiently large number to produce trustworthy probabilities.

variables will need to be linearly rescaled. In the current research, Gower indices were not calculated if measurement scales were deemed too different to be meaningfully transformed. This was the case for the job performance dimensions when compared with any variables other than each other, as these only had three possible levels of measurement. Age and length of service were also removed from analysis of Gower indices, because of the arbitrary assignation of minimum and maximum possible values.

6.4.2 Results and Discussion

Gower indices for cognitive ability subtest scores showed 73-74% agreement between scores (see table 6.16; Gower = .73, p = .046 for VR and NR; Gower = 74%, p = .029 for VR and AR; Gower = 74%, p = .019 for NR and VR), supporting the hypothesis that cognitive abilities are related.

Table 6.16: Gower indices for cognitive ability subtest scores

		VR	**NR**
NR	Gower index	.73*	-
	N	*39*	-
AR	Gower index	.74*	.74*
	N	*41*	*39*

*p<.05

Gower indices of similarity for job performance dimensions showed team performance level having a mean absolute agreement of 72% with individual performance ratings (p = .064). These similarities are explained by the clusters of data points.

Table 6.17: Gower indices for job performance ratings

		IPR	TPL
TPL	Gower index	.72^	-
	N	*43*	-
VAD	Gower index	.62	.64
	N	*43*	*43*

^p<.1

Gower indices suggested 74% mean similarity in verbal reasoning scores and overall performance, which was the same percentage of agreement found for abstract reasoning scores and overall performance (see table 6.18). Inspection of scatter plots showed that both verbal and abstract reasoning had a cluster of data points around higher levels of reasoning ability, but while verbal ability showed a relatively systematic relationship throughout the data set, abstract reasoning had a cluster and then apparently random variability at other levels of each variable. In this instance, the Gower agreement indicated the clustering of data points, but does not provide much information about the relationships between variables, because the variables relate to instruments utilising different scales in order to measure different constructs.

Table 6.18: Gower indices for cognitive ability subtest scores and overall performance ratings

		VR	NR	AR
OP	Gower	.74*	.66	.74*
	N	*43*	*39*	*41*

^p<.1 *p<.05 **p<.01

References

Ackerman, P. L. (1997). Personality, self-concept, interests, and intelligence: Which construct doesn't fit? *Journal of Personality, 65,* 171-204. doi: 10.1111/j.1467-6494.1997.tb00952.x

Ackerman, P. L., & Beier, M. E. (2005). Knowledge and intelligence. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 125-139). Thousand Oaks, CA: Sage.

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121,* 219-245. doi: 10.1037/0033-2909.121.2.219

Ackerman, P. L., & Humphreys, L. G. (1990). Individual differences in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 1, pp. 223-282). Palo Alto, CA: Consulting Psychologists Press.

Alfonso, V. C., Flanagan, D. P., & Radwan, S. (2005). The impact of the Cattell-Horn-Carroll theory on test development and interpretation of cognitive and academic abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 185-202). New York, NY: Guilford Press.

Arthur, W., Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93,* 435-442. doi: 10.1037/0021-9010.93.2.435

Astin, A. W. (1964). Criterion-centered research. *Educational and Psychological Measurement, 24,* 807-822. doi: 10.1177/001316446402400408

Austin, J. T., Humphreys, L. G., & Hulin, C. L. (1989). Another view of dynamic criteria: A critical reanalysis of Barrett, Caldwell, and Alexander. *Personnel Psychology, 42,* 583-596. doi: 10.1111/j.1744-6570.1989.tb00670.x

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917-1992. *Journal of Applied Psychology, 77,* 836-874. doi: 10.1037/0021-9010.77.6.836

Baron, M. R., & Kenny, D. A. (1986). The moderator-mediator distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173-1182.

Barrett, G. V., Caldwell, M. S., & Alexander, R. A. (1985). The concept of dynamic criteria: A critical reanalysis. *Personnel Psychology, 38,* 41-56. doi: 10.1111/j.1744-6570.1985.tb00540.x

Barrett, P., & Eysenck, H. J. (1992). Brain evoked potentials and intelligence: The Hendrickson paradigm. *Intelligence, 16,* 361-381. doi: 10.1016/0160-2896%2892%2990015-J

Barrett, P., Eysenck, H. J., & Lucking, S. (1986). Reaction time and intelligence: A replicated study. *Intelligence, 10,* 9-40. doi: 10.1016/0160-2896%2886%2990025-5

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1-26. doi: 10.1111/j.1744-6570.1991.tb00688.x

Bates, T. C., & Eysenck, H. J. (1993). Intelligence, inspection time, and decision time. *Intelligence, 17,* 523-531. doi: 10.1016/0160-2896%2893%2990016-X

Bell, E. (2008). *Theories of performance*. Los Angeles, CA: Sage.

Bell, S. T. (2007). Deep-level composition variables as predictors of team performance: A meta-analysis. *Journal of Applied Psychology, 92,* 595-615. doi: 10.1037/0021-9010.92.3.595

Bellinger, D. C., & Adams, H. F. (2001). Environmental pollutant exposures and children's cognitive ability. In R. J. Sternberg & E. L. Grigorenko (Eds.), *Environmental effects on cognitive abilities* (pp. 157-188). Mahwah, NJ: Lawrence Erlbaum Associates.

Berg, C. A., & Klaczynski, P. A. (2002). Contextual variability in the expression and meaning of intelligence. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The general factor of intelligence: How general is it?* (pp. 381-412). Mahwah, NJ: Lawrence Erlbaum Associates.

Berger, M. (1982). The 'scientific approach' to intelligence: An overview of its history with special reference to mental speed. In H. J. Eysenck (Ed.), *A model for intelligence* (pp. 13-44). Berlin: Springer-Verlag.

Bergman, M. E., Donovan, M. A., Drasgow, F., Overton, R. C., & Henning, J. B. (2008). Test of Motowidlo et al.'s (1997) theory of individual differences in task and contextual performance. *Human Performance, 21,* 227-253. doi: 10.1080/08959280802137606

Bernardin, H. J. (1977). Behavioral expectation scales versus summated scales: A fairer comparison. *Journal of Applied Psychology, 62,* 422-427. doi: 10.1037/0021-9010.62.4.422

Bernardin, H. J., & Beatty, R. W. (1984). *Performance appraisal : Assessing human behavior at work*. Boston, MA: Kent.

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6,* 205-212.

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology, 65,* 60-66. doi: 10.1037/0021-9010.65.1.60

Berry, J. W. (2001). Contextual studies of cognitive adaptation. In J. M. Collis & S. Messick (Eds.), *Intelligence and personality: Bridging the gap in theory and measurement* (pp. 319-334). Mahwah, NJ: Lawrence Erlbaum Associates.

Bertua, C., Anderson, N., & Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. *Journal of Occupational and Organizational Psychology, 78,* 387-409. doi: 10.1348/096317905x26994

Binet, A., Simon, T., & Drummond, W. B. (1914). *Mentally defective children*. London, England: Arnold.

Binet, A., Simon, T., & Kite, E. S. (1916a). *The development of intelligence in children (The Binet-Simon Scale)*. Baltimore, MD: Williams & Wilkins.

Binet, A., Simon, T., & Kite, E. S. (1916b). *The intelligence of the feeble-minded*. Baltimore, MD: Williams & Wilkins.

Binet, A., Simon, T., & Town, C. H. (1912). *A method of measuring the development of the intelligence of young children*. Lincoln, IL: Courier Co.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74,* 478-494. doi: 10.1037/0021-9010.74.3.478

Blumberg, M., & Pringle, C. D. (1982). The missing opportunity in organizational research: Some implications for a theory of work performance. *Academy of Management Review, 7,* 560-569.

Boring, E. G. (1923). Intelligence as the tests test it. *New Republic, 34,* 35-37.

Borman, W. C. (1974). The rating of individuals in organizations: An alternate approach. *Organizational Behavior & Human Performance, 12,* 105-124. doi: 10.1016/0030-5073%2874%2990040-3

Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology, 60,* 556-560. doi: 10.1037/0021-9010.60.5.556

Borman, W. C. (1978). Exploring upper limits of reliability and validity in job performance ratings. *Journal of Applied Psychology, 63,* 135-144. doi: 10.1037/0021-9010.63.2.135

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology, 64,* 410-421. doi: 10.1037/0021-9010.64.4.410

Borman, W. C., Bryant, R. H., & Dorio, J. (2010). The measurement of task performance as criteria in selection research. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 439-461). New York, NY: Routledge/Taylor & Francis Group.

Borman, W. C., Buck, D. E., Hanson, M. A., Motowidlo, S. J., Stark, S., & Drasgow, F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology, 86,* 965-973. doi: 10.1037/0021-9010.86.5.965

Borman, W. C., Hanson, M. A., & Hedge, J. W. (1997). Personnel selection. *Annual Review of Psychology, 48,* 299-337. doi: 10.1146/annurev.psych.48.1.299

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organisations* (pp. 71-98). San Francisco, CA: Jossey-Bass.

Borman, W. C., & Motowidlo, S. J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance, 10,* 99-109. doi: 10.1207/s15327043hup1002_3

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*: Cambridge University Press.

British Psychological Society. (2003). *Review of general reasoning test GRT2*. Retrieved from http://www.psychtesting.org.uk/test-registration-and-test-reviews/test-reviews.cfm?page=evalsummary&Test_ID=80

Brody, N. (2000). History of theories and measurements of intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 16-33). New York, NY: Cambridge University Press.

Bruchez-Hall, C., & Gruber, H. E. (1994). Piaget, Jean. (1896-1980). In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 809-814). New York, NY: Macmillan.

Budd, R. (1993). *General, critical and graduate test battery: The technical manual*. Technical Manual retrieved from http://www.opragroup.com/images/opra/pdf/downloads/TechnicalManuals/grt1grt2crtb1.pdf

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology, Vol 1* (2nd ed., pp. 687-732). Palo Alto, CA: Consulting Psychologists Press.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organisations* (pp. 35-70). San Francisco, CA: Jossey-Bass.

Carroll, J. B. (1943). The factorial representation of mental ability and academic achievement. *Educational and Psychological Measurement, 3,* 307-332. doi: 10.1177/001316444300300127

Carroll, J. B. (1944). The analysis of verbal behavior. *Psychological Review, 51,* 102-119. doi: 10.1037/h0061404

Carroll, J. B. (1953). *The study of language: a survey of linguistics and related disciplines in America*. Cambridge, MA: Harvard University Press.

Carroll, J. B. (1954). Individual differences. *Annual Review of Psychology, 5,* 127-148. doi: 10.1146/annurev.ps.05.020154.001015

Carroll, J. B. (1964). Linguistics and the psychology of language. *Review of Educational Research, 34,* 119-126. doi: 10.2307/1169750

Carroll, J. B. (1973). Implications of aptitude test research and psycholinguistic theory for foreign-language teaching. *International Journal of Psycholinguistics, 2,* 5-14.

Carroll, J. B. (1978). How shall we study individual differences in cognitive abilities? Methodological and theoretical perspectives. *Intelligence, 2,* 87-115. doi: 10.1016/0160-2896%2878%2990002-8

Carroll, J. B. (1987). Psychometric approaches to cognitive abilities and processes. In S. H. Irvine & S. E. Newstead (Eds.), *Intelligence and cognition: Contemporary frames of reference* (pp. 217-251). Dordrecht, Netherlands: Martinus Nijhoff.

Carroll, J. B. (1993). *Human cognitive abilities : a survey of factor-analytic studies*: Cambridge University Press.

Carroll, J. B., & Burke, M. L. (1965). Parameters of paired-associate verbal learning: Length of list, meaningfulness, rate of presentation, and ability. *Journal of Experimental Psychology, 69,* 543-553. doi: 10.1037/h0022006

Carroll, J. B., & Maxwell, S. E. (1979). Individual differences in cognitive abilities. *Annual Review of Psychology, 30,* 603-640. doi: 10.1146/annurev.ps.30.020179.003131

Cattell, R. B. (1935). *Cattell group intelligence scale*. Oxford, England: Harrap.

Cattell, R. B. (1940). A culture-free intelligence test. *Journal of Educational Psychology, 31,* 161-179. doi: 10.1037/h0059043

Cattell, R. B. (1958). Extracting the correct number of factors in factor analysis. *Educational and Psychological Measurement, 18,* 791-838. doi: 10.1177/001316445801800412

Cattell, R. B. (1961). Fluid and Crystallized Intelligence. In J. J. Jenkins & D. G. Paterson (Eds.), *Studies in individual differences: The search for intelligence* (pp. 738-746). East Norwalk, CT: Appleton-Century-Crofts.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54,* 1-22. doi: 10.1037/h0046743

Cattell, R. B. (1980). The heritability of fluid, gf, and crystallised, gc, intelligence, estimated by a least squares use of the MAVA method. *British Journal of Educational Psychology, 50,* 253-265.

Cattell, R. B. (1986). *Intelligence: its structure, growth, and action*. Amsterdam, North-Holland: Elsevier Science.

Chen, M. J. (1994). Chinese and Australian concepts of intelligence. *Psychology and Developing Societies, 6,* 103-117. doi: 10.1177/097133369400600202

Chen, M. J., Braithwaite, V., & Jong Tsun, H. (1982). Attributes of intelligent behavior: Perceived relevance and difficulty by Australian and Chinese students. *Journal of Cross Cultural Psychology, 13,* 139-156. doi: 10.1177/0022002182013002001

Chen, M. J., & Chen, H.-c. (1988). Concepts of intelligence: A comparison of Chinese graduates from Chinese and English schools in Hong Kong. *International Journal of Psychology, 23,* 471-487. doi: 10.1080/00207598808247780

Chernyshenko, O. S. (2005). *Report on psychometric evaluation of the general reasoning test (GRT2) for New Zealand organisation XX: Measurement equivalance across ethnic and gender groups*. Christchurch, New Zealand: Drasgow Consulting Group.

Chernyshenko, O. S., Stark, S., & Drasgow, F. (2008). *Adaptive General Reasoning test (AdaptGRT): Report to Psychometrics Ltd*. Urbana, IL: Drasgow Consulting Group.

Cleveland, J. N., & Colella, A. (2010). Criterion validity and criterion deficiency: What we measure well and what we ignore. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 551-567). New York, NY: Routledge/Taylor & Francis Group.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159. doi: 10.1037/0033-2909.112.1.155

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation : design & analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Cowan, N. (2005). Understanding intelligence: A summary and an adjustable-attention hypothesis. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 469-488). Thousand Oaks, CA: Sage.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York, NY: Harper & Row.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302. doi: 10.1037/h0040957

Dalessio, A. T. (1998). Using multisource feedback for employee development and personnel decisions. In J. W. Smither (Ed.), *Performance appraisal: State of the art in practice* (pp. 278-330). San Francisco, CA: Jossey-Bass.

Danthiir, V., Roberts, R. D., Schulze, R., & Wilhelm, O. (2005). Mental speed: On frameworks, paradigms, and a platform for the future. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 27-46). Thousand Oaks, CA: Sage.

Davidson, J. E., & Downing, C. L. (2000). Contemporary models of intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 34-49). New York, NY: Cambridge University Press.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education, 44,* 109-117. doi: 10.1111/j.1365-2923.2009.03425.x

De Dreu, C. K. W., Harinck, F., & van Vianen, A. E. M. (1999). Conflict and performance in groups and organizations. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology 1999, Vol 14* (pp. 369-414). New York, NY: John Wiley & Sons.

Deadrick, D. L., Bennett, N., & Russell, C. J. (1997). Using hierarchical linear modeling to examine dynamic performance criteria over time. *Journal of Management, 23,* 745-757. doi: 10.1177/014920639702300603

Deadrick, D. L., & Gardner, D. G. (2008). Maximal and typical measures of job performance: An analysis of performance variability over time. *Human Resource Management Review, 18,* 133-145. doi: 10.1016/j.hrmr.2008.07.008

Deadrick, D. L., & Madigan, R. M. (1990). Dynamic criteria revisited: A longitudinal study of performance stability and predictive validity. *Personnel Psychology, 43,* 717-744. doi: 10.1111/j.1744-6570.1990.tb00680.x

DeNisi, A. S. (1996). *A cognitive approach to performance appraisal : A program of research*. London, England: Routledge.

DeNisi, A. S., Cafferty, T. P., & Meglino, B. M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior & Human Performance, 33,* 360-396. doi: 10.1016/0030-5073%2884%2990029-1

Dickens, W. T., & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological Review, 108,* 346-369. doi: 10.1037/0033-295X.108.2.346

Dorsey, D. W., Cortina, J. M., & Luchman, J. (2010). Adaptive and citizenship-related behaviours at work. In J. L. Farr & N. Tippins (Eds.), *Handbook of employee selection* (pp. 463-487). New York, NY: Routledge/Taylor & Francis Group.

Drasgow, F., & Hulin, C. L. (1990). Item response theory. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology, Vol 1* (2nd ed., pp. 577-636). Palo Alto, CA: Consulting Psychologists Press.

Drasgow, F., Leucht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 471-515). Westport, CT: Praeger.

DuBois, C. L. Z., Sackett, P. R., Zedeck, S., & Fogli, L. (1993). Further exploration of typical and maximum performance criteria: Definitional issues, prediction, and White-Black differences. *Journal of Applied Psychology, 78,* 205-211.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review, 95,* 256-273. doi: 10.1037/0033-295X.95.2.256

Edenborough, R. (2005). *Assessment methods in recruitment, selection, and performance : a manager's guide to psychometric testing, interviews, and assessment centres*. Sterling, VA: Kogan Page.

Edwards, J. R. (1991). Person-job fit: A conceptual integration, literature review, and methodological critique. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology, 1991, Vol 6* (pp. 283-357). Oxford, England: John Wiley & Sons.

EEO Trust (Producer). (2007, 25/09/2009). Workforce Profile Spreadsheet. Retrieved from http://www.eeotrust.org.nz/toolkits/index.cfm

Elliott, E. S., & Dweck, C. S. (1988). Goals: An approach to motivation and achievement. *Journal of Personality and Social Psychology, 54,* 5-12. doi: 10.1037/0022-3514.54.1.5

Embretson, S. E. (2010). Measuring psychological constructs with model-based approaches: An introduction. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 1-7). Washington, DC: American Psychological Association.

Embretson, S. E., & McCollam, K. M. S. (2000). Psychometric approaches to understanding and measuring intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 423-444). New York, NY: Cambridge University Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Eysenck, H. J. (1971). *The IQ argument: Race, intelligence and education*. Oxford, England: Library Press.

Eysenck, H. J. (1982). *A model for intelligence*. Berlin: Springer-Verlag.

Eysenck, H. J. (1986). Intelligence: The new look. *Psychologische Beitrage, 28,* 332-365.

Eysenck, H. J. (1988a). The biological basis of intelligence. In S. H. Irvine & J. W. Berry (Eds.), *Human abilities in cultural context* (pp. 87-104). New York, NY: Cambridge University Press; US.

Eysenck, H. J. (1988b). The concept of "intelligence": Useful or useless? *Intelligence, 12,* 1-16. doi: 10.1016/0160-2896%2888%2990019-0

Farr, J. L., & Newman, D. A. (2001). Rater selection: Sources of feedback. In D. W. Bracken, C. W. Timmreck & A. H. Church (Eds.), *The handbook of multisource feedback: The comprehensive resource for designing and implementing MSF processes* (pp. 96-113). San Francisco, CA: Jossey-Bass.

Fay, D., & Frese, M. (2001). The concept of personal initiative: An overview of validity studies. *Human Performance, 14,* 97-124. doi: 10.1207/S15327043HUP1401_06

Fay, D., & Sonnentag, S. (2010). A look back to move ahead: New directions for research on proactive performance and other discretionary work behaviours. *Applied Psychology: An International Review, 59,* 1-20. doi: 10.1111/j.1464-0597.2009.00413.x

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66,* 127-148. doi: 10.1037/0021-9010.66.2.127

Fletcher, C., & Baldry, C. (1999). Multi-source feedback systems: A research perspective. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology 1999, Vol 14* (pp. 149-193). New York, NY: John Wiley & Sons.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101,* 171-191. doi: 10.1037/0033-2909.101.2.171

Frese, M., Garst, H., & Fay, D. (2007). Making things happen: Reciprocal relationships between work characteristics and personal initiative in a four-wave longitudinal structural equation model. *Journal of Applied Psychology, 92,* 1084-1102. doi: 10.1037/0021-9010.92.4.1084

Gardner, H. (1993). *Frames of mind : the theory of multiple intelligences* (2nd ed.). London, England: Fontana.

Gardner, H. (1999). *Intelligence reframed: Multiple intelligences for the 21st century*. New York, NY: Basic Books.

Gelman, S. A. (1994). Competence versus performance. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence*. New York, NY: Macmillan.

Ghiselli, E. E. (1956). Dimensional problems of criteria. *Journal of Applied Psychology, 40,* 1-4. doi: 10.1037/h0040429

Glew, D. J. (2009). Personal values and performance in teams: An individual and team-level analysis. *Small Group Research, 40,* 670-693. doi: 10.1177/1046496409346577

Gottfredson, L. S. (2002a). g: Highly general and highly practical. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The general factor of intelligence: How general is it?* (pp. 331-380). Mahwah, NJ: Lawrence Erlbaum Associates.

Gottfredson, L. S. (2002b). Where and why g matters: Not a mystery. *Human Performance, 15,* 25-46.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics, 27,* 857-871.

Grigorenko, E. L. (2002). Other than g: The value of persistence. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The general factor of intelligence: How general is it?* (pp. 299-327). Mahwah, NJ: Lawrence Erlbaum Associates.

Guilford, J. P. (1967). *The nature of human intelligence*. New York, NY: McGraw-Hill.

Guilford, J. P. (1988). Some changes in the structure-of-intellect model. *Educational and Psychological Measurement, 48,* 1-4. doi: 10.1177/001316448804800102

Gustafsson, J.-E. (2001). On the hierarchical structure of ability and personality. In J. M. Collis & S. Messick (Eds.), *Intelligence and personality: Bridging the gap in theory and measurement* (pp. 25-42). Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., & Bejar, I. I. (1983). *Applications of item response theory*. Vancouver, Canada: Educational Research Institute of British Columbia.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Haslam, S. A. (2001). *Psychology in organisations: The social identity approach* (1st ed.). London, England: Sage.

Hattie, J. A. C. (2007). *Police standards and assessment practice (to probationary constable level)* (Technical report). Wellington, New Zealand: New Zealand Police.

Hebb, D. (1942). The effect of early and late brain injury upon test scores, and the nature of adult intelligence. *Proceedings of the American Philosophical Society, 85,* 275-292.

Hedge, J. W., Borman, W. C., & Birkeland, S. A. (2001). History and development of multisource feedback as a methodology. In D. W. Bracken, C. W. Timmreck & A. H. Church (Eds.), *The handbook of multisource feedback: The comprehensive resource for designing and implementing MSF processes* (pp. 15-32). San Francisco, CA: Jossey-Bass.

Hedge, 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. doi: 10.1037/0021-9010.73.1.68

Hill, K. L. (2001). *Frameworks for sport psychologists: Enhancing sport performance*. Champaign, IL: Human Kinetics.

Hofstee, W. K. B. (2001). Intelligence and personality: Do they mix? In J. M. Collis & S. Messick (Eds.), *Intelligence and personality: Bridging the gap in theory and measurement* (pp. 43-60). Mahwah, NJ: Lawrence Erlbaum Associates.

Hogan, J., & Holland, B. (2003). Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of Applied Psychology, 88,* 100-112. doi: 10.1037/0021-9010.88.1.100

Hogan, J. C., Hogan, R., & Gregory, S. (1992). Validation of a sales representative selection inventory. *Journal of Business and Psychology, 7,* 161-171. doi: 10.1007/BF01013926

Hogan, J. C., Hogan, R., & Murtha, T. (1992). Validation of a personality measure of managerial performance. *Journal of Business and Psychology, 7,* 225-237. doi: 10.1007/BF01013931

Hollenbeck, J. R. (2000). A structural approach to external and internal person-team fit. *Applied Psychology: An International Review, 49,* 534-549. doi: 10.1111/1464-0597.00030

Horn, J. L., & Blankson, N. (2005). Foundations for better understanding of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 41-68). New York, NY: Guilford Press.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57,* 253-270. doi: 10.1037/h0023816

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica, 26,* 107-129. doi: 10.1016/0001-6918%2867%2990011-X

Horn, J. L., & Knapp, J. R. (1973). On the subjective character of the empirical base of Guilford's structure-of-intellect model. *Psychological Bulletin, 80,* 33-43. doi: 10.1037/h0034681

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D. P. Flanagan, J. L. Genshaft & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 53-91). New York, NY: Guilford Press.

Hough, L. M., & Oswald, F. L. (2000). Personnel selection: Looking toward the future--Remembering the past. *Annual Review of Psychology, 51,* 631-664. doi: 10.1146/annurev.psych.51.1.631

Hulsheger, U. R., Maier, G. W., & Stumpp, T. (2007). Validity of general mental ability for the prediction of job performance and training success in Germany: A meta-analysis. *International Journal of Selection and Assessment, 15,* 3-18.

Hunt, S. T. (1996). Generic work behavior: An investigation into the dimensions of entry-level, hourly job performance. *Personnel Psychology, 49,* 51-83. doi: 10.1111/j.1744-6570.1996.tb01791.x

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72-98. doi: 10.1037/0033-2909.96.1.72

Hutchinson, T. P. (1991). *Controversies in item response theory*. Adelaide: Rumsby Scientific.

Inhelder, B., & Sinclair, H. (1994). Piagetian theory of intellectual development. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 814-821). New York, NY: Macmillan.

Irvine, S. H., & Berry, J. W. (Eds.). (1988). *Human abilities in cultural context*. New York, NY: Cambridge University Press.

Jensen, A. R. (1969). Intelligence, learning ability and socioeconomic status. *The Journal of Special Education, 3,* 23-35. doi: 10.1177/002246696900300103

Jensen, A. R. (1978). The current status of the IQ controversy. *Australian Psychologist, 13,* 7-27. doi: 10.1080/00050067808415562

Jensen, A. R. (1987). Intelligence as a fact of nature. *Zeitschrift fur Padagogische Psychologie/ German Journal of Educational Psychology, 1,* 157-169.

Jensen, A. R. (1989). The relationship between learning and intelligence. *Learning and Individual Differences, 1,* 37-62. doi: 10.1016/1041-6080%2889%2990009-5

Jensen, A. R. (1992). Understanding g in terms of information processing. *Educational Psychology Review, 4,* 271-308. doi: 10.1007/BF01417874

Jensen, A. R. (1998). *The g factor : the science of mental ability*. Westport, CT: Praeger.

Jensen, A. R. (2002). Galton's legacy to research on intelligence. *Journal of Biosocial Science, 34,* 145-172. doi: 10.1017/S0021932002001451

Jensen, A. R. (2004). Obituary: John Bissell Carroll. *Intelligence, 32,* 1-5. doi: 10.1016/j.intell.2003.10.001

Kamin, L. J., & Grant-Henry, S. (1987). Reaction time, race, and racism. *Intelligence, 11,* 299-304. doi: 10.1016/0160-2896%2887%2990013-4

Kanfer, R. (1990). Motivation theory and industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 1, pp. 75-170). Palo Alto, CA: Consulting Psychologists Press.

Kanfer, R., & Ackerman, P. L. (2005). Work competence: A person-oriented perspective. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 336-353). New York, NY: Guilford Press.

Kanfer, R., Wolf, M. B., Kantrowitz, T. M., & Ackerman, P. L. (2010). Ability and trait complex predictors of academic and job performance: A person-situation approach. *Applied Psychology: An International Review, 59,* 40-69. doi: 10.1111/j.1464-0597.2009.00415.x

Kaufman, J. C., & Sternberg, R. J. (Eds.). (2010). *The Cambridge handbook of creativity*. New York, NY: Cambridge University Press.

Kline, P. (1991). *Intelligence: The psychometric view*. Florence, KY: Taylor & Frances/Routledge.

Kline, P. (1993). *The handbook of psychological testing*. London, England: Routledge.

Kline, P. (2001). Ability and temperament. In J. M. Collis & S. Messick (Eds.), *Intelligence and personality: Bridging the gap in theory and measurement* (pp. 113-118). Mahwah, NJ: Lawrence Erlbaum Associates.

Kristof, A. L. (1996). Person-organization fit: An integrative review of its conceptualizations, measurement, and implications. *Personnel Psychology, 49,* 1-49. doi: 10.1111/j.1744-6570.1996.tb01790.x

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology, 86,* 148-161. doi: 10.1037/0022-3514.86.1.148

Kuncel, N. R., Ones, D. S., & Sackett, P. R. (2010). Individual differences as predictors of work, educational, and broad life outcomes. *Personality and Individual Differences, 49,* 331-336. doi: 10.1016/j.paid.2010.03.042

Kyllonen, P. C. (2002). g: Knowledge, speed, strategies, or working-memory capacity? A systems perspective. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The general factor of intelligence: How general is it?* (pp. 415-445). Mahwah, NJ: Lawrence Erlbaum Associates.

Landy, F. J., Barnes, J. L., & Murphy, K. R. (1978). Correlates of perceived fairness and accuracy of performance evaluation. *Journal of Applied Psychology, 63,* 751-754. doi: 10.1037/0021-9010.63.6.751

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87,* 72-107. doi: 10.1037/0033-2909.87.1.72

Lang, J. W., Kersting, M., Hulsheger, U. R., & Lang, J. (2010). General mental ability, narrower cognitive abilities, and job performance: The perspective of the nested-factors model of cognitive abilities. *Personnel Psychology, 63,* 595-640. doi: 10.1111/j.1744-6570.2010.01182.x

Latham, G. P., & Wexley, K. N. (1977). Behavioral observation scales for performance appraisal purposes. *Personnel Psychology, 30,* 255-268. doi: 10.1111/j.1744-6570.1977.tb02092.x

Lebas, M., & Euske, K. (2002). A conceptual and operational definition of performance. In A. Neely (Ed.), *Business performance measurement* (pp. 65-79): Cambridge University Press.

LePine, J. A. (2003). Team adaptation and postchange performance: Effects of team composition in terms of members' cognitive ability and personality. *Journal of Applied Psychology, 88,* 27-39. doi: 10.1037/0021-9010.88.1.27

LePine, J. A., Hollenbeck, J. R., Ilgen, D. R., & Hedlund, J. (1997). Effects of individual differences on the performance of hierarchical decision-making teams: Much more than g. *Journal of Applied Psychology, 82,* 803-811. doi: 10.1037/0021-9010.82.5.803

LePine, J. A., & Van Dyne, L. (2001). Voice and cooperative behavior as contrasting forms of contextual performance: Evidence of differential relationships with big five personality characteristics and cognitive ability. *Journal of Applied Psychology, 86,* 326-336.

Lievens, F., & Chan, D. (2010). Practical intelligence, emotional intelligence, and social intelligence. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 339-359). New York, NY: Routledge/Taylor & Francis Group.

Locke, E. A. (Ed.). (1986). *Generalizing from laboratory to field settings : research findings from industrial-organizational psychology, organizational behavior, and human resource management*: Lexington Books.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist, 57,* 705-717. doi: 10.1037/0003-066X.57.9.705

Lohman, D. F. (2001). Issues in the definition and measurement of abilities. In J. M. Collis & S. Messick (Eds.), *Intelligence and personality: Bridging the gap in theory and measurement* (pp. 79-98). Mahwah, NJ: Lawrence Erlbaum Associates.

London, M., Mone, E. M., & Scott, J. C. (2004). Performance management and assessment: Methods for improved rater accuracy and employee goal setting. *Human Resource Management, 43,* 319-336. doi: 10.1002/hrm.20027

Longenecker, C. O., Gioia, D. A., & Sims, J. H. P. (1987). Behind the mask: The politics of employee appraisal. *Academy of Management Executive, 1,* 183-193.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lubinski, D. (2004). John Bissell Carroll (1916-2003). *American Psychologist, 59,* 43-44. doi: 10.1037/0003-066X.59.1.43

MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review, 55,* 95-107. doi: 10.1037/h0056029

Mangels, J. A., Butterfield, B., Lamb, J., Good, C., & Dweck, C. S. (2006). Why do beliefs about intelligence influence learning success? A social cognitive neuroscience model. *Social Cognitive and Affective Neuroscience, 1,* 75-86. doi: 10.1093/scan/nsl013

Marcus, B., Goffin, R. D., Johnston, N. G., & Rothstein, M. G. (2007). Personality and cognitive ability as predictors of typical and maximum managerial performance. *Human Performance, 20,* 275-285.

Matthews, G., Zeidner, M., & Roberts, R. D. (2005). Emotional intelligence: An elusive ability? In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 79-99). Thousand Oaks, CA: Sage.

Mayer, J. D., Salovey, P., & Caruso, D. (2000). Models of emotional intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 396-420). New York, NY: Cambridge University Press.

McCrae, R. R., & Costa, P. T. (1985). Comparison of EPI and psychoticism scales with measures of the five-factor model of personality. *Personality and Individual Differences, 6,* 587-597. doi: 10.1016/0191-8869%2885%2990008-X

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgement tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60,* 63-91.

McDaniel, M. A., & Whetzel, D. L. (2007). Situational judgment tests. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement : Industrial psychology in human resources management* (pp. 235-257). New York, NY: Taylor & Francis Group/Lawrence Erlbaum Associates.

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In D. P. Flanagan, J. L. Genshaft & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 151-119). New York, NY: Guilford Press.

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 136-181). New York, NY: Guilford Press.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114,* 449-458. doi: 10.1037/0033-2909.114.3.449

Mohammed, S., Cannon-Bowers, J., & Foo, S. C. (2010). Selection for team membership: A contingency and multilevel perspective. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 801-822). New York, NY: Routledge/Taylor & Francis Group.

Morgeson, F. P., Delaney-Klinger, K., & Hemingway, M. A. (2005). The importance of job autonomy, cognitive ability, and job-related skill for predicting role breadth and job performance. *Journal of Applied Psychology, 90,* 399-406. doi: 10.1037/0021-9010.90.2.399

Morgeson, F. P., Reider, M. H., & Campion, M. A. (2005). Selecting individuals in team settings: The importance of social skills, personality characteristics, and teamwork knowledge. *Personnel Psychology, 58,* 583-611. doi: 10.1111/j.1744-6570.2005.655.x

Motowidlo, S. J. (2003). Job performance. In W. C. Borman, D. R. Ilgen & R. J. Klimoski (Eds.), *Handbook of psychology: Industrial and organizational psychology, Vol 12* (pp. 39-53). Hoboken, NJ: John Wiley & Sons.

Motowidlo, S. J., Borman, W. C., & Schmit, M. J. (1997). A theory of individual differences in task and contextual performance. *Human Performance, 10,* 71-83. doi: 10.1207/s15327043hup1002_1

Motowidlo, S. J., Brownlee, A. L., & Schmit, M. J. (2008). Effects of personality characteristics on knowledge, skill, and performance in servicing retail customers. *International Journal of Selection and Assessment, 16,* 272-280. doi: 10.1111/j.1468-2389.2008.00433.x

Motowidlo, S. J., & Van Scotter, J. R. (1994). Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology, 79,* 475-480. doi: 10.1037/0021-9010.79.4.475

Mumford, M. D., Connelly, M. S., Helton, W. B., Strange, J. M., & Osburn, H. K. (2001). On the construct validity of integrity tests: Individual and situational factors as predictors of test performance. *International Journal of Selection and Assessment, 9,* 240-257. doi: 10.1111/1468-2389.00177

Murphy, K. R. (1989). Is the relationship between cognitive ability and job performance stable over time? *Human Performance, 2,* 183-200. doi: 10.1207/s15327043hup0203_3

Murphy, K. R., & Cleveland, J. (1995). *Understanding performance appraisal : Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.

Murphy, K. R., Cleveland, J. N., & Mohler, C. J. (2001). Reliability, validity, and meaningfulness of multisource ratings. In D. W. Bracken, C. W. Timmreck & A. H. Church (Eds.), *The handbook of multisource feedback: The comprehensive resource for designing and implementing MSF processes* (pp. 130-148). San Francisco, CA: Jossey-Bass.

Murphy, K. R., Garcia, M., Kerkar, S., Martin, C., & Balzer, W. K. (1982). Relationship between observational accuracy and accuracy in evaluating performance. *Journal of Applied Psychology, 67,* 320-325. doi: 10.1037/0021-9010.67.3.320

Murphy, K. R., Myors, B., & Wolach, A. (2009). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (3rd ed.). New York, NY: Routledge/Taylor & Francis Group.

Nathan, B. R., & Lord, R. G. (1983). Cognitive categorization and dimensional schemata: A process approach to the study of halo in performance ratings. *Journal of Applied Psychology, 68,* 102-114. doi: 10.1037/0021-9010.68.1.102

Neely, A. (Ed.). (2002). *Business performance measurement*: Cambridge University Press.

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A., Brody, N., Ceci, S. J., et al. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51,* 77-101. doi: 10.1037/0003-066X.51.2.77

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures : issues and applications*. Thousand Oaks, CA: Sage.

Neubauer, A. C., & Fink, A. (2009). Intelligence and neural efficiency: Measures of brain activation versus measures of functional connectivity in the brain. *Intelligence, 37,* 223-229. doi: 10.1016/j.intell.2008.10.008

No authorship, i. (1921). Intelligence and its measurement: A symposium. *Journal of Educational Psychology, 12,* 123-147. doi: 10.1037/h0076078

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Ogbu, J. U., & Stern, P. (2001). Caste status and intellectual development. In R. J. Sternberg & E. L. Grigorenko (Eds.), *Environmental effects on cognitive abilities* (pp. 3-37). Mahwah, NJ: Lawrence Erlbaum Associates.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Salgado, J. F. (2010). Cognitive abilities. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 255-275). New York, NY: Routledge/Taylor & Francis Group.

Ones, D. S., & Viswesvaran, C. (2007). A research note on the incremental validity of job knowledge and integrity tests for predicting maximal performance. *Human Performance, 20,* 293-303.

Ones, D. S., Viswesvaran, C., & Dilchert, S. (2005). Cognitive ability in selection decisions. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 431-468). Thousand Oaks, CA: Sage.

Organ, D. W., & Paine, J. B. (1999). A new kind of performance for industrial and organizational psychology: Recent contributions to the study of organizational citizenship behavior. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology 1999, Vol 14* (pp. 337-368). New York, NY: John Wiley & Sons Ltd.

Ortiz, S. O., & Dynda, A. M. (2005). Use of intelligence tests with culturally and linguistically diverse populations. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 545-556). New York, NY: Guilford Press.

Parker, S. K., & Wall, T. D. (2002). Work design: Learning from the past and mapping a new terrain. In N. Anderson, D. S. Ones, H. K. Sinangil & C. Viswesvaran (Eds.), *Handbook of industrial, work and organizational psychology, volume 1: Personnel psychology* (pp. 90-109). Thousand Oaks, CA: Sage.

Penley, L. E., Alexander, E. R., Jernigan, I., & Henwood, C. I. (1991). Communication abilities of managers: The relationship to performance. *Journal of Management, 17,* 57-76. doi: 10.1177/014920639101700105

Perkins, D. N., & Tishman, S. (2001). Dispositional aspects of intelligence. In J. M. Collis & S. Messick (Eds.), *Intelligence and personality: Bridging the gap in theory and measurement* (pp. 233-257). Mahwah, NJ: Lawrence Erlbaum Associates.

Pervin, L. A. (2001). Persons in context: Defining the issues, units, and processes. In J. M. Collis & S. Messick (Eds.), *Intelligence and personality: Bridging the gap in theory and measurement* (pp. 307-317). Mahwah, NJ: Lawrence Erlbaum Associates.

Petrill, S. A. (2002). The case for general intelligence: A behavioral genetic perspective. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The general factor of intelligence: How general is it?* (pp. 281-298). Mahwah, NJ: Lawrence Erlbaum Associates.

Piaget, J. (1947). *The psychology of intelligence*. Oxford, England: Armand Colin.

Piaget, J. (1951). The biological problem of intelligence. In D. Rapaport (Ed.), *Organization and pathology of thought: Selected sources* (pp. 176-192). New York, NY: Columbia University Press.

Piaget, J. (1961). The genetic approach to the psychology of thought. *Journal of Educational Psychology, 52,* 275-281. doi: 10.1037/h0042963

Piaget, J. (1980). *Intelligence and adaptation: Organic selection and phenocopy* (S. Eames, Trans.). Paris: Hermann.

Ployhart, R. E., & Hakel, M. D. (1998). The substantive nature of performance variability: Predicting interindividual differences in intraindividual performance. *Personnel Psychology, 51,* 859-901. doi: 10.1111/j.1744-6570.1998.tb00744.x

Psytech International Ltd. (2006). *General and graduate reasoning tests*. Technical Manual retrieved from http://www.psytech.co.uk/downloads/manuals/grt2Man.pdf

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology, 69,* 581-588. doi: 10.1037/0021-9010.69.4.581

Pulakos, E. D. (2007). Performance measurement. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement : Industrial psychology in human resources management* (pp. 293-317). New York, NY: Taylor & Francis Group/Lawrence Erlbaum Associates.

Pulakos, E. D., Arad, S., Donovan, M. A., & Plamondon, K. E. (2000). Adaptability in the workplace: Development of a taxonomy of adaptive performance. *Journal of Applied Psychology, 85,* 612-624. doi: 10.1037/0021-9010.85.4.612

Pulakos, E. D., & O'Leary, R. S. (2010). Defining and measuring results of workplace behavior. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 513-529). New York, NY: Routledge/Taylor & Francis Group.

Putka, D. J., & Sackett, P. R. (2010). Reliability and validity. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 9-49). New York, NY: Routledge/Taylor & Francis Group.

Ramey, C. T., Ramey, S. L., & Lanzi, R. G. (2001). Intelligence and experience. In R. J. Sternberg & E. L. Grigorenko (Eds.), *Environmental effects on cognitive abilities* (pp. 83-115). Mahwah, NJ: Lawrence Erlbaum Associates.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmark Paedagogiske Institut.

Reb, J., & Greguras, G. J. (2010). Understanding performance ratings: Dynamic performance, attributions, and rating purpose. *Journal of Applied Psychology, 95,* 213-220. doi: 10.1037/a0017237

Ree, M. J., Carretta, T. R., & Steindl, J. R. (2002). Cognitive ability. In N. Anderson, D. S. Ones, H. K. Sinangil & C. Viswesvaran (Eds.), *Handbook of industrial, work and organizational psychology, volume 1: Personnel psychology* (pp. 219-232). Thousand Oaks, CA: Sage.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than g. *Journal of Applied Psychology, 79,* 518-524. doi: 10.1037/0021-9010.79.4.518

Reeve, C. L., Heggestad, E. D., & Lievens, F. (2009). Modeling the impact of test anxiety and test familiarity on the criterion-related validity of cognitive ability tests. *Intelligence, 37,* 34-41. doi: 10.1016/j.intell.2008.05.003

Robertson, I. T., & Smith, M. (2001). Personnel selection. *Journal of Occupational and Organizational Psychology, 74,* 441-472. doi: 10.1348/096317901167479

Roe, R. A. (1999). Work performance: A multiple regulation perspective. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology 1999, Vol 14* (pp. 231-335). New York, NY: John Wiley & Sons.

Rotundo, M., & Spector, P. E. (2010). Counterproductive work behavior and withdrawal. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 489-511). New York, NY: Routledge/Taylor & Francis Group.

Ruzgis, P., & Grigorenko, E. L. (1994). Cultural meaning systems, intelligence, and personality. In R. J. Sternberg & P. Ruzgis (Eds.), *Personality and intelligence* (pp. 248-270). New York, NY: Cambridge University Press.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88,* 413-428. doi: 10.1037/0033-2909.88.2.413

Sackett, P. R., & DeVore, C. J. (2002). Counterproductive behaviors at work. In N. Anderson, D. S. Ones, H. K. Sinangil & C. Viswesvaran (Eds.), *Handbook of industrial, work and organizational psychology, volume 1: Personnel psychology* (pp. 145-164). Thousand Oaks, CA: Sage.

Sackett, P. R., & Lievens, F. (2008). Personnel selection. *Annual Review of Psychology, 59,* 419-450. doi: 10.1146/annurev.psych.59.103006.093716

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology, 73,* 482-486. doi: 10.1037/0021-9010.73.3.482

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F., & Rolland, J. P. (2003). A meta-analytic study of general mental ability validity for different occupations in the European community. *Journal of Applied Psychology, 88,* 1068-1081. doi: 10.1037/0021-9010.88.6.1068

Salgado, J. F., Viswesvaran, C., & Ones, D. S. (2002). Predictors used for personnel selection: An overview of constructs, methods and techniques. In N. Anderson, D. S. Ones, H. K. Sinangil & C. Viswesvaran (Eds.), *Handbook of industrial, work and organizational psychology, volume 1: Personnel psychology* (Vol. 1, pp. 165-199). Thousand Oaks, CA: Sage.

Schaie, K. W., & Zuo, Y.-L. (2001). Family environments and adult cognitive functioning. In R. J. Sternberg & E. L. Grigorenko (Eds.), *Environmental effects on cognitive abilities* (pp. 337-361). Mahwah, NJ: Lawrence Erlbaum Associates.

Schleicher, D. J., Venkataramani, Y., Morgeson, F. P., & Campion, M. A. (2006). So you didn't get the job... Now what do you think? Examining opportunity-to-perform fairness perceptions. *Personnel Psychology, 59,* 559-590.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262-274. doi: 10.1037/0033-2909.124.2.262

Schmidt, F. L., & Hunter, J. E. (2002). Meta-analysis. In N. Anderson, D. S. Ones, H. K. Sinangil & C. Viswesvaran (Eds.), *Handbook of industrial, work and organizational psychology, volume 1: Personnel psychology* (pp. 51-70). Thousand Oaks, CA: Sage.

Schmidt, F. L., & Hunter, J. E. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86,* 162-173. doi: 10.1037/0022-3514.86.1.162

Schmidt, F. L., & Kaplan, L. B. (1971). Composite vs multiple criteria: A review and resolution of the controversy. *Personnel Psychology, 24,* 419-434. doi: 10.1111/j.1744-6570.1971.tb00365.x

Schmidt, F. L., Shaffer, J. A., & Oh, I.-S. (2008). Increased accuracy for range restriction corrections: Implications for the role of personality and general mental ability in job and training performance. *Personnel Psychology, 61,* 827-868.

Schmitt, N., Cortina, J. M., Ingerick, M. J., & Wiechmann, D. (2003). Personnel selection and employee performance. In W. C. Borman, D. R. Ilgen & R. J. Klimoski (Eds.), *Handbook of psychology: Industrial and organizational psychology, Vol 12* (pp. 77-105). Hoboken, NJ: John Wiley & Sons.

Schulte, M. J., Ree, M. J., & Carretta, T. R. (2004). Emotional intelligence: Not much more than g and personality. *Personality and Individual Differences, 37,* 1059-1068. doi: 10.1016/j.paid.2003.11.014

Schulz-Hardt, S., Mojzisch, A., & Vogelgesang, F. (2008). Dissent as a facilitator: Individual- and group-level effects on creativity and performance. In C. K. W. De Dreu & M. J. Gelfand (Eds.), *The*

psychology of conflict and conflict management in organizations (pp. 149-177). New York, NY: Taylor & Francis Group/Lawrence Erlbaum Associates.

Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Ashland, OH: Hogrefe & Huber.

Shultz, K. S., & Whitney, D. J. (2005). *Measurement theory in action : case studies and exercises*. Thousand Oaks, CA: Sage.

Smith, D., & Bar-Eli, M. (Eds.). (2007). *Essential readings in sport and exercise psychology*. Champaign, IL: Human Kinetics.

Smith, P. C., & Kendall, L. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47,* 149-155. doi: 10.1037/h0047060

Smither, J. W., & London, M. (2009). *Performance management : putting research into action*. San Francisco, CA: Jossey-Bass.

Snow, R. E. (1994). Abilities and aptitudes. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence*. New York, NY: Macmillan.

Snow, R. E. (1996). Aptitude development and education. *Psychology, Public Policy, and Law, 2,* 536-560. doi: 10.1037/1076-8971.2.3-4.536

Somech, A., Desivilya, H. S., & Lidogoster, H. (2009). Team conflict management and team effectiveness: The effects of task interdependence and team identification. *Journal of Organizational Behavior, 30,* 359-378. doi: 10.1002/job.537

Sonnentag, S. (2000). Excellent performance: The role of communication and cooperation processes. *Applied Psychology: An International Review, 49,* 483-497. doi: 10.1111/1464-0597.00027

Sonnentag, S., & Frese, M. (2005). Performance concepts and performance theory. In S. Sonnentag (Ed.), *Psychological management of individual performance* (pp. 1-25). Chichester, England: John Wiley & Sons.

Sonnentag, S., & Volmer, J. (2010). What you do for your team comes back to you: A cross-level investigation of individual goal specification, team-goal clarity, and individual performance. *Human Performance, 23,* 116-130. doi: 10.1080/08959281003622164

Spearman, C. E. (1904). 'General intelligence,' objectively determined and measured. *The American Journal of Psychology, 15,* 201-293. doi: 10.2307/1412107

Spearman, C. E. (1927). *The abilities of man : their nature and measurement*. London, England: Macmillan.

Spearman, C. E., & Jones, L. W. (1950). *Human ability : a continuation of "The abilities of man"*. London, England: Macmillan.

Spicer, J. (2005). *Making sense of multivariate data analysis*. Thousand Oaks; CA: Sage.

Stankov, L. (2002). g: A diminutive general. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The general factor of intelligence: How general is it?* (pp. 19-37). Mahwah, NJ: Lawrence Erlbaum Associates.

Stern, W. (1949). The intelligence quotient. In W. Dennis (Ed.), *Readings in general psychology* (pp. 338-341). New York, NY: Prentice-Hall.

Stern, W., & Spoerl, H. D. (1938). *General psychology: From the personalistic standpoint*. New York, NY: MacMillan.

Stern, W., & Whipple, G. M. (1914). *The psychological methods of testing intelligence*. Baltimore, MD: Warwick & York.

Sternberg, R. J. (1985). *Beyond IQ :A triarchic theory of human intelligence*. New York, NY: Cambridge University Press.

Sternberg, R. J. (1990). *Metaphors of mind : conceptions of the nature of intelligence*: Cambridge University Press.

Sternberg, R. J. (1997). *Successful Intelligence*. New York, NY: Plume.

Sternberg, R. J. (2000). *Practical intelligence in everyday life*: Cambridge University Press.

Sternberg, R. J. (2001). Successful intelligence: Understanding what Spearman had rather than what he studied. In J. M. Collis & S. Messick (Eds.), *Intelligence and personality: Bridging the gap in theory and measurement* (pp. 347-373). Mahwah, NJ: Lawrence Erlbaum Associates.

Sternberg, R. J. (2003a). WICS: A model of leadership in organizations. *Academy of Management Learning & Education, 2*, 386-401.

Sternberg, R. J. (2003b). *Wisdom, intelligence, and creativity synthesized*. New York, NY: Cambridge University Press.

Sternberg, R. J. (2005). Intelligence, competence, and expertise. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 15-30). New York, NY: Guilford Press.

Sternberg, R. J., Conway, B. E., Ketron, J. L., & Bernstein, M. (1981). People's conceptions of intelligence. *Journal of Personality and Social Psychology, 41,* 37-55. doi: 10.1037/0022-3514.41.1.37

Sternberg, R. J., & Detterman, D. K. (Eds.). (1986). *What is intelligence? Contemporary viewpoints on its nature and definition*. Norwood, NJ: Ablex.

Sternberg, R. J., & Gardner, M. K. (1982). A componential interpretation of the general factor in human intelligence. In H. J. Eysenck (Ed.), *A model for intelligence* (pp. 231-254). Berlin: Springer-Verlag.

Sternberg, R. J., & Grigorenko, E. L. (Eds.). (2002). *The general factor of intelligence : How general is it?* Mahwah, NJ: Lawrence Erlbaum Associates.

Sternberg, R. J., & Hedlund, J. (2002). Practical intelligence, g, and work psychology. *Human Performance, 15,* 143-160.

Sternberg, R. J., & Kaufman, J. C. (1998). Human abilities. *Annual Review of Psychology, 49,* 479-502. doi: 10.1146/annurev.psych.49.1.479

Sternberg, R. J., & Wagner, R. K. (1993). The g-ocentric view of intelligence and job performance is wrong. *Current Directions in Psychological Science, 2,* 1-5. doi: 10.1111/1467-8721.ep10770441

Sternberg, R. J., Wagner, R. K., Williams, W. M., & Horvath, J. A. (1995). Testing common sense. *American Psychologist, 50,* 912-927. doi: 10.1037/0003-066X.50.11.912

Strelau, J., Zawadzki, B., & Piotrowska, A. (2001). Temperament and intelligence: A psychometric approach to the links between both phenomena. In J. M. Collis & S. Messick (Eds.), *Intelligence and personality: Bridging the gap in theory and measurement* (pp. 61-78). Mahwah, NJ: Lawrence Erlbaum Associates.

Suss, H.-M., & Beauducel, A. (2005). Faceted models of intelligence. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 313-332). Thousand Oaks, CA: Sage.

Tannenbaum, S. I., & Yukl, G. (1992). Training and development in work organisations. *Annual Review of Psychology, 43,* 399-441. doi: 10.1146/annurev.ps.43.020192.002151

Taylor, P. J. (2003). Performance management and appraisal. In M. O'Driscoll, P. Taylor & T. Kalliath (Eds.), *Organsational psychology in Australia and New Zealand* (pp. 78-103). Melbourne: Oxford University Press.

Taylor, P. J., Keelty, Y., & McDonnell, B. (2002). Evolving personnel selection practices in New Zealand organisations and recruitment firms. *New Zealand Journal of Psychology, 31,* 8-18.

Taylor, P. J., 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.

Terman, L. M. (1911). The Binet-Simon scale for measuring intelligence: Impressions gained by its application. *Psychology Clinic, 5,* 199-206.

Terman, L. M. (1916). *The measurement of intelligence: An explanation of and a complete guide for the use of the Stanford revision and extension of the Binet-Simon Intelligence Scale*. Boston, MA: Houghton, Mifflin and Company.

Terman, L. M., Lyman, G., Ordahl, G., Ordahl, L. E., Galbreath, N., & Talbert, W. (1917). *The Stanford revision and extension of the Binet-Simon scale for measuring intelligence*. Baltimore, MD: Warwick & York.

Thorndike, E. L. (1898). *Animal intelligence: An experimental study of the associative processes in animals*. New York, NY: Columbia University Press.

Thorndike, E. L. (1901). *The human nature club: An introduction to the study of mental life* (2nd ed.). New York, NY: Longmans, Green and Co.

Thorndike, E. L. (1903). *Educational psychology*. New York, NY: Lemcke & Buechner.

Thorndike, E. L. (1904). *Theory of mental and social measurements*. New York, NY: The Science Press.

Thorndike, E. L. (1920). Intelligence and its uses. *Harper's Magazine, 140,* 227-235.

Thorndike, R. L. (1994). g. *Intelligence, 19,* 145-155. doi: 10.1016/0160-2896%2894%2990010-8

Thorndike, R. M. (1994). Thorndike, Edward L. (1874-1949). In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence*. New York, NY: Macmillan.

Thurstone, L. L. (1923). The nature of general intelligence and ability. *British Journal of Psychology, 14,* 241-247.

Thurstone, L. L. (1933). *The theory of multiple factors*. New York, NY: Harcourt Brace & Company.

Thurstone, L. L. (1938). *Primary mental abilities*: University of Chicago Press.

Undheim, J. O., & Horn, J. L. (1977). Critical evaluation of Guilford's structure-of-intellect theory. *Intelligence, 1,* 65-81. doi: 10.1016/0160-2896%2877%2990027-7

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York, NY: Springer.

van der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review, 113,* 842-861. doi: 10.1037/0033-295X.113.4.842

van Knippenberg, D. (2000). Work motivation and performance: A social identity perspective. *Applied Psychology: An International Review, 49,* 357-371. doi: 10.1111/1464-0597.00020

van Knippenberg, D., & Schippers, M. C. (2007). Work group diversity. *Annual Review of Psychology, 58,* 515-541. doi: 10.1146/annurev.psych.58.110405.085546

van Scotter, J., Motowidlo, S. J., & Cross, T. C. (2000). Effects of task performance and contextual performance on systemic rewards. *Journal of Applied Psychology, 85,* 526-535. doi: 10.1037/0021-9010.85.4.526

Vernon, P. A., & Jensen, A. R. (1984). Individual and group differences in intelligence and speed of information processing. *Personality and Individual Differences, 5,* 411-423. doi: 10.1016/0191-8869%2884%2990006-0

Vernon, P. E. (1969). *Intelligence and cultural environment*. London, England: Methuen.

Viswesvaran, C. (2002). Assessment of individual job performance: A review of the past century and a look ahead. In N. Anderson, D. S. Ones, H. K. Sinangil & C. Viswesvaran (Eds.), *Handbook of industrial, work and organizational psychology, volume 1: Personnel psychology* (pp. 110-126). Thousand Oaks, CA: Sage.

Viswesvaran, C., & Ones, D. S. (2000). Perspectives on models of job performance. *International Journal of Selection and Assessment, 8,* 216-226.

Viswesvaran, C., & Ones, D. S. (2005). Job performance: Assessment issues in personnel selection. In A. Evers, N. Anderson & O. Voskuijl (Eds.), *The Blackwell handbook of personnel selection* (pp. 354-375). Malden, MA: Blackwell

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

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90,* 108-131. doi: 10.1037/0021-9010.90.1.108

Vogel, R. M., & Feldman, D. C. (2009). Integrating the levels of person-environment fit: The roles of vocational fit and group fit. *Journal of Vocational Behavior, 75,* 68-81. doi: 10.1016/j.jvb.2009.03.007

von Mayrhauser, R. T. (1994). Yerkes, Robert M. (1876-1956). In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 1165-1175). New York, NY: Macmillan.

Vygotsky, L. S. (1962). *Thought and language* (E. Hanfmann & G. Vakar, Trans.). Oxford, England: Wiley.

Vygotsky, L. S., Rieber, R. W., & Hall, M. J. (1997). *The collected works of L. S. Vygotsky, Vol. 4: The history of the development of higher mental functions*. New York, NY: Plenum Press.

Wagner, R. K., & Sternberg, R. J. (1985). Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology, 49,* 436-458. doi: 10.1037/0022-3514.49.2.436

Waiter, G. D., Deary, I. J., Staff, R. T., Murray, A. D., Fox, H. C., Starr, J. M., et al. (2009). Exploring possible neural mechanisms of intelligence differences using processing speed and working memory tasks: An fMRI study. *Intelligence, 37,* 199-206. doi: 10.1016/j.intell.2008.09.008

Wallace, S. (1965). Criteria for what? *American Psychologist, 20,* 411-417. doi: 10.1037/h0022446

Wasserman, J. D., & Tulsky, D. S. (2005). A history of intelligence assessment. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 3-22). New York, NY: Guilford Press.

Wherry, R. J., & Bartlett, C. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology, 35,* 521-551. doi: 10.1111/j.1744-6570.1982.tb02208.x

Whiting, S. W., Podsakoff, P. M., & Pierce, J. R. (2008). Effects of task performance, helping, voice, and organizational loyalty on performance appraisal ratings. *Journal of Applied Psychology, 93,* 125-139. doi: 10.1037/0021-9010.93.1.125

Wildman, J. L., Bedwell, W. L., Salas, E., & Smith-Jentsch, K. A. (2011). Performance measurement at work: A multilevel perspective. In S. Zedeck (Ed.), *APA handbook of industrial and*

organizational psychology, Vol 1: Building and developing the organization (pp. 303-341). Washington, DC: American Psychological Association.

Wilhelm, O., & Engle, R. W. (2005). Intelligence: A diva and a workhorse. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 1-9). Thousand Oaks, CA: Sage.

Wood, P. Q., & Englert, P. (2009). Intelligence compensation theory: A critical examination of the negative relationship between conscientiousness and fluid and crystallised intelligence. *The Australian and New Zealand Journal of Organisational Psychology, 2,* 19-29. doi: 10.1375/ajop.2.1.19

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 111-153). Westport, CT: Praeger.

Yerkes, R. M. (1912). The intelligence of earthworms. *Journal of Animal Behavior, 2,* 332-352. doi: 10.1037/h0072456

Yerkes, R. M. (1943a). *Chimpanzees: a laboratory colony*. New Haven, CT: Yale University Press.

Yerkes, R. M. (1943b). The Yale Laboratories of Primate Biology. *Scientific Monthly, New York, 56,* 287-290.

Yerkes, R. M., Bridges, J. W., & Hardwick, R. S. A. (1915). *Point scale for measuring mental ability*. Oxford, England: Warwick & York.

Yerkes, R. M., & Learned, B. W. (1925). *Chimpanzee intelligence and its vocal expressions*. Oxford, England: Williams & Wilkins Co.

Zeidner, M., & Matthews, G. (2005). Evaluation anxiety: Current theory and research. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 141-163). New York, NY: Guilford Press.

Zyphur, M. J., Chaturvedi, S., & Arvey, R. D. (2008). Job performance over time is a function of latent trajectories and previous performance. *Journal of Applied Psychology, 93,* 217-234.