SOME METHODOLOGICAL ISSUES IN DEVELOPING A TEST OF GENERAL ABILITY FOR THE ARMY

Anthony James Cotton

Thesis submitted in complete fulfilment of the requirements for the degree of Master of Science at the Australian National University

May 1992

Declaration

I declare that this thesis reports my original work, and that no part of it has been previously accepted or presented for the award of a degree or diploma by any university. To the best of my knowledge, no material previously published or written by another person is included except where due acknowledgement is given.

A.J. Cotton
May, 1992

Acknowledgements

I would like to thank the following individuals for the parts they played in the completion of this thesis. Doctor Don Byrne who, as my supervisor, was prepared to take me on as a student at the ANU. My principal advisers for the thesis, Professor Reg Marsh and Doctor Richard Bell who provided me with excellent guidance in both the writing of the thesis and in dealing with the experimental and methodological issues that arose during the course of the work. Naturally any errors are entirely my own responsibility.

I would also like to thank Colonel Wal Hall, the Director of Psychology - Army, who provided me with valuable support and encouragement, particularly in the initial stages; and my Commanding Officer, Lieutenant Colonel Ron Furry without whose continuing support the thesis would not have been completed. Finally, I would like to thank my wife Leanne who has been extremely supportive and understanding throughout.

Abstract

The aim of this study was to investigate the model of intelligence and the test theory used in the actual development of a test of general ability. It was hypothesised that a three factor model of intelligence, comprising verbal, numeric, and spatial abilities, would provide as good a fit to the data as a similar model including fluid intelligence. It was also hypothesised that a test based on Classical Test Theory (CTT) and one based on Item Response Theory (IRT), although having different item compositions, would perform similarly in predicting scores on an existing general ability test. It was further hypothesised that, for the IRT based test, examinees' number right score would provide an adequate approximation of the formal IRT ability estimates.

A 52 item test was administered to a sample (N=209) to investigate the factor structure and to develop the IRT and CTT tests. A factor analysis of the items and a reliability analysis of the scales showed good support for the three factor model of intelligence over the four factor model. Two twenty item tests were then developed from this item set and administered to a second sample (N=371). These results showed strong support for both hypotheses, namely, there was little difference between the CTT and IRT based tests in the amount of criterion variance they predicted; and number right scores and IRT ability estimates showed extremely high correlations.

TABLE OF CONTENTS

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

The Australian Army Psychology Corps is one of the largest organisations of psychologists in Australia, employing nearly 50 psychologists and a similar number of clerical and administrative staff. The Corps has a number of roles and one of the principal ones is to act as the "quality control" for the Army's recruiting efforts screening in excess of 15,000 people per annum for the various avenues of entry to the Australian Regular Army (ARA). The role of the screening is twofold; first to identify those candidates who lack the necessary traits to complete initial training; and secondly, to provide adequate information for the future training and allocation of the individual.

The selection procedures for the different avenues of entry to the Army have a common foundation, that is:

a. some measure of general ability,

b. some measure of ability or aptitude that is specific to the avenue of entry,

c. some measure of personality or coping ability,

d. an interview with an Army psychologist, and

e. a final screening process where the examinee is considered in competition with those others applying for the same entry.

There are three types of entrant to the Army; General Entry (GE) applicants, Officer and Apprentice applicants. GEs provide by far the bulk of Army entrants with about 6,000 tested each year.

1.1 General Entry Selection

Applicants for general entry to the Army must be between 17 and 34 years of age when they enlist and must otherwise be literate enough to complete the application form. This is a straightforward document requiring the applicant to provide information about family history, nationality, basic employment and education history.

After making initial contact with an Army recruiter the individual will complete and submit an application form and will be given a date to present for "testing" where he or she will undergo further screening for entry to the Army.

On the test day the examinee will go through the following process (in the order specified):

a. a preliminary medical screening;

b. a psychometric test battery consisting of a 14 item test of written comprehension called the Recruiting Office Form 1

(RO1), a 100 item test of general ability called the Army General Classification Test (AGC) and a 52 item anxiety checklist called the Self Descriptive Inventory (SDI), all of which have all been developed specifically for the Army;

c. a structured interview with an Army psychologist which covers family history, education, employment history, motivation, and expectations for recruit training;

d. a complete medical examination; and

e. an interview with the Enlistment Officer which covers similar ground to what the psychologist investigates and also addresses drug use, criminal record, etc.

All of the information is collated by the Enlistment Officer who has the final say on the individual's success or otherwise on the day. The successful examinee will be enlisted at some later stage (normally two to six weeks depending on circumstances).

1.1.1 Recruit Training

Recruit training involves an intensive 13 week training course covering topics ranging from physical fitness, to navigation to elements of military law. As well as completing these studies, the recruit spends considerable time (often hours per day) preparing and maintaining his or her equipment and uniforms.

Recruit training places considerable other pressures on the recruit. In particular the socialisation process the recruit undergoes and the generally dangerous training undertaken, for example rifle shooting, grenade practise, and gas training. The recruit training environment requires a great deal of the recruits' personal resources, over and above their cognitive abilities.

1.1.2 Allocation to Initial Trade Training

Further psychometric testing is conducted during recruit training which involves tests of arithmetic and mathematical achievement, clerical aptitude and mechanical comprehension. The results of these tests are combined with the recruit's initial testing to provide a psychometric profile which is fundamental to the allocation of the recruit to initial employment training.

Immediately on completion of recruit training the soldier will commence Initial Employment Training (IET). Employments available for IET range from Medical Assistant and Rifleman, which have few prerequisites for allocation in terms of cognitive ability, through to Electronics Technician which requires a high level of general ability and mathematics ability.

1.2 The Importance of Psychological Testing

The initial psychological testing serves two purposes: it is used to screen the individual for an complex selection/rejection process, and it is used for allocating those selected to a very broad range of employments. Finally, it must also be able to provide valid information and stand up to scrutiny at indeterminate points in the future.

Two important points need to be considered; first, while there is a cognitive component in recruit training, the bulk of the stressors are essentially non-cognitive, the communal living, the discipline, the physical requirements. Secondly, the Army enlists large numbers of individuals and only those who are considered manifestly unsuitable are screened out (e.g. the bottom 5-6% based on performance on the AGC, among other criteria).

Thus, on the surface of it, the screening instrument need only provide a measure of the individual’s functioning level at the particular cut point chosen to maximise its effectiveness. In this case at the very lower end of the scale. Certainly there can be only limited predictive value in a test of cognitive ability when the bulk of the stressors for the recruit are non-cognitive. More may be gained by making the test highly accurate around the cutting point rather than providing a reasonable level of information across a more broad range of ability.

This would make the task of the initial test of general ability seem, on the surface, fairly easy because the test would only need to provide discrimination at the lower end of the ability spectrum. Unfortunately the test also needs to be able to identify potentially successful candidates for intensive and lengthy vocational courses (electronics trade as an example). Thus the test needs considerably more versatility than it would seem on first appearances.

The Current Instrument. The current instrument used by the Army for the selection and allocation of general entry soldiers is a 100 item spiral omnibus format test of general ability, the Army General Classification test (AGC). This test was purpose built for the Army and has been in use for about 30 years, but until 1987 was only used for allocating recruits and their future career management. In 1987 it was introduced for selection as well.

1.3 The Current Problem

Due to changes in the availability of another test used by AA Psych Corps, the AGC was reviewed in 1989. This highlighted the age of the instrument and some other limitations of the test and the testing procedure (e.g. the time taken to administer the test, the fact that only a single general ability score was available, etc). It was decided that the test was lacking in face validity and that it should be replaced.

This decision led to a set of test guide-lines being developed by the Directorate of Psychology - Army, the salient features of these were that:

a. the test should provide (at least) three separate scores, namely verbal, numerical and spatial abilities;

b. the test should be stable and not dated over 10-20 years if possible;

c. the test should be free of (unspecified) bias; and

d. the test should take advantage of the current test development and administration technology.

These guidelines raised a number of issues for the development of the test. The first was an organisational constraint that the test must be able to discriminate between different abilities, this dictates much of the form of the construction of the test. Also the construction of the test is necessarily constrained, seriously, by the other three requirements as they are fundamental to the design of any test. These organisational constraints highlighted the two aspects that are fundamental to any test design project, namely the model of ability that provides the foundation of the test and the test development methodology to be used in constructing the test. These would have to be investigated as part of the test development process.

1.4 Chapter Summary

The Army tests a broad range of individuals for entry to the Army. The results of this testing is not only used for selecting individuals, it is also critical in the allocation of soldiers and forms a permanent record for the future management of the soldier.

A review of the existing selection test found that it lacked face validity and needed to be replaced. Test guide-lines were developed and these indicated that it was necessary to investigate both the model of ability on which the new test would be based and the test development methodology which was to be used in the development of the test.

CHAPTER 2

MODELS OF INTELLIGENCE

Intelligence, along with personality, is one of the most written about topics in psychology. The list of individuals who have written in the field reads like a "Who's Who" of psychology; Spearman, Thurstone, Terman, Cattell, Eysenck, Wechsler, Burt, Vernon; the list seems endless. Unfortunately, with so many researchers working in the field, one tends to get a range of theories proliferating, not to mention a range of definitions of intelligence.

There are many definitions of intelligence and they cover a wide range of approaches from the totally operational definition of Boring (1929) that intelligence is whatever is measured by intelligence tests, to "physiological" definitions concerning the connections of the synapses in the brain, to "biological" definitions that describe intelligence in terms of its impact on the organism's ability to adapt to its environment, to psychological definitions that refer to the types of thought processes that comprise "intelligent behaviour" (usually abstract thinking). Many different approaches have been taken to the study of intelligence over the years; the developmental approach of Binet and Piaget, the psychometric approach, the "neo-behaviourist" approach (Butcher, 1968), and the recent componential approach of the cognitive scientists.

Despite the proliferation of theories, there are two main paradigms in intelligence today; these are the information-processing (or cognitive) theories of people like Sternberg (1985), Jensen (1982), and Pellegrino and Kail (1982), and the "traditional" psychometric approach of people like Thurstone (1938), Guilford (1967), Cattell (1971), and Horn (1988).

2.1 The Cognitive Paradigm

The cognitive approach to intelligence can be seen as largely the result of the enormous growth in cognitive psychology in recent years. Although relatively recent, the origins of this approach date back to some of the earliest work done in the area of measuring human differences, namely that of Francis Galton in the latter half of the 19th Century. Indeed, Spearman's early work took as much cognizance of "Sensory Discrimination" as it did of academic performance (Spearman, 1904).

The early researchers in the experimental psychology field were primarily concerned with measuring individuals on basic sensory discriminations, e.g. using visual and auditory stimuli, and tactile stimuli such as telling the distance between two pin pricks on one's finger. Using these procedures the hypothesis developed, particularly among psychologists with some physiological training, that there was "...a general cognitive capacity probably dependant upon the number, complexity of connections and organization of nerve-cells in the cerebral cortex." (Butcher, 1968, p15).

Although the early work was strongly physiological in nature, the approach today is an area of cognitive science. The cognitive theorist is interested in understanding the mental "processes" that occur during "intelligent behaviour" and the theories in the field concentrate on the components or rules, that are used in solving cognitive tasks.

The concept of an "information-processing component" is fundamental to the cognitive approach. An information-processing component is a cognitive operation or function that is a building block of intelligent behaviour. Generally components include memory functions, information acquisition processes (perception) and information-processing processes.

Components have been used for the development of tests (Smith, 1986; Irvine, Dann and Anderson, 1990), for the analysis of existing tests (Carpenter, Just and Shell, 1990), the analysis of general intelligence (Hunt, 1982; Detterman, 1987; Sternberg and Gardner, 1982; Detterman, 1982) and the analysis of the factors of intelligence (Maybery, 1990; Pellegrino and Kail, 1982).

There are many theories of cognition, some dealing with specific cognitive tasks, e.g. Pellegrino and Glaser's theory of inductive reasoning (Pellegrino & Glaser, 1980, 1982), while others deal with more general cognitive aspects of intelligence. Two of the more influential writers in the field are Sternberg and Carroll who have both contributed significantly to the work in the area, particularly in the area of general theories of human cognition.

2.1.1 Sternberg's Triarchic Theory

Sternberg's (1985) theory is perhaps the best-known of the cognitive theories. It is based on three basic components of intelligence (or cognition), hence the term "triarchic", these are: the "meta-components", the "performance components", and the "knowledge acquisition components".

Meta-components. These are the higher level components that are responsible for the planning and monitoring of intelligent behaviour. They are responsible for a range of tasks that include: deciding the nature of the task at hand, selecting the lower order components required to complete the task, selecting a mental representation for the information, and monitoring the solution of the problem.

Performance Components. These are the lower order components referred to above that act on the instructions of the meta-components. While there are some general performance components, e.g. encoding the nature of the stimulus, most are specific to the task.

Knowledge Acquisition Components. These are the components required for acquiring information and storing it in memory. Sternberg (1985) cites three as the "most important" for intelligence, these are: selective encoding, selective combination, and selective comparison. Generally these components act together in a highly interactive way which makes them difficult to analyse separately.

Sternberg (1980) defines four ways in which these components can interact: direct activation of one by another, indirect activation of one by another through the action of a third, direct feedback from one to another, and indirect feedback from one to another again through the action of a third component. All control in the system rests with the meta-components which are the only type of components that are able to receive feedback from their own type.

2.1.2 Carroll's Theory

Carroll's theory differs from that of Sternberg in that the former is more specific about the components of intelligence where Sternberg takes a broader view of his components.

After investigating a number of the major tests available Carroll (1981), identified a set of ten information-processsing components. These were:

a. Monitor - during a task, this process drives the operation of the other processes.

b. Attention - this is based on the individual's expectations about the types of stimuli that are to be presented.

c. Apprehension - this is the registering of a stimulus.

d. Perceptual Integration - this is the process of integrating the perception of the stimulus and comparing it with previous representations.

e. Encoding - this is the forming of an internal representation of the stimulus.

f. Comparison - this is used to ascertain whether two stimuli are the same or similar.

g. Co-representation Formation - this is used to form a new representation, related to a pre-existing representation.

h. Co-representation Retrieval - this is the process of finding, in memory, a representation related to another representation.

i. Transformation - this is used to change a representation based on some specified rule.

j. Response Execution - this is the process of producing a response, either overt or covert, to a mental representation.

Carroll (1981) states that these cover all of the basic processes involved in a range of cognitive processes and these components can be used to form the basis of a componential task performance.

2.1.3 Application of the Cognitive Approach to the Current Problem

According to Sternberg (1985) the cognitive approach is interested in mental **processes** whereas the psychometric approach is interested in mental **structure.** The cognitive psychologist is interested in how a task is performed, and is less interested in how people differ on their performance on the task, unless it aids the understanding of the underlying process.

Despite this, the cognitive approach has been applied to the study of individual differences. Snow (1979) identified four areas of information-processing where individuals may differ. He labelled these: parameter differences, or differences on the components used; sequence differences, or differences in the order in which the same components are executed; route differences where different sequences of components are used; and strategic differences.

Sternberg (1977) suggested six areas where individuals differ: the components they use, their rules for combining components, the order of component processing, the mode of component execution, the time taken to execute a component, and the mental representation on which the component acts. Both of these examine individual differences on the components of information-processing, but the fundamental aim is to aid the understanding of cognition rather than the differentiation of individuals.

Another consideration in the application of the cognitive approach to the study of individual differences is the types of tasks involved. Due to the requirement to measure response latency as well as response accuracy, the tasks used in the cognitive approach are more effectively presented in an interactive computer environment. In fact the recent growth in availability of this type of equipment is in part responsible for the recent upsurge in work in this area.

Although there have been cases where a cognitive analysis of the types of tasks found in traditional pencil-and-paper tests have been conducted, these are necessarily post hoc analyses (Carpenter, Just and Shell, 1990). There have been attempts to apply the principles of the cognitive approach to pencil-and-paper tests but they are few and far between (Irvine, Dann, and Anderson, 1990; Smith, 1990).

The information-processing approach is inappropriate for the present Army project for two reasons:

a. Cognitive theorists are primarily concerned with identifying the fundamental components of "intelligence" and as yet have focussed little attention on using their techniques for differentiating between individuals. Some progress in this area has been made, the work of Irvine, Dann, and Anderson (1990) for the British Army is one example, but they are, as yet, few.

b. As stated above the techniques and tasks are much more suited for computer administration, involving such things as measuring latencies, real time pattern matching, etc. The current project called basically for the replacement of a pencil-and-paper test with another pencil-and-paper test. There was no provision for the acquisition of the equipment necessary to really take advantage of this approach.

This is not to say that the information-processing approach is without merit, on the contrary it is continually making advances in the understanding of how humans process information and behave cognitively. The current problem for the Army, however, is to differentiate between individuals and (generally) try to predict a pattern of human behaviour, i.e. successful completion of recruit training. This is a task for which the psychometric approach provides a much more appropriate foundation.

2.2 Psychometric Models

The psychometric approach is the traditional approach to intelligence and can be seen as being characterised by two things: first, the use of "higher" order tasks such as analogies, series etc rather than the simple perceptual or cognitive tasks used by the cognitive theorists; and secondly, the use of correlational techniques (in particular factor analysis) developed by Galton (1888), Pearson (1900), and Spearman (1904b) and extended by Burt (1940), and Thurstone (1947).

Like the cognitive approach, the psychometric area has, as a result of the number of different individuals working in the field, developed a wide range of theories to account for the structure of intelligence in humans. One of the major divisions in the many models proposed is between those that allow a general intelligence factor and those that don't (Gustafsson, 1984). These two schools of thought have also been called the British and American schools because of their location and the whole controversy was initiated as a result of the work of Spearman (1904a).

2.2.1 Spearman's Model

Spearman (1904a) was the first researcher to propose a general factor in human intelligence. His work was, in many ways, a reaction to what he perceived as the failure of experimental psychologists to find a connection between the laboratory and real life. He felt that this was due to the poor experimental methods of the researchers at the time. Generally he felt that there was a lack of problem definition, a general failure to take into account experimental error, and poor understanding and use of correlation techniques. After providing solutions to these problems Spearman examined performance on sensory discrimination tasks, a variety of academic subjects, and ratings of intellectual ability in an attempt to identify the source of individual differences of intelligence. In this study he found an almost perfect relationship between "general discrimination" and "general intelligence" and went on to claim that there existed a general factor which was common to all areas of intellectual activity.

This general factor of Spearman's, labelled "g", can be seen as representing an individual's "mental energy" (Carroll, 1991, p558), and is present in all intelligence tests along with a factor labelled "s" which represents that component that is specific to that test alone. This became known as Speaman's "Two Factor Theory" of intelligence. Spearman also developed the "tetrad difference" (Spearman, 1904b) method of analysing a matrix of correlations to demonstrate that g and s are the only significant factors.

After Spearman much of the work that included a general factor of intelligence was conducted by British psychologists, in particular Sir Cyril Burt and P.E. Vernon.

2.2.2 The Work of Burt and Vernon

While Burt (1940) generally supported Spearman's two factor theory, he divided the general or common factor (as he called it) into universal and group factors, the latter common to "a certain group of traits" (p103). The specific factors he divided into singular (relating to a single trait) and accidental (relating to a particular testing occasion) factors. Burt considered the combination of these four types of factors as a "...fundamental logical postulate..." (p103), expressing it as a mathematical equation.

Burt's model was supported empirically by his discovery much earlier (Burt, 1909) of evidence of a sensory discrimination group factor in addition to g. Later Burt (1939) was to provide fairly strong evidence for the existence of verbal and numeric group factors in the abilities of school age children. Although the general factor accounted for more variance than the group factors combined, their contribution was none-the-less significant.

Although Vernon (1950) supported Spearman's concept of general intelligence, he felt that Spearman had neglected the existence of group factors. This was because the sample sizes employed by Spearman were too small for the residual correlations between the tests and the general factor to be statistically significant. As Vernon (1950) states, "...lack of statistical significance does not disprove the existence of additional factors; it only fails to prove it." (p 14).

Vernon had empirical support for his position from an earlier (1947) study he had conducted where he analysed the results of an administration of thirteen tests to a sample of about 1,000 Army recruits. From this, the existence of two group factors was quite clear, although the variance accounted for by the general factor he also identified was double that of the group factors. Vernon labelled these; v:ed, a verbal-numerical-educational factor, and k:m, a practical-spatial-mechanical factor. He further demonstrated that these factors would divide into minor group factors if the analysis was thorough enough (in this case v:ed was divided into verbal and numerical abilities). Vernon's model is graphically depicted in Figure 1.

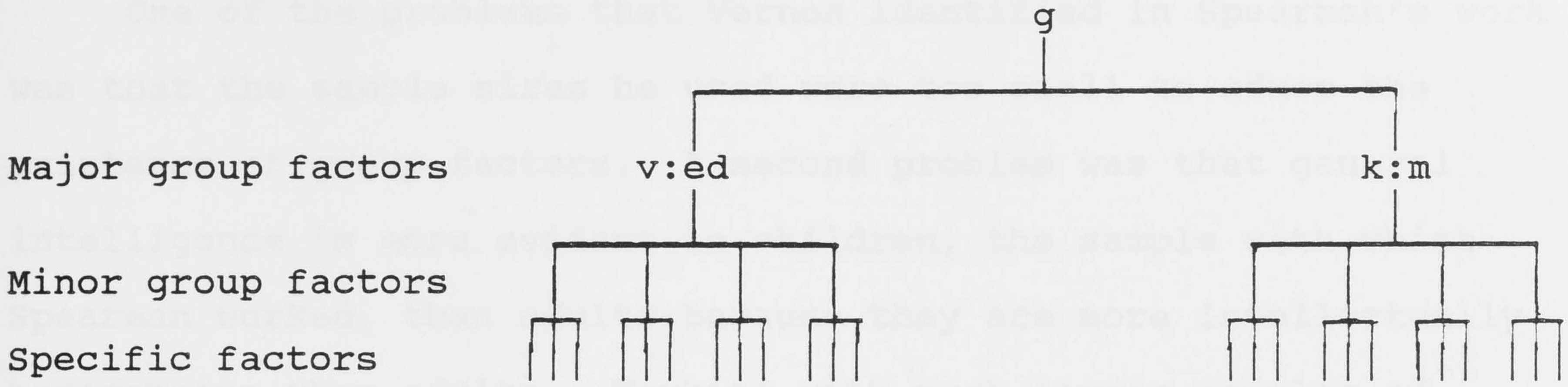

Figure 2-1: Vernon's (1950) Hierarchical Model of Abilities

Despite this support for the concept of a general factor of intelligence from Britain, in the US the concept received much less support. With the refinement of factor analysis, particularly the work done by Thurstone (1938 and 1947) in developing the procedure to allow multiple common factors and in developing the concept of "simple structure", the emergence of theories of intelligence based on a multitude of relatively equal factors occurred.

Many models of this type have been developed, but three of the most important (for a variety of reasons) were Thurstone's Primary Mental Abilities (1938), Guilford's Structure of Intellect model (1967), and the theory of fluid and crystallised intelligence developed by Cattell and Horn (Horn and Cattell, 1966).

2.2.3 Thurstone's Primary Mental Abilities

One of the problems that Vernon identified in Spearman's work was that the sample sizes he used were too small to educe the existence of group factors. A second problem was that general intelligence is more evident in children, the sample with which Spearman worked, than adults because they are more intellectually homogeneous than adults. Working with much larger samples of adults, Thurstone applied the new factor analytic techniques to large test batteries and was able to identify a number of replicable factors (Thurstone, 1940; Thurstone and Thurstone, 1941). These he reduced to seven Primary Mental Abilities (PMA): Verbal Comprehension, Numerical Facility, Induction, Deduction, Spatial Ability, Perceptual Speed, and Rote Memory.

While in his original analyses Thurstone kept his factors orthogonal, that is uncorrelated, he later allowed the factors to be correlated (oblique factors), something which greatly facilitated the attainment of simple structure within the set of factors. Simple structure, as defined by Thurstone, really means that only a small number of tests should load on any one factor and all other tests should have very small (approaching zero) loadings with the factor. Similarly, each test should load on preferably only one factor with near zero loadings on the other factors.

Simple structure is one of the primary sources of difference between this model and that of Spearman and the British school. In particular, achieving simple structure means that it is unlikely that a general factor will appear that correlates with all of the tests. Rather, each test will correlate with one or two factors and each factor will cover a separate but possibly overlapping set of tests (Vernon, 1950).

2.2.4 Guilford's Structure of Intellect Model

Guilford's model developed as an extension of Thurstone's PMAs. His work with the US military during World War 2 verified Thurstone's work and identified more primary abilities so that Guilford identified some 25 abilities (Guilford, 1985). His work after the war verified the existence of most of these factors and added to the list, bringing the total of identified factors to about 40. It was at this time that Guilford decided that some organisation of these abilities was required.

He did this by defining abilities in terms of three facets: their **contents**, that is the type of information featured; their **products**, that is the form that the types of information took; and their **operations**, the kind of mental processes that were involved. He defined different types of operations, contents and products, some empirically and some theoretically, and he conceptualised the model so that primary abilities were represented as the conjunction of a content, product and an operation.

Guilford's facets or categories were operationally defined (Guilford, 1985) as:

a. Operations - cognition, memory, divergent production, convergent production, and evaluation.

b. Contents - visual, symbolic, semantic, and behavioural.

c. Product - unit, class, relation, system, transformation, and implication.

Calling his model the Structure-of-Intellect (SOI) model Guilford further conceptualised it as a cube, with the operations, products and contents representing the three dimensions of the structure. This leads to a total of 120 primary abilities that can be represented by the model (5 x 4 x 3 = 120). It also led to the belief that the facets were orthogonal (Guilford's penchant for the use of orthogonal rotation methods in factor analysis also contributed to this belief).

Guilford later refuted this, allowing his factors to correlate with one another and showing that higher order abilities could be extracted from the SOI model (Guilford, 1984). Other researchers have also demonstrated the obliqueness of the SOI model (Kelderman, Mellenbergh and Elshout, 1981). Despite this he stopped short of supporting a hierarchical model as suggested by Burt and Vernon. In fact Guilford was quite opposed to the idea of a general factor of intelligence finding "Of some 48,000 correlations between pairs of tests, about 18% were below .10, many of them below zero..." (Guilford, 1985, p238).

Although theoretically elegant, Guilford's model has been broadly criticised on a variety of methodological and conceptual grounds (cf Horn and Knapp, 1973; Undheim and Horn, 1977; Vernon, 1979). Certainly more recent work, using more sophisticated research methods, have found his work wanting in a number of areas (Bachelor, 1989). Despite this, Guilford's model was a serious attempt at providing some order to the many factorial models of intelligence that exist.

2.2.5 Fluid and Crystallised Intelligence

One of the more popular models of ability is that of Cattell and Horn (Horn and Cattell, 1966; Cattell, 1971). This model is based on factor analysis of several orders. That is an oblique factor analysis of a set of tests yields a number of first order, or primary, factors. The intercorrelations of these factors are subsequently factor analysed to yield second order factors (secondaries), and so on.

This process has yielded a two level hierarchical model with five secondaries, these are:

a. fluid intelligence - Gf;

b. crystallised intelligence - Gc;

c. visualization capacity - Gv;

e. general speededness - Gs; and

g. general fluency - Gr.

Of these the most commonly identified secondaries are those pertaining to fluid and crystallised intelligence. Indeed the original model proposed that there was no single general factor of intelligence, but that general intelligence was a combination of Gf and Gc (Cattell, 1971) a position that is still argued (Horn, 1988).

2.2.6 Common Methodological Problems

All of the psychometric models are based on some form of factor analysis, which has been typically applied in an exploratory fashion to a set of standard tests. Until recently, and certainly in most, if not all, of the models discussed here, this analysis used a principal components analysis (PCA) of the matrix of intercorrelations of the tests for the extraction of the initial set of factors (singular or multiple). PCA provides only a rigid (i.e. orthogonal) rotation of the coordinate axes of the matrix of correlations. By deriving these in such a way as to maximise the amount of variation explained by each subsequent component, one is left with a set of linear composites equal in number to the number of variables in the original matrix, in decreasing order of the amount of variation they account for in the original data. It is then up to the researcher to try and determine the minimum number of components that adequately describe the data.

There are fundamentally two problems with this approach. First, the derived components relate only to the variables in the original data set. That is, in strict theoretical terms, the components obtained are not generalisable outside that particular data set. Secondly, there is no provision for statistically testing the fit of the model chosen. While there are many "tests" for the "correct" number of components such as the scree test (Cattell, 1966) and the Kaiser criterion (Kaiser, 1960), they are not tests in the sense of traditional inferential statistics.

The advent of more sophisticated common factor analysis techniques has gone some way to reducing the problems faced by the early factor analysts. In particular, the development of confirmatory common factor analysis (as opposed to exploratory component analysis) by Joreskog (1969) has allowed a more rigourous, statistically testable application of the factor analytic technique. This is not to say that common factor analysis is a perfect analytic tool (Marsh, Balla, and McDonald 1988); also, there are good arguments for the use of both component analysis and factor analysis (Velicer and Jackson, 1990). Confirmatory common factor analysis has, however, allowed some more statistically rigourous analyses of some common models of intelligence to be undertaken (e.g. Bachelor, 1989).

Another, related, development that has had a significant impact on psychometric methods has been the development of sophisticated techniques for covariance structure analysis (Joreskog, 1970). This has allowed some especially powerful and interesting results to be obtained from analyses of existing models of intelligence (Gustafsson, 1984; Undeheim and Gustafsson, 1987).

2.2.7 A Unifying Model

The recent revision of Cattell's model by Gustafsson and Undheim (1987; Gustafsson, 1984; Undheim, 1981a, 1981b) has provided an interesting alternative to the models of Cattell and Horn and Vernon. Basically Gustafsson has found that Cattell and Horn's Gf, or fluid intelligence, is statistically identical to what Vernon (1950) called general intelligence.

Gustafsson (1984) felt that "The kinds of tests identified to measure Gf comes very close to the kind of tests that Vernon lists as measures of g in his model." (p 184). After identifying Gc with Vernon's v:ed and Gv with k:m and then allowing a third order factor, labelled g, he had what he felt was a resolution of the main differences between the two major hierarchical models of ability. Gustafsson recognised that the relationship between g and Gf was an empirical question and, following the work of Undheim (1981a, 1981b), he used covariance structure analysis to examine his model.

Gustafsson (1984) found support for his model, that is three secondaries (Gf, Gc, and Gv) and one tertiary factor g. He also found that, statistically, g was identical to Gf. This result was later verified in a series of experiments (Undheim and Gustafsson, 1987).

2.2.8 Selecting a Model

Of the psychometric models of intelligence available, those of Vernon and Cattell-Horn are, as described by Gustafsson, the "...two most important hierarchical models..." (1984, p184). Gustafsson also provides strong evidence that his combination of the two models is empirically sound and theoretically reasonable. How then does the model fit the organisational constraints set for our project, that is the provision of a score for verbal, numeric and spatial abilities?

Although not tightly defined in the original instruction, one could conceptualise our required spatial ability as Gv from the Cattell-Horn model. The remaining two required factors, verbal and numeric ability, are consistent with Vernon's (1947) findings that these were the minor group factors that comprised v:ed, or in this case Gc. Thus we would have our numeric and verbal abilities combining to form Gc and our spatial abilities representing Gc with general ability as a higher-order factor.

2.2.9 Testing the Model

Although Undheim and Gustafsson (1987) provide fairly sound support for their model, it was based on an analysis of existing standard tests. In the current situation the project will be starting from scratch, that is, writing the items rather than using existing tests, therefore it seems prudent to test the adequacy of the model using item data rather than test score data.

The main aim of this test will be to determine whether we should include some measure of fluid intelligence as a separate factor from general intelligence. Despite the evidence of Gustafsson, fluid intelligence is one of the most enduring features of the Cattell-Horn model. Also, given the way it has been conceptualised, that is as an individual's capacity to deal with new and unfamiliar problems, and given the wide roles to which the current test will be put, Gf may be one of the most important abilities that we could measure.

Testing the model will involve comparing the adequacy of fit of the three factor model (verbal, numeric, and spatial abilities) against the fit of the four factor model (verbal, numeric, spatial, and fluid abilities). Given that all of the secondaries correlate with general intelligence, the three factor model should provide at least as good a fit to a set of test items as the four factor model.

After examining the different types of items that are generally used in psychometric tests two conclusions were drawn:

a. there are generally three broad types of stimuli or content areas in test items, namely verbal, numerical and spatial;

b. test items can address abilities that are specific to the stimulus type, eg addition in a numerical item, word knowledge in a verbal item, or they can address abilities that can be generalised over stimulus, eg inductive reasoning as in analogy problem.

This latter set of abilities are those that are typically used to define Gf and thus it was decided to write four types of items, labelled as follows:

a. numerical operations, eg arithmetic, clock questions, etc;

b. verbal operations, eg word knowledge;

c. spatial operations, eg paper form board, paper folding, etc; and

d. fluid ability items, eg analogies, series, etc using all three stimulus types.

By combining these in a test we should be able to test the model by conducting a confirmatory factor analysis on the set of items and hypothesising two factor structures; a four factor structure comprising verbal, numeric, spatial and fluid ability, and a three factor structure of verbal, numeric and spatial ability. In the three factor model, the fluid ability items of appropriate stimulus type would be included in the other scales; for example, the verbal reasoning items would be included with the verbal operations items to form the verbal factor.

The Hypothesis. Our hypothesis is that the three factor structure will produce at least as good a representation of this set of items as the four factor model.

2.3 Chapter Summary

There are two main paradigms of intelligence extant today; the cognitive paradigm and the psychometric paradigm. The former seeks primarily to understand the processes involved in intelligent behaviour. Its models are characterised by identifying the basic components of intelligent behaviour and examining their operations through the accuracy and latency of responses to very simple cognitive tasks. The cognitive paradigm was seen as being difficult to operationalise for the current project.

The psychometric paradigm traditionally uses factor analytic techniques to identify the underlying structure of ability tests and items. It can be characterised by its use of more complex cognitive tasks than the cognitive paradigm and its greater reliance on differentiating between individuals.

There are many models of intelligence under the psychometric paradigm ranging from one to over one hundred factors underlying intelligence. Of all of these, Gustafsson's combination of the hierarchical models of Cattell-Horn and Vernon is considered the most empirically sound and theoretically reasonable. Furthermore it fits the organisational requirements quite well and was therefore chosen to provide the basis for the test development project.

The importance of fluid ability is difficult to discount and the current project will be developing a test from scratch rather than using developed instruments (as had Gustafsson in developing his model) and therefore it was decided that the model should be tested. This would be done by examining the factor structure of a set of items and comparing a structure that includes a fluid ability scale with one that did not. It was hypothesised that the latter would provide as good a fit to the data as the former.

CHAPTER 3

TEST THEORY

Having determined the model of ability that will be used to develop the test the next feature that is critical in any test development project is the theory chosen with which one will develop the test. This includes the theory underlying the selection of items to be included in the test, the calibration of these items and the test, and the development of the scale of measurement of the test.

Why have a test theory? Writing appropriate questions, grouping them together and counting up the number that an individual examinee gets right does not generally produce an outcome that is consistent with other tests aimed at the same target group. So consistency (reliability) is the first goal of having a test theory.

There are two main components to any test theory; the technique for selecting "good" items, and the technique by which one generates a "score" for an individual.

3.0.1 Selecting Items

The selection of the items to include in a test is obviously fundamental to the development of the test. There are two features to this process; selecting items that tap the ability that you are trying to measure, and selecting items that will give you a meaningful **measure** of the particular ability you are trying to measure.

There is hardly any point in asking a question about the history of Australia, for example, if you are trying to measure someone's ability to manipulate two dimensional figures. Thus the first step in selecting items is to draw items from appropriate content areas. This is largely an experiential and theoretical issue. The test developer, after deciding what content areas are to be measured, selects items that he believes are appropriate to that content area.

Having selected a pool of items, the test developer tests these on a sample of individuals who are representative of the intended test audience. The items must be tested because what the developer considers representative of a content area may not accord with what is really the case. Also, the sample must be somewhat representative so that the information from the sample at least pertains to the group to whom it will be administered.

3.0.2 Scoring the Test

Once the test has been developed and we are satisfied that the items in the test all tap the appropriate content area, we can administer the test. Then the next question is how are we to use the test to differentiate between individuals? This is, after all, the reason for which we have written our test.

Do we simply add up the number of items that an individual answers correctly and assign that as the person's "score"? How do we group individuals in this case; if one examinee answers one more item correctly than another examinee, is that sufficient to differentiate between them for a selection decision? What degree of error, if any, is there in the measurements made by our test?

All of these considerations reinforce the need for a test theory. Our test is a measuring instrument. Unfortunately, we don't have the luxury of being able to directly measure the things that psychologists typically want to measure. In some cases, that which we wish to measure can only be defined in terms of the instrument used to take the measurement. One is reminded of Boring's (1929) operational definition of intelligence as whatever the intelligence test measures.

Our test needs to measure that which we claim it measures, and it must be able to measure that ability consistently over repeated uses. It is for these and the above reasons that we need a test theory. Today there are two predominant test theories; Classical Test Theory (CTT) and Item Response Theory (IRT).

3.1 Classical Test Theory

This is the "traditional" test development model and until the 1970's nearly all of the psychometric tests in use were based on Classical test theory.

3.1.1 The Basic Assumptions of Classical Test Theory

Classical test theory as described by people like Gulliksen (1950) and Nunnally (1978) is fundamentally a "true score theory" (Lord and Novick, 1968). The theory states that for any trait we attempt to measure, each individual so measured has a "true score". This is based on the assumption that there is an infinite domain of items relevant to the trait and that the true score is the number of correct answers the person would make if they were administered all of the items in this domain of items (Kline, 1986).

Obviously it is not possible to administer every item from an infinite domain of items and therefore random samples of items, tests, are used. Thus an individual's score on a test will be comprised of two components, their true score on that trait and an error component (as the test is only a random sample from the item domain); and the equation for an individual's test score is:

$$X_i = T + E_i \tag{1}$$

Where X_i is the observed score on test i, T is the individual's true score on the trait (note that this is a constant) and E_i is the error component associated with this test.

Being a random sample from the domain of test items, it is assumed that the errors will be random and that the distribution of the errors will be independent of the individual's true score. It is also assumed that the distribution of errors has a mean of zero, and that the error of measurement for an individual on any two tests is uncorrelated. These are expressed mathematically as follows:

$$r(T,E_i) = 0 \tag{2}$$

$$E(E_i) = 0 \tag{3}$$

$$r(E_i,E_j) = 0 \tag{4}$$

Where r indicates the correlation coefficient, and E in Equation 3 is the expectation operator.

These assumptions form the basic structure of Classical Test Theory. It is on these equations that the methodology of Classical Test Theory is based.

3.1.2 The Methodology of Classical Test Theory

The methodology of Classical Test Theory involves the following steps:

a. developing a pool of items,

b. trialling these on a sample of subjects that is representative of the test's intended audience to obtain appropriate item statistics,

c. selecting those items which meet the test's design specifications (normally based on the obtained item statistics), and

d. assembling these into a test which is calibrated (or normed) on a representative sample of subjects.

The first of these steps, developing a pool of items, is common to any test development procedure. It is unrealistic to expect that all of the items that one writes will be exactly what is required. Rather an experimental approach is taken and one develops a pool of items that are trialled on a representative sample, with the expectation that some of the items will not be adequate for the task.

Item writing should be considered both a science and an art and is fundamental to the test development process. Briefly, item writing is an extremely complex process that requires a good knowledge of the subject matter, of the types of items that can be used, and, overall, considerable planning (see, for example, Tinkleman, 1971; Wesman, 1971). Needless to say, good item writing can contribute considerably to the development of the test (Kline, 1986).

Having developed a pool of items, they are then trialled on a representative sample to obtain two item statistics that are fundamental to the item selection process; the item difficulty and the item reliability.

Item Difficulty. The item difficulty is simply measured as the proportion of subjects who correctly answer an item; this also represents the easiness of the item and for this reason has also been more correctly called the item facility value. The value of this statistic is that an item with too few or too many individuals correctly responding will provide little discrimination between individuals, that is, too little information. Kline (1986) recommends that items with facilities between 0.2 and 0.8 can be considered for further use in a test but that items falling outside these limits provide too little discrimination to be of any practical value.

Item Reliability. An item's reliability, or discrimination as it is sometimes known, is its correlation with the overall test score. In the case of items scored dichotomously (i.e. right or wrong, as is generally the case with ability test items) this is the biserial correlation (Lord & Novick, 1968). This statistic tells how much the item contributes to the scale score and gives an indication of how important the item is to the scale score. The higher the item reliability the better the particular item relates to that scale as measured with that sample. There is, theoretically, no upper bound for item reliability, and Kline (1986) recommends that any item with a reliability in excess of 0.3 can be further considered.

Selection of items for high reliability needs to be tempered, however, by consideration of what Cattell (1972) calls "bloated specifics". Cattell holds that choosing items with too high reliabilities can lead to selecting only a very narrow range of item types, and while this would lead to a scale with high reliability, the scale's items would be too specific to properly measure the trait in which we were interested. For example, if the test concerned were for verbal ability, then selecting only word knowledge items would lead to a scale with high reliability but it would be an imperfect measure of the verbal ability trait. Dealing with "bloated specifics" is an experiential issue and requires the test developer to ensure that an adequate range of test item types are included in the test.

Having trialled a set of items and selected those which meet the prescribed item statistic guidelines, these are now assembled into a test and the test is trialled. This trial results in two outcomes; an estimate of the test reliability, and, evidence of whether further refinement of the test is needed.

Test Reliability. A test's reliability, is the extent to which it reflects an individual's true score on a trait, that is, the correlation of the observed score with the true score. It can be shown that the square root of a test's average correlation with all other tests in the domain of interest is its reliability, (Kline, 1986; Nunnally, 1978). Thus reliability is defined, operationally, in terms of parallel measures. There are two types of reliability coefficients; coefficients of stability which examine parallel measures over time, and coefficients of equivalence which examine the equivalence of two parallel measures. Unfortunately a test's reliability can only ever be estimated as we can never test all of the tests in the universe, but sampling theory shows that such estimates are more than adequate.

Coefficient Alpha. One of the most common means of estimating a test's reliability is through coefficient alpha (Cronbach, 1970). This is an estimate of the correlation of the test with another test of the same length from the universe of items (Kline, 1986) and is calculated as follows:

$$\text{alpha} = \frac{k}{k - 1}\left[1 - \frac{\text{SUM}[\text{var}_i]}{[\text{var}_y]}\right] \qquad (5)$$

Where k is the number of items in the test, SUM[var_i] is the sum of item variances and [var_y] is the variance of the test. The Kuder-Richardson 20 (KR-20) formula is a special case of coefficient alpha for dichotomously scored items.

The Effect of Test Length. With the conceptualization of true score as an individual's score on a test of infinite length, then it is obvious that the longer the test, the more accurate the estimate of the individual's true score. Thus the number of items in the test has a direct influence on a test's reliability and this has been operationalised in the Spearman-Brown Prophesy formula (Guilford, 1956). This formula, used originally to calculate the reliability of a test after a split-half reliability study, relates the number of items in a test directly to the reliability of the test.

Test Refinement. One of the other considerations from the test trial is a refinement of the test. All of the item statistics so far calculated, the item difficulty and item reliability, are sample dependent statistics. That is, they relate to the sample of items and individuals on which they were calculated, a factor that has been identified as one of the major limitations of CTT (Hambleton and Swaminathan, 1985). Thus, an item's statistics may change when the test is trialled on another sample. Given that the item trial sample was relatively representative, and the same applies to the test trial sample, any significant changes in the performance of an item can most likely be attributed to a problem with the item and the item will be normally be discarded.

After any final refinements to the test, if required, the test will usually have to be normed. Cut-off points, or standard scores, are calculated from the distribution of test scores to represent certain percentages of the sample. There are a variety of norms available, the IQ scale has a mean of 100 and a standard deviation of 15, the T scale has a mean of 50 and a standard deviation of 10, etc. As with the original item statistic these norms are sample dependent and therefore a test that is used over some length of time will require periodic re-norming to ensure that the cut-offs chosen still reflect the appropriate percentages of the population.

3.1.3 Limitations of Classical Test Theory

Hambleton and Swaminathan (1985), perhaps two of the harshest critics of CTT, cite five limitations of CTT; these are:

a. that the item statistics fundamental to the model are sample dependent;

b. that comparisons of individuals are limited to situations where the individuals are administered the same or parallel tests;

c. that test reliability, a concept fundamental to CTT, is defined in terms of parallel measures, something that, they claim, is difficult to realise in real life;

d. that CTT provides no indication of how an individual might perform on a particular item; and finally

e. that the model assumes that each individual's error variance is identical.

Lumsden (1976), with his "Flogging Wall Test", also provided an erudite description of what he saw as a major problem with the concept of a test's reliability, namely that a test's reliability is related to the individual's score on the test. This is because the test is of finite length and therefore test scores of those individuals near the extremes of scores on the test must suffer either a floor (at the low score extreme) or ceiling (at the high extreme) effect which reduces the variance of scores at the extremes and therefore reduces the calculated reliabilities of these scores.

Finally, Lord and Novick (1968) cite three points of view with regard to the concept of true scores: that of Thorndike (1964) who feels that true scores are of no **theoretical** interest; that of

Loevinger (1957) who feels that since true scores aren't directly measurable, the observed score is the only meaningful notion and that true scores have no practical interest; and their own, which is basically supportive of the concept. Thus the fundamental premise of CTT has also been found wanting by some people, in some cases.

Some of these criticisms are more problematic than others. In particular, the sample dependent nature of the item statistics, the requirement for at least parallel measurements to be able to meaningfully compare individuals, and Lumsden's (1976) concerns about the concept of test reliability.

3.1.4 Benefits of Classical Test Theory

Despite these criticisms CTT has survived for many years and has had an enormous impact on psychology as a science and has proved to be robust in almost all practical test situations.

From an administrator's point of view scoring a CTT based test is a quick and simple matter. From the test developer's point of view the item statistics required are simple to calculate and don't require inordinantly large samples. Also, being based on simple linear statistical models, the theory behind CTT is easy to understand.

3.2 Item Response Theory

Item Response Theory (IRT) is the other predominant test theory in use today and can be seen as the main alternative test theory to the Classical Test Theory described above.

IRT is a subset of a wider measurement theory called Latent Trait theory. This assumes that an individual's performance on a particular measurement is completely determined by a set of latent, or unobservable, traits. Many psychological theories are based on this concept of latent traits but often without any requirement that the traits actually exist (Lord and Novick, 1968).

Item Response Theory requires two assumptions be made of the data: the first concerns the dimensionality of the latent space, and the second is known as the assumption of local independence.

3.2.1 Dimensionality

The assumption pertaining to the dimensionality of the latent space is fundamental to all IRT models, and in testing terms can be considered as follows: given a set of n test items and k traits, denoted by the vector: $A = (a_1, a_2, \ldots, a_k)$; each examinee can then be described by a point in the k-dimensional space, the latent space, described by A. Next, consider all of the populations of interest, if the (joint) distribution of items scores for examinees with the same value of A is also the same then A is said to "span" the latent space, and the latent space is said to be "complete".

The regression of item score on A, that is the plot of average item score for given values of A, is called the item characteristic function. Because the distribution of item scores is the same across populations for a given value of A, the item characteristic function is also invariant across populations. As a consequence of this, any parameter used to describe the item characteristic function is also invariant across populations.

Thus, if it can be determined that the items in a test can be described by a complete latent space, then the parameters that describe the relation between item scores and the latent traits will be invariant across the populations of interest. For binary items, the item characteristic function specifies exactly how the observed responses relate to the latent traits, and because of this, it is possible to make inferences about the latent traits directly from the item responses (Lord and Novick, 1968).

3.2.2 Local Independence

Whenever more than one test item is being considered, the assumption of local independence is considered necessary for useful theoretical work (Lord and Novick, 1968). In practical terms, local independence mean that the relative position of items within a test has no bearing on examinees' performance on them.

In more theoretical terms, local independence can be defined as meaning that within any group of examinees with the same value for A, the distribution of item scores are independent of each other. This doesn't mean that the items are unrelated to each other, rather it means that the items are only related to each other through the latent variables $a_1, a_2, \ldots, a_k$. This is equivalent to saying that the latent variables span the latent space because, if the item scores were not independent (conditional on A) then this would mean that there were some variable(s) in the latent space other than the k latent variables we have considered.

For binary items, this has considerable importance. In particular, because binary items are scored 0 or 1; then for a pattern of item responses: $V = (U_1, U_2, \ldots U_k)$, $U_n = 1$ or 0; the distribution of V (conditional on A) is:

$$P(V|A) = \prod_{g=1}^{n} P_g^{u_g} Q_g^{1-u_g} \qquad (6)$$

Now from this, if for some population of examinees, A has a distribution g(A), then P(V), the **unconditional** distribution of V, is given by:

$$P(V) = \int g(A) \prod_{g=1}^{n} P_g^{u_g} Q_g^{1-u_g} \, dA \qquad (7)$$

Because we can draw a sample from P(V) (i.e. obtain an empirical estimate of P(V)) we can use equation (7) to make inferences about g(A).

Thus the assumption of local independence allows one to take a sample of item responses and make inferences about the latent (unobservable) traits underlying the examinees' performance.

3.2.3 Item Response Models

There are a variety of models for item response theory that generally differ in one of three ways. These are:

a. the dimensionality of the model,

b. the number of parameters involved, and

c. the form of the function involved.

Model Dimensionality. This was one of the most controversial features of IRT, because early IRT models required that the latent space be unidimensional. This found harsh criticism from Goldstein (1980) and was perhaps largely the result of confusion over the meaning of unidimensionality. McDonald's (1981) definition of unidimensionality in terms of the common factor model and Hambleton and Swaminathan's (1985) qualification of this to mean "one dominant factor" provided perhaps the best compromise between the strict requirements of the IRT model and the practical considerations of the test developer. While multi-dimensional models are available, they provide unique difficulties in their scoring and the interpretation of item responses, and the bulk of the current models use McDonald's (1981) conceptualization of unidimensionality to allow the assumption of unidimensionality to be made.

Model Parameters. The most simple of IRT models, the Rasch model (Rasch, 1960), uses only two parameters in describing the item characteristic function; one representing the individual's standing on the latent trait (their ability), the other represents the amount of the trait required to correctly answer the item (the item difficulty). Other parameters that have been considered in IRT models include; an item discrimination parameter that allows for an item to provide different levels of discrimination at different levels of the latent trait; a parameter to take into account the probability that an examinee with very low levels of the latent trait correctly answers (i.e. guesses) an item, this is usually called the psuedo-guessing parameter; and a parameter that allows for an examinee of very high ability answering an "easy" item incorrectly, though this parameter is less common than the others.

Function Form. There are two main function forms that are used for IRT models; the normal ogive and the logistic models. The details of the normal ogive model were first developed by Lawley (1943) and is given by:

$$P_g(A) = P_g(A, a_g, b_g) = \text{PHI}\,[L_g(A)] = \int_{-L_g(A)}^{\infty} \text{phi}(t)\, dt \qquad (8)$$

Where $L_g(A) = a_g(A - b_g)$ and phi(t) is the normal frequency function. The a_g parameter indicates the amount of information an item provides about A, the discrimination parameter discussed above, and is assumed to be finite and positive. The b_g parameter is related to the level of ability at which the item discriminates most effectively (the difficulty parameter) and for the normal ogive model $P_g(b_g) = 0.5$ (Lord and Novick, 1968).

The logistic test model was introduced by Birnbaum (1968), and is one that very closely approximates the normal ogive. The logistic model is as follows:

$$PSI(x) = e^x/(1 + e^x) = 1/(1 + e^{-x}) \qquad (9)$$

It has been shown (Lord & Novick, 1968) that for all values of x, the normal ogive and the logistic model, when scaled by a value of 1.7, differ by no more than 0.01 in value. Given this we can now write:

$$P_g(A) = PSI[1.7a_g(A - b_g)] = (1 + \exp[-1.7a_g(A - b_g)])^{-1} \qquad (10)$$

This equation (10) is obviously much easier to work with than equation (8) above and therefore the normal ogive model provides a mathematically convenient approximation of the normal ogive and an IRT model in its own right. Both models are unidimensional and of the two the two parameter logistic model is by far the more commonly implemented. Asa result only this model will be discussed further.

3.2.4 The Methodology of Item Response Theory

The methodology of IRT follows an essentially similar format to that of CTT. The steps involved are:

a. develop a pool of items,

b. decide the IRT model to use,

c. trial the items on a sample of examinees,

d. discard those which do not fit the model, and

e. calibrate the refined test.

Apart from selecting which IRT model is to be used the steps involved are basically identical to those for CTT, it is the way items are calibrated and, in particular, the apparent lack of a need to develop norms for the final test that sets IRT apart.

Model Selection. As discussed above there are two types of models commonly used in IRT, the normal ogive and the logistic model. Of the two, the logistic model is by far the simpler to implement and therefore is the most commonly available. The approximation of this to the normal ogive is so close, as shown above, that, for practical purposes, it is the same. The next most important criterion for model selection is how many parameters will be used in the model and this will be dictated by the type of test that is being developed. For example, a multiple choice test can be considered prone to guessing and therefore the test developer may want to take this into account by including a psuedo-guessing parameter, whereas a free-response test is much less prone to this sort of error in measurement. A final consideration in deciding the number of parameters in the model is the sample size available for calibration purposes. The more parameters included the larger the sample size required to adequately estimate the parameters (see for example, Hambleton and Cook, 1983; Lord, 1983).

Item Trials. This is essentially the same process as for a CTT test development procedure. The pool of items is administered to a group of examinees to obtain estimates of the fit of the model to the data. The requirements for this sample are similar to that for the CTT based test, a representative sample will ensure that the model parameter estimates will be as accurate as possible at the desired ability level. If the test is trialled on a sample that is very different from that to which the final test will be administered, ability estimates for the final group can be made (identical to the original estimates up to a linear transformation) but they will not be as accurate as those made for the trial group. For example, if the test is trialled on a sample of examinees with very high levels of a trait then the ability estimates for individuals with high trait levels will be very accurate compared to those for individuals with relatively low levels of the trait.

Parameter Estimation. Parameter estimation is normally made using some form of maximum likelihood estimation method. The two most common are marginal maximum likelihood (MML) originated by Bock and Lieberman (1970) and operationalised in the BILOG computer software package (Mislevy and Bock, 1990); and the earlier Joint Maximum Likelihood approach suggested by Birnbaum (1968) and operationalised in the computer program LOGIST (Wingersky, Barton, and Lord, 1982). Of these two, the BILOG implementation has been shown to be slightly more robust to violations of the assumptions IRT (Ackerman, 1987) and also more effective with a wider range of sample sizes, in terms of items and subjects (Mislevy and Stocking, 1989).

Test Refinement. As with the CTT based test, it is to be expected that not all of the items chosen in the initial pool will adequately fit the model. There will therefore be a requirement to discard some items from this pool as was done in the development of a CTT based test. One of the features of the MML procedure as implemented in BILOG is that it allows calculation of goodness-of-fit indices for the individual items in the test.

Estimating an Examinee's Ability. Having refined the test it is now ready for use and here one of the major differences between the two test development methods appears. This is that there is allegedly no need to administer the test to another sample to develop norms for the test. This is because once the item parameters have been determined, and they are said to be invariant across samples, an examinee's ability can be calculated, in standard score form, directly from the ICC. This estimate will also be invariant across samples of items taken from the originally calibrated item pool (Hambleton and Swaminathan, 1985).

3.2.5 Limitations of Item Response Theory

IRT is mathematically elegant and, providing the appropriate assumptions can be met, allows the test developer to make very strong statements about an individual's ability level based purely on their responses to a set of test items. This is not to say that IRT is without its critics. The earliest criticisms of IRT centred on the assumptions it required of the data and the fact that few items seemed to fit the chosen models.

The Assumptions of IRT. The early criticisms of IRT concerned the assumption of unidimensionality that was a requirement of the models (Goldstein, 1980). As stated above, McDonald's (1981) description of the unidimensionality issue in terms of the common factor model clarified many of these criticisms.

Item Fit to IRT Models. Another common criticism of IRT was that very few items seemed to fit the models used, and this was particularly true of the Rasch model, the prime concern being that legitimate items would be discarded because of poor fit to the model. A close examination of the Rasch model shows that not only is it a special case of the two parameter logistic model but that it is also a very strong measurement model in its own right (see, for example, Andrich, 1988). As such it makes quite stringent requirements of the data, but in return allows the user to make very strong statements about the data. Certainly other IRT models, for example the two parameter logistic model, are not so strict on the requirements of the data for the model and consequently show much higher proportions of items fitting.

In practical terms IRT presents a few other problems, most prominent of these is scoring the test. As stated above, once the item parameters of the test have been determined, scoring the test for a new examinee is simply a case of inputting the examinee's response pattern into the ICC. Unfortunately this is not a simple process and normally requires the application of sophisticated computer software. This may not be a problem for a computer administered test but the reality is that, in most situations, tests are still manually administered and scored.

The rationale for scoring a test is to simplify the data pertaining to an examinee's responses to an item set. The range of an examinee's responses to a set of v items is v-dimensional, while any scoring formula, say t = t(v), used on the responses yields a one-dimensional range of values. So scoring a test provides a real simplification of the available data. The problem now is to simplify the data without losing any information. Fortunately, statistical theory provides a class of statistics called sufficient statistics that serve this exact purpose.

Sufficient Statistics. A sufficient statistic is one which summarizes all of the information in a sample concerning a target parameter. Formally a statistic is a sufficient statistic for a parameter if the conditional distribution of the sample values (given the statistic) does not depend on the target parameter (Hogg and Craig, 1978). For example, the sample mean is a sufficient statistic for the population mean of a normally distributed population. In the case of the psychological test we are looking for a statistic that can be used to provide a more efficient estimate of the examinee's ability. Birnbaum (1968) has shown that for tests with equivalent items the number correct is a sufficient statistic for the examinee's ability, and that for the two parameter logistic test model, the sum of item scores weighted by their discrimination parameter, a_g, is a sufficient statistic.

Indeed one of the attractions of the Rasch model, was that some function of an examinee's number right score could be used as an estimate of their ability.

For a test that is to be manually administered and marked this offers some hope, but it still requires work. Unless the Rasch model is going to be used, with its concomitant rigourous demands on the data, the test developer still has to deal with a **weighted** sum of item scores which still have to be entered into a complex equation. Really, the test administrator wants to be able to simply add up the number right on a test and use this as the examinee's score. Is it possible to do this and still retain the benefits of using an IRT test development method? Two issues need to be addressed for this to happen; the need for a weighted sum and the form of the scoring function.

Weighted Composites. In terms of the weighted composite, there is a considerable amount of literature available on the comparison between different weighting schemes (Wilks, 1938) and most show that unit weights are as good or better than differential weights for prediction purposes (Wainer, 1976; Dawes and Corrigan, 1974; Einhorn and Hogarth, 1975). Indeed Dinero and Haertel (1977) explored this question in a testing setting using simulated data. They found that items with varying discriminations estimated using the Rasch model gave little loss of information. So it seems that it may be possible to simply use the unweighted test score as an estimator of ability even if we select items based on the two parameter logistic model.

Form of the Function. While the form of the function for the estimation of the examinee's ability in the two parameter logistic model is fairly complex (see Lord and Novick, 1968, p 429), it is only a non-decreasing function in the test score. Given this, the unweighted number right score should provide good correlation with the ability estimate and therefore should be an adequate substitute for the ability estimate.

3.3 Comparing the Models

From a practical point of view the CTT based test is easy to develop and administer. The IRT based test is more complex but allows the test developer to make very strong statements about an examinee's ability from their responses to a set of test items. Which is better? Unfortunately the literature yields only one (somewhat dated) study where the relative merits of CTT and IRT were directly compared (Douglass, Khavari, & Farber, 1979).

Douglass et al (1979) found that, although the two test development procedures produced different tests, in terms of their item composition, there was little difference in the correlations of the tests with an external criterion. Although this work was done with a clinical instrument rather than an ability test and used the Rasch model as the IRT test model for test development, the results are of relevence here. Certainly there is no reason to assume that the results of Douglass et al (1979) should not be repeated here, that is, tests developed using CTT and IRT may well have different item compositions but should be equally useful in predicting an external criterion.

Even if this is the case the sample-free nature of the IRT estimates are attractive to the test developer. An ideal situation would be to somehow combine the ease of administration of the CTT test with the power of the IRT test. The discussion above has lead to the conclusion that a test may be developed using an IRT model, in this case the two parameter logistic model, but we might obtain adequate ability estimates using a simple number right statistic.

We therefore have three models to consider; firstly, the CTT model, secondly, the "standard" two parameter logistic model, and thirdly, items modelled on the two parameter logistic model but with ability estimated via the simple number right score. Our primary concerns are that each of the models should be able to produce equivalent estimates of examinee ability.

As always the difficulty with these questions is what criteria are to be used to ascertain the adequacy, or otherwise, of the ability estimates? In the current situation it is fortunate that a ready criterion presents itself, namely the ability estimates provided by the existing Army selection test, the AGC. This test is to be replaced by the new test, not because of any perceived inadequacy in the performance of the test, but rather because it lacks face validity due to its age and the opportunity presents itself to replace it. As such it is critical that the new test replicate the performance of the AGC as closely as possible.

The Hypotheses. Our first hypothesis is that a test developed using CTT, although different in item composition, will not produce appreciably better estimates of examinee ability, defined in terms of performance on the existing selection test than a test developed using an IRT model. The second hypothesis is that a test developed using IRT will produce the same ability estimates, as defined above, whether the full ability estimation procedure is used or a simple number right test score is used.

3.4 Chapter Summary

There are two test development theories predominant; Classical Test Theory and Item Response Theory. CTT assumes that for any trait on which we attempt to measure an individual there will be a true score for that person which is a constant. The measure we take can only ever approximate this true score and the central problem in CTT is to build a test that will make this estimate as accurate as possible. We try to minimise the error of measurement. CTT has been criticised on a number of grounds, in particular the estimates it makes are sample dependent.

Item response theory, states that an examinee's performance on a test item can be completely determined by their standing on the traits underlying that item. If a set of dichotomously scored test items occupies a complete latent space then inferences about an examinee's standing on the traits underlying those items can be made directly from the examinee's responses to those items. To achieve this in practical terms, IRT makes strong assumptions about the dimensionality of the data and the conditional independence of the items. If these assumptions can be met, however, the IRT parameter estimates obtained are (theoretically) sample-free, up to a linear transformation of the estimates.

Although CTT is the traditional test development model and is simple to implement, IRT offers much for the test developer. It is not without its problems though, in particular it requires complex scoring formulae and therefore is more difficult to implement in a pencil-and-paper test form. Investigation of the scoring formulae indicate, however, that simple number right score may provide an adequate approximation of the ability estimates provided by the full model.

It was therefore decided to test two hyptoheses; whether IRT produced a better measure than a CTT based test and whether simple number right score could provide an adequate replacement for the less practical ability estimates from the IRT model.

CHAPTER 4

METHOD

There were two phases to the project; the development of tests using the different strategies (i.e. the classical model and the two parameter logistic model), and the comparison of the scores obtained from the different strategies (including the modified two parameter item response model) with the existing instrument.

4.1 The Instrument

A set of 52 items provided the calibration test (called the T0) for the development of the final test (the T1) from which the concurrent validities would be calculated. The T0 included 18 verbal items (code and word knowledge items), 10 spatial (paper form board, rotations and unfolding items), 10 numeric items (arithmetic and clock items), and 14 reasoning (analogies and series items). Of the reasoning items, there were five with verbal content, five spatial and four with numeric content. The items were organised in a spiral omnibus format without regard to any ordering for difficulty. The T1 included items selected after the first part of the analyses and was also designed as a spiral omnibus format test. This was also administered to a sample of Army examinees.

4.2 Analyses

Four analyses were conducted:

a. a factor analysis of the item set was conducted to test the dimensionality of the item pool, a reliability analysis of the scales was also conducted as a confirmatory procedure for the factor analysis;

b. a CTT analysis of the items was conducted;

c. an IRT analysis of the items was conducted; and

d. a comparison was made of the relative effectiveness of each of the different test development strategies using regression analysis.

Factor Analysis. As discussed in Chapter 2, a need was seen to ascertain the dimensionality of the item pool. In particular a three and four factor representation of the item space were to be compared. The problems of factor analysing test (binary) data is well documented, the basic problem being that neither the tetrachoric correlation nor the phi coefficient are considered a suitable base for factor analysis. A technique developed by Christoffersson (1975) uses the distribution of joint probabilities and the generalised least squares principle to conduct a multiple-factor analysis of dichotomised data. This approach was shown to be equivalent to the "harmonic-least-squares" approach implemented by Fraser (1988) in his computer program NOHARM and this is the program that was used here. NOHARM allows a confirmatory as well as an exploratory analysis to be done and the former approach will be used here. The fit of each model to the data will be compared by examining the root mean square of the residual (RMSR) inter-item correlations and by examining individual item communalities. The results of this analysis will provide the factor structure of the T1.

CTT Analyses. CTT uses three main statistical tools for identifying "good" items; a difficulty index, item-total correlations and a measure of the scale's reliability to test the usefulness of the scale. The most common difficulty index is simply the proportion of examinees correctly answering an item, this is often called the "p-value". The item-total correlation simply shows how well an individual item contributes to the scale score to which it belongs. The reliability of a scale can be seen as the internal consistency of the set of items that form the scale. Item difficulty values of 0.2 and 0.8 (Kline, 1986) were used as boundaries, outside of which items would be discarded as being of little use, a lower bound was set on the item-total correlation of 0.2 for suitability for further consideration.

IRT Analysis. The IRT analysis was conducted using the micro-computer based program BILOG 3 (Mislevy and Bock, 1990). This software provides a range of options for the analysis of items and tests and the two parameter logistic model was chosen as the model for the analysis. BILOG uses the MML estimation procedure mentioned in the previous chapter. MML assumes the independence of item responses conditional on the examinees' ability level, that is for examinees with the same level of the trait under investigation (see Equation 6 from Chapter 3).

Item Selection. BILOG provides a range of goodness-of-fit statistics for individual items dependent on the number of items in the scale being examined. These include; testing the goodness-of-fit of the model directly for very short tests (10 or fewer items), where all or nearly all of the 2^n item response patterns appear in the data, using a likelihood ratio chi-square statistic; for tests of 11 to 20 items, Mislevy and Bock (1990) claim that no reliable test exists for testing the overall fit of the model but standardised posterior residuals can be calculated for individual items for testing the model; and, for sufficiently long tests, more than 20 items, a likelihood ratio chi-square statistic can be calculated from the estimated ability levels of the examinees based on the model and their ability level as estimated from the model.

Regression Analyses. The utility of the three test development models will be compared, as discussed, by comparing their concurrent validities with the current Army selection test, the AGC. Number right scores for the CTT test and the modified two parameter model and scale scores for the two parameter logistic model will be calculated (for each of the three scales) and then raw scores on the AGC will be regressed onto these. The resulting R^2 values will be compared for each model to indicate which test development model provides the best fit to the criterion.

4.3 Sample

T0 Sample. The T0 was administered to 209 male Army GE examinees at two test sites. The group ranged in age from 16 to 32 years and had a median age of 18 years. Through an administrative error school level was only recorded at one of the test sites (104 cases); 56.7% had completed Year 10 or below, 19.2% had completed Year 11 and 24% had completed Year 12. Scores on the current Army

GE selection tests (the AGC, a 100 item spiral omnibus format test of general ability) ranged from 14 to 87 with a mean score of 50.923 and standard deviation of 13.86. This compares reasonably well with the figures from all examinees for 1989; range 0-98, mean 50.54, standard deviation 14.85.

T1 Sample. The T1 was administered to a sample of 371 Army examinees. There were 357 males and 14 females in the sample (the number of females was small enough that it was considered unlikely that any sex differences would effect the results). The mean age for this sample was 18.9 years, and of these 66% had completed Year 12 at school, 10% Year 11 and 24% Year 10 or less at school. Scores on the AGC ranged from 25 to 95 with a mean of 61.13 and a standard deviation of 12.42. These figures compare less well with the general figures than those for the T0 sample.

CHAPTER 5

RESULTS

5.1 Sample

There were considerable differences in the composition of the samples used for the initial item selection, the TPAB-T0 sample, and that used for the final test calibration, the TPAB-T1 sample. This was caused by the different entry types of the two groups.

Due to policy changes between when the T0 was administered and when the T1 was administered, GE examinees were not being processed for entry to the Army when the T1 was administered. The only avenue of entry open, and therefore the only source of subjects available, was a special form of Reserve entry, the Ready Reserve. Because of the requirements of this form of service, high school graduates were specifically targeted for the Ready Reserve. This resulted in a much higher proportion of Year 12 graduates (66% in the T1 sample versus only 24% for the T0 sample) which probably also contributed to the significantly higher mean AGC score for the T1 sample ($t = 9.117$, $p < 0.005$).

Although a confounding variable, this difference in the sample composition means that one of the main claims of IRT will be able to be considered, namely that parameter estimates are invariant across groups while those of CTT are not.

5.2 First Phase of the Research

5.2.1 The Factor Structure of the Item Set

Two analyses were conducted to investigate the factor structure of the item set. First, a confirmatory factor analysis using the NOHARM computer program (Fraser, 1988) was conducted to determine whether a three or a four factor model was required to adequately describe the item set. Then a reliability analysis was conducted as a confirmatory procedure for the factor analysis.

Two factor structures were investigated:

a. a three factor model where items were allocated to factors on the basis of their content, ie verbal, numerical and spatial; and

b. a four factor model which included a reasoning factor along with the three above.

For completeness a single factor structure was also included in the factor analysis.

The main means provided by the NOHARM factor analysis package for testing the fit of the model is by the analysis of the residual correlation matrix through the root mean square of the residuals (RMSR). Fraser (1988) states that an RMSR "...in the order of the typical standard error of the residuals (4 times the reciprocal of the square root of the sample size)..." (p2) indicates that the hypothesised model should not be rejected. Table 5-1 below shows the RMSR for the models plus Fraser's suggested value.

Table 5-1: RMSR Values

Model	Value
Three Factor Model	.01069
Four Factor Model	.01063
Single Factor Model	.01093
Suggested Value	.27669

As can be seen from this, all of the models yield RMSR values well below Fraser's (1988) suggested value and there is little difference between the values.

As there is little **overall** difference between the models, the next step was to examine how **individual items** fared under each model. This was done by examining the unique variances for each item under the different models (these statistics are at Annex A) and discarding items that fit poorly to the hypothesised structures. Items were discarded if they yielded high unique variances (>0.8) in two of the three factor solutions. A total of 26 items were dropped from further consideration under this criterion.

Of the remaining 26 items, the items fit the three factor model best in 20 cases and the four factor model in eight cases (two items had identical communalities under both models). In no case did any item fit the single factor model best. The conclusion to be drawn from this analysis is that; first, separate abilities are required to best explain the data at hand, and secondly, that a reasoning, or fluid ability, scale is not required to adequately fit the data, over and above the fit provided by the scales related to the other secondary abilities, namely verbal, numeric and spatial ability.

To augment the factor analysis, a reliability analysis of the three and four scale models was conducted. The scale reliabilities for the models are in Table 5-2 below. This also includes the Spearman-Brown (SB) prophesy formula (Guilford, 1956) value for the three scales common to both models. This calculation is based on increasing the original scales by the number of appropriate items from the reasoning scale.

Table 5-2: Reliabilities for Three and Four Factor Models

Factor	Reliability		
	Four	Three	SB Value
Verbal	0.608	0.674	0.664
Spatial	0.520	0.610	0.619
Numerical	0.656	0.726	0.732
Reasoning	0.641	-	-

The Spearman-Brown prophesy formula measures how much the reliability of a test would improve if it were lengthened by a specific amount. This is based on lengthening an existing test with items from the **same scale** and was originally developed for estimating the reliability of a complete test from its split-half reliability.

In this case, the scales in the three factor model were lengthened by including the items of appropriate content from the reasoning scale. The results in Table 2 show that in all cases the scales from the three factor model show an increase in reliability in the same order as that specified by the Spearman-Brown formula.

From this we can conclude that the reasoning items contribute to the internal consistency of the other three scales, a conclusion that was confirmed by examining the individual item statistics. These showed that the majority of the reasoning items contributed positively to the reliabilities for the three-factor scales. This supports the conclusion drawn from the factor analysis that a scale devoted specifically to reasoning, or fluid ability, is not necessary to adequately describe the item set.

5.2.2 CTT Analysis

The selection of items using the CTT procedure followed the "standard" CTT parameter cutoffs of p-value greater than 0.2 and less than 0.8 and item-total correlations greater than 0.2 (Kline, 1989). A total of 31 items "passed" these criteria, 15 verbal, 9 numeric and 7 spatial items. From this set a twenty item test comprising seven verbal items, six spatial and seven numeric items was constructed. Complete CTT item statistics are at Annex B.

5.2.3 IRT Analysis

IRT parameters were then estimated for the complete item set within their individual scales. This was a two stage process where poorly fitting items were discarded after the first calibration and the analysis was conducted again on the reduced item set (item calibration statistics are at Annex C). From this second calibration a twenty item test was constructed containing seven verbal, six spatial and seven numeric items.

5.2.4 Supplementary Analyses

Two supplementary analyses were conducted. First, a Rasch analysis of the items fitting the two parameter logistic model was conducted to examine the differences in fit of the items to the two models. Comparative statistics for the two calibrations are at Annex D, and these generally show that many of the items chosen as fitting the two parameter logistic model would not have been chosen under the Rasch model.

Secondly, for comparison purposes simple number correct scores for the scales in the second IRT calibration were correlated with the two parameter ability estimates from these "scales". The resulting values are in Table 5-3, below.

Table 5-3: Correlations between Number Correct and Ability Estimate - T0

Scale	Correlation
Numeric	0.967
Verbal	0.950
Spatial	0.939

Interestingly, these correlations all show very strong relationships between the number correct score and the ability estimate provided by the two parameter logistic model. This bodes well for the comparison of the two parameter logistic model with the normal and modified scoring procedures.

<u>Common Items.</u> The final test, the T1, contained 27 items, nine each of the three scales and there were 13 items common to both the CTT and IRT sets (five of the verbal and four each of the numeric and spatial items).

5.3 Second phase of the Research

5.3.1 Regression Analysis

Having developed the instruments, AGC scores were regressed against the three scales to compare the three models. The hypotheses in which we were interested were; whether the CTT and IRT tests gave comparable concurrent validities despite having different item compositions; and whether, for the IRT test, the number correct score provided reasonable estimates of examinees' abilities. Table 5-4 below contains the R-square values for the three different tests.

Table 5-4: R-square Values for Regression Analyses

Test	R-square
CTT	0.428
IRT (Ability Estimate)	0.417
IRT (Number Right)	0.399

As can be seen from the results in Table 4, the IRT based test predicted almost as much variance in the dependent variable as that from the CTT based test (41.7% as opposed to 42.8%) and that simple number correct scores based on the IRT based test was not far below the amount of variance predicted by the other two models (39.9%). Also, with one exception, all three scales entered the regression equations for all three models (see Table 5-5). The only non-significant scale, using the "traditional" alpha level of 0.5, is the spatial subtest for the IRT test when the simple number right score is used as the ability estimate. In all cases the scales entered the equations in the same order, namely numeric, verbal and spatial.

Table 5-5: Regression Equations for the Models

Model: CTT test;

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
Intercept	1	32.067777	2.10474370	15.236	0.0001
Numeric	1	3.481933	0.35802845	9.725	0.0001
Verbal	1	1.363631	0.40731851	3.348	0.0009
Spatial	1	1.363060	0.39694888	3.434	0.0007

Model: IRT test (ability estimates)

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
Intercept	1	61.134771	0.49420314	123.704	0.0001
Numeric	1	5.251413	0.60731244	8.647	0.0001
Verbal	1	3.063147	0.58115971	5.271	0.0001
Spatial	1	1.329721	0.55943467	2.377	0.0180

Model: IRT test (number correct score)

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
Intercept	1	23.431491	2.80503610	8.353	0.0001
Numeric	1	3.993169	0.46499314	8.588	0.0001
Verbal	1	1.899013	0.45032997	4.217	0.0001
Spatial	1	1.079507	0.55013166	1.962	0.0505

Generally all three test development models performed equally well with the CTT predicting slightly more variance in the criterion than the IRT ability estimates, which were, in turn, a little better than number right scores based on the IRT test.

The comparison of IRT ability estimates and number correct scores also showed remarkable similarity in performance. Correlating the ability estimates made for each scale yielded the results in Table 5-6.

Table 5-6: Correlations between Number Correct and Ability Estimate - T1

Scale	Correlation
Numeric	0.970
Verbal	0.976
Spatial	0.961

These values are even higher than those for the T0, and all were significantly different from zero.

5.3.2 Supplementary Analyses

As mentioned above, there were significant differences in the characteristics of the two samples which allowed the opportunity to compare the stability of the item parameter estimates for each of the two test development models. The CTT parameter estimates and the IRT final (i.e. rescaled) parameter estimates are in Annex E.

Stability of CTT Item Parameter Estimates. In all but one case the item facility estimates from the T1 sample were higher than those for the T0 sample. The two sets of parameters also showed a reasonably linear relationship with each other, see Figure 5-1 below.

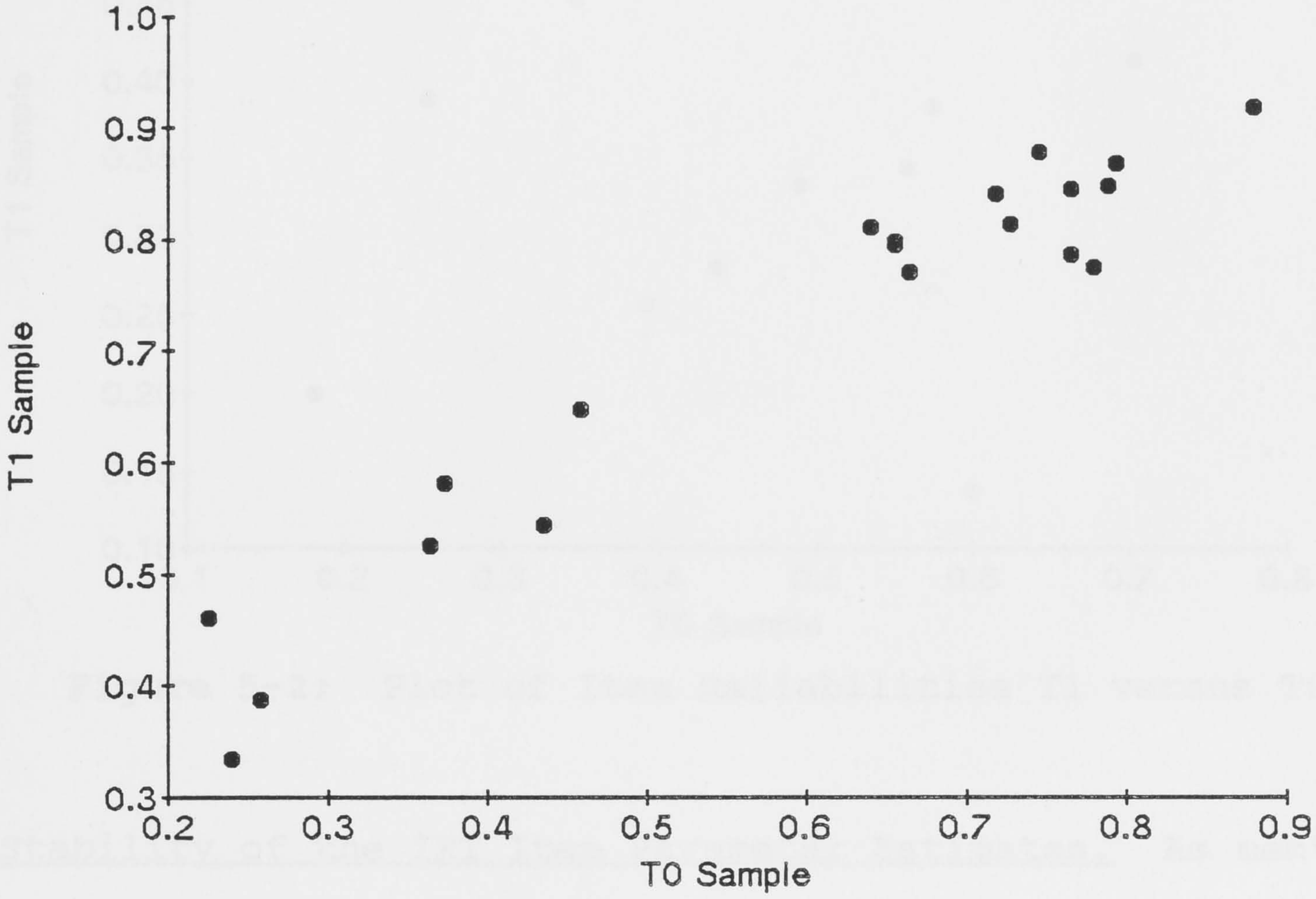

Figure 5-1: Plot of Item p-values T0 versus T1

This is as expected, given that the T1 sample scored significantly higher on the AGC and the T1 correlates well with the AGC. Unfortunately, the behaviour of the item reliabilities is less easy to predict, with the majority being lower in the T1 sample, with the exception of the Spatial scale where all but one were greater in the later sample. The plot of the two sets of estimates is at

Figure 5-2 below and shows no clear relationship between the two sets of estimates.

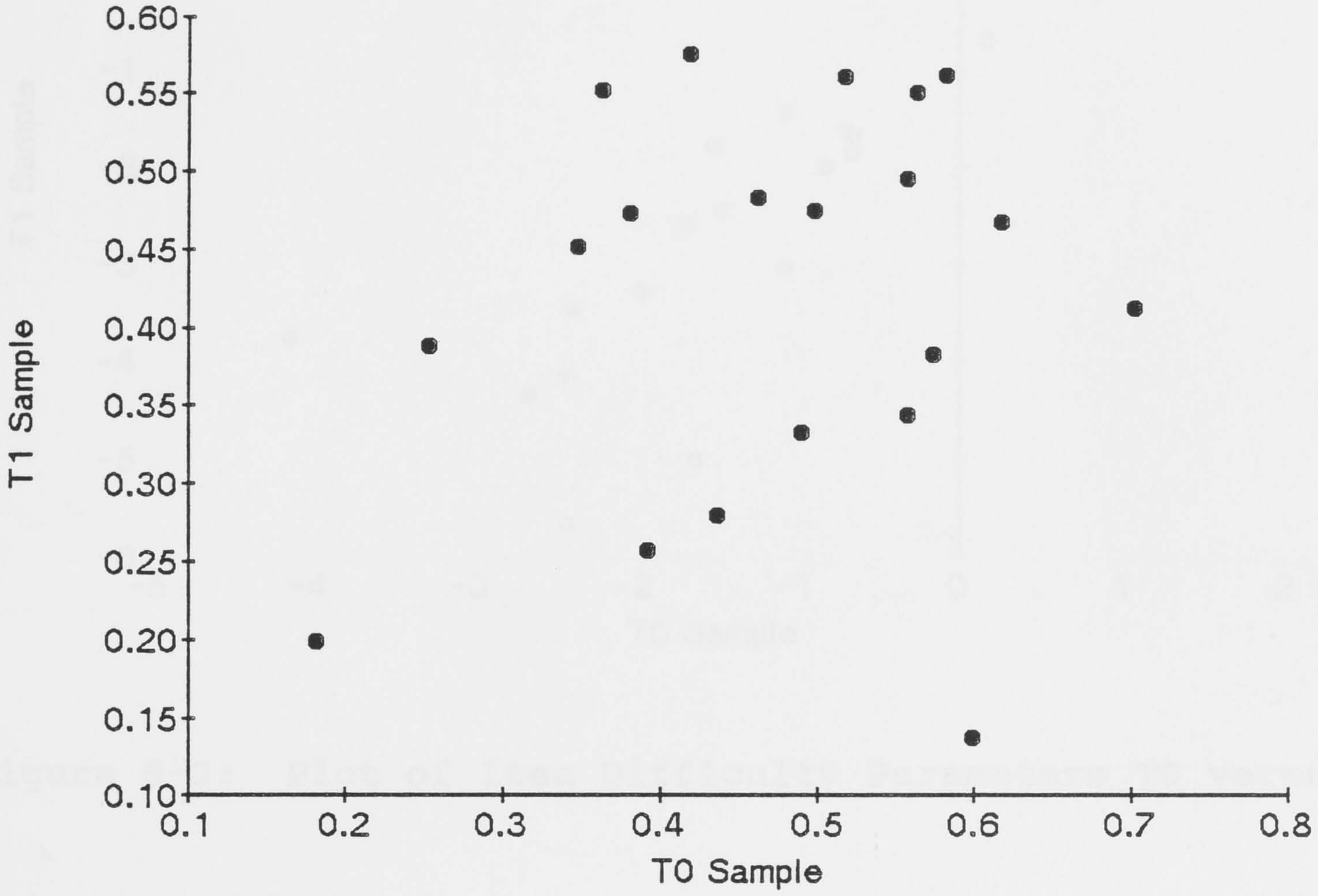

Figure 5-2: Plot of Item Reliabilities T1 versus T0

Stability of the IRT Item Parameter Estimates. As mentioned, one of the theoretical strengths of IRT is that the item parameter estimates are sample free. However, as can be seen in Annex E differences in difficulty parameter estimates across the two samples ranged from 0.21 to 3.391. There would also appear to be a linear relationship between the two sets of estimates as shown in Figure 5-3 below.

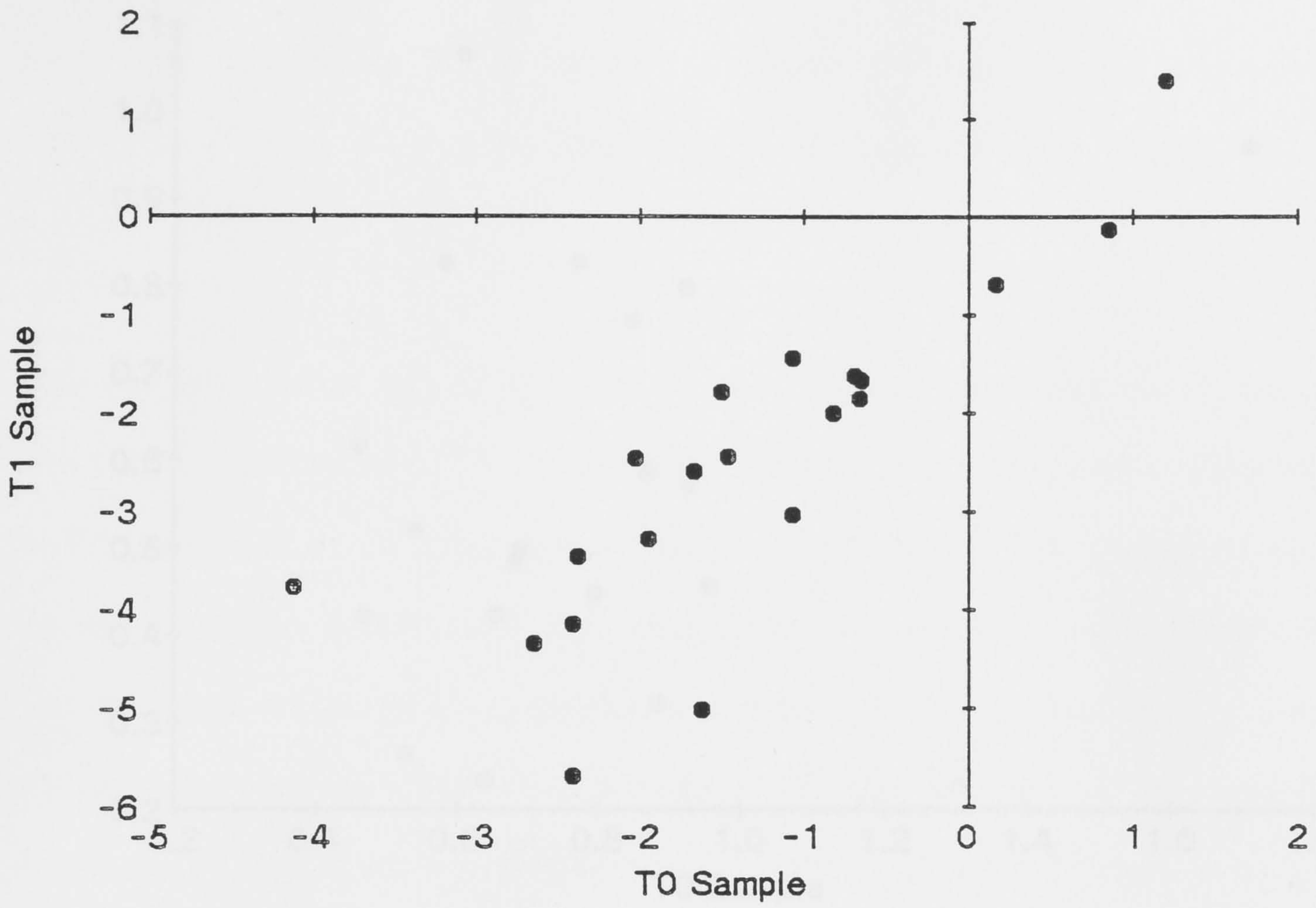

Figure 5-3: Plot of Item Difficulty Parameters T0 versus T1

Differences in the discrimination parameter estimates were smaller and ranged from 0.045 to 0.749, this isn't surprising however as these parameters generally have a narrower range than the difficulty parameter. In this case however, there appears to be no relationship between the two sets of parameter estimates (see Figure 5-4).

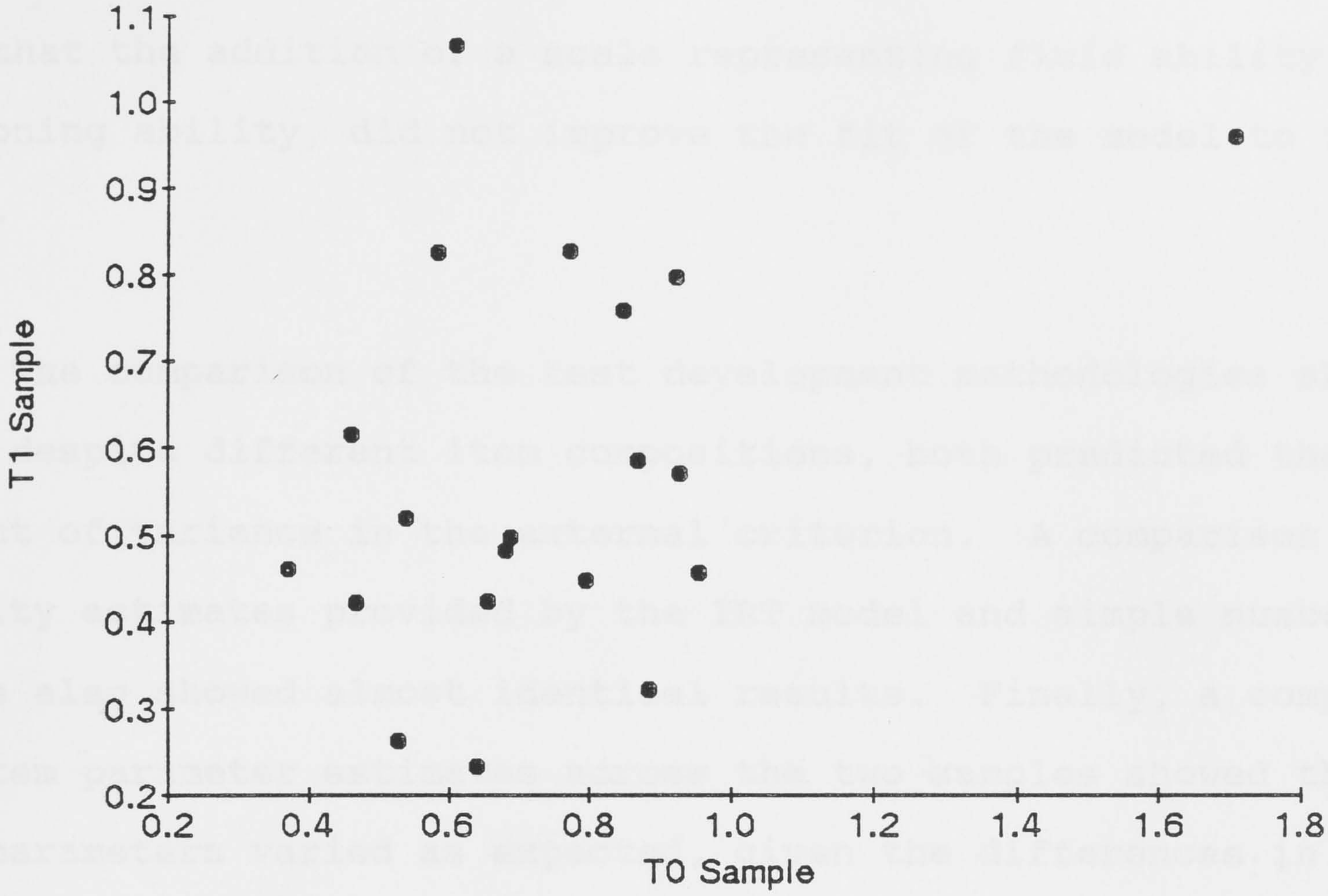

Figure 5-4: Plot of Item Discrimination Parameters T0 versus T1

Despite the apparent differences in the parameter estimates, 95% confidence intervals based on the standard errors of these estimates gave only four difficulty parameter estimates of all of the estimates which were different across samples.

5.4 Chapter Summary

Two sets of analyses were presented; an analysis of the structure of the item set, and a comparison of the CTT and IRT test development methodologies.

The analysis of the structure of the item set showed that a three factor structure consisting of scales representing verbal, numeric and spatial ability adequately accounted for the data set and that the addition of a scale representing fluid ability, called reasoning ability, did not improve the fit of the model to the data.

The comparison of the test development methodologies showed that despite different item compositions, both predicted the same amount of variance in the external criterion. A comparison of the ability estimates provided by the IRT model and simple number right score also showed almost identical results. Finally, a comparison of item parameter estimates across the two samples showed that the CTT parameters varied as expected, given the differences in the samples, and that the IRT parameter estimates varied in a similar fashion but that these variations were within 95% confidence intervals based on the standard errors of the estimates.

CHAPTER 6

DISCUSSION

This study was a predominantly practical one; there was a requirement to produce a selection test for the Army. This also provided the opportunity to examine some methodological issues in test development. In particular two fundamental aspects of test development; the model of ability underlying the test and the test construction theory used, have been examined.

6.1 Study Findings

6.1.1 The Model of Intelligence

The aim of this analysis was to determine if it was necessary to include fluid ability as a separate scale to provide an adequate measure of general ability in addition to the verbal, numeric, and spatial ability scales required for the test development project. The place of fluid ability in the group of secondary abilities has been well documented, though some more recent work (Undheim and Gustafsson, 1987) has provided a new view on this. Rather, fluid intelligence has been seen as possibly a manifestation of general ability and not a secondary ability at all. It was hypothesised, therefore, that a factor structure consisting of three factors; verbal, numeric, and spatial ability, should provide as good an account of the data as the same structure with a fourth scale, fluid ability.

In a departure from the usual research in this area, the analyses were conducted at the item level rather than the test level. That is, in most previous research it has been usual for batteries of tests to be analysed, for example Gustafsson (1984) analysed a battery of thirteen ability tests and three standard achievement tests, and Undheim and Gustafsson (1987) analysed a battery of 26 tests. In the current study it was the items comprising the tests that were analysed.

The results showed quite strong support for the hypothesised three factor structure for the item compared with the four factor structure. It should be noted, however, that the differences evident in the factor analysis were at the item level, that is in the item communalities or the amount of information the factor structure provided about each item, rather than in the overall fit of the two models. The reliability analysis of the scales also showed strong support for the three factor model finding that the items in the fluid ability scale were equally at home in scales related to their content area, that is numeric, verbal or spatial.

That items traditionally seen as relating to fluid ability were shown to be equally useful as items in the other scales support the contention that, rather than being a separate secondary ability, fluid ability is related to all of the secondaries. The ability of the fluid items to contribute to both a separate scale, as well as the other scales indicates that the items contain variance unique to both. Therefore the results previously mentioned (Gustafsson, 1984; Undheim & Gustafsson, 1987) are not surprising as it would be expected that the scales would all correlate with the fluid ability scale.

6.1.2 Test Theory

Two a priori hypotheses were examined in the area of the test theory used to develop the test: First, whether CTT and IRT would produce similar estimates of examinee ability, and secondly, whether the simple number right score, on an IRT based test, would produce the same estimate of examinee ability as the estimates yielded by the complete IRT calibration procedure. Due to substantive differences in the samples used for the first and second test administrations it was also possible to examine one of the (claimed) major advantages of IRT, namely the sample-free nature of the item parameter estimates.

The Comparability of CTT and IRT Ability Estimates. The results quite clearly supported the hypothesis that, despite differences in the item composition of the two tests, both the CTT based test and the IRT test would predict the same amount of variance in the criterion measure, scores on the current selection test (see Table 5-4). This is a clear replication of the results of Douglass et al (1979) when they compared the Rasch model and CTT. Thus, in terms of the external validity of the two tests, as measured in this study, there was no difference in the measures provided by the two test methodologies.

Similarity of Ability Estimates. The study also strongly supported the second hypothesis that number right scores for the

IRT based test would be very similar to the ability estimates made from the full calibration. The correlations between the estimates for all three scales were all extremely high (see Table 5-6) and both scoring strategies yielded similar concurrent validities with the criterion measure (see Table 5-4). These results provide clear support for the notion that the more complex IRT ability estimation procedure can be adequately approximated by number right score. This is based on the idea that unit weights can be used in the linear combination of item scores, the differences in item discrimination parameters can be ignored, and that because the formula for the ability estimate is a non-decreasing function, simple number right score will provide an adequate approximation of the ability estimate.

<u>Stability of the Item Parameter Estimates.</u> This analysis provided mixed results. First, as expected, the CTT item parameter estimates showed considerable variation over the two test administrations. These changes were in the direction that would be predicted, given that the T1 sample was "smarter" on the external criterion. That is, the p-values were greater in the second sample and, generally, the item reliabilities were reduced as there was less variation in the responses of the second group (because the items were generally easier for this group there were fewer different responses in the second sample). The changes in the p-values were considerably more linear, and therefore more predictable, than the changes in the item reliabilities. The parameter estimates provided by the IRT calibration also demonstrated considerable differences. But these were within confidence intervals based on the standard errors of the parameter

estimates. The differences did, however, follow very similar patterns to those seen in the CTT parameter estimates. The changes in the b-values were roughly linear and generally these estimates were larger for the T1 sample. The a-value estimates, on the contrary, showed almost no relationship.

6.2 Implications of the Findings

6.2.1 Models of Intelligence

The fact that the three scales provide as adequate a fit of the data as the four factor model indicates that, in practical terms, the fluid ability scale contributes little information independent of the other secondaries in the model. Thus, to build as efficient a test as possible for the present purposes, three scales should be included; numeric, verbal and spatial ability.

In current terms of the structure of intelligence the results have provided some support for the theory that fluid ability should not be considered a separate secondary ability. Rather it is related to **all secondary abilities** because it is, in fact, a manifestation of general ability to which all secondary abilities are subordinate.

The results here showed that items traditionally considered to be in the domain of fluid ability are equally related to other secondary abilities based on their content, that is whether they are of verbal, spatial or numeric content. Any scale constructed containing these items could reasonably be expected to correlate with fluid intelligence. While examining this at the test level may yield these correlations between the secondaries, including fluid ability, examining it at the item level may provide better insight. The overlap in items between fluid ability and the other secondary abilities examined here may provide an indication of how fluid ability is a manifestation of g rather than being a separate secondary ability.

As items which measure fluid intelligence contribute to both a fluid ability scale as well as a content related scale, perhaps we can conclude that performance on a test item can be conceptualised as requiring a content component and an operation component. Thus performance on a test item would involve varying levels of both components. There could also be some interaction between the operation component and the content component. For example, performance on an arithmetic item would require operations that are highly content-specific, whereas performance on an analogy problem would require operations that were more content-independent, because the operations required are similar across content types.

Thus individuals may have different levels of the content component, that is some would have a greater facility with, say, arithmetic, which would be tempered by different levels of the operation component, that is some would have better general test taking skills than others. Individuals with high levels of the operation component would be generally good on intelligence type tests, they would have high general ability. Similar models of test items, in terms of components which differentially effect the difficulty of the item, have been formulated under the name "Component Latent Trait Models" (see, for example, Embretson, 1984; Embretson & Wetzel, 1987).

This conception of performance on intelligence test items is obviously consistent with Spearman's two factor model, that is all tests have a common and specific component. It is also compatible with the Cattellian models of primary and secondary abilities with the difference that fluid ability is made up of content-independent items drawn from the range of secondary abilities and therefore the secondary abilities are subordinate to fluid ability. Secondary ability scales contain items that require content-specific and content-independent operations, but that are all within the same content area.

In terms of the cognitive paradigm this concept of item performance would mean that instead of looking at item performance as a single piece of data, there are the content and operation components to be considered and the operation component can be broken into operations that were content-specific and those that were content-independent. The basic components of information processing would effect each separately; for example, cognitive speed might be more relevant to the content-independent operations than the content-specific operations while the reverse may be true for, say, long term memory. As mentioned above, the type of models discussed by Embretson (1984) reflect this to some extent.

Thus the concept discussed not only provides a possible explanation for the data at hand, but it also appears consistent with two of the major paradigms of intelligence; the psychometric and the cognitive paradigms.

6.2.2 Test Theory

In terms of test theory, the results of the study are more practical in their application but no less wide ranging. The results have shown that, in an applied test development setting with examinees of a general range of abilities and using the types of tests one can expect to find in common testing situations there are few practical differences between Classical Test Theory and Item Response Theory. The two approaches produce tests that yield very similar external validities, and their parameter estimates behave similarly under conditions of varying the calibration sample composition. Finally, it has also been shown that simple number right score produces very close approximations to the more complex IRT ability estimates.

The implications of this are that for most test development situations the simpler CTT development procedures are probably adequate for the task. Certainly the finding with regard to the closeness of number right scores and the IRT ability estimates is a very positive one for the practical test developer. This means that in a pencil-and-paper test situation, tests could be developed using the sophisticated techniques of IRT but that a simple scoring procedure can be used which will give a good approximation of the estimates that would have been made using the full IRT model. Given the high degree of linearity between the two, all that is required of the test developer is to develop norm tables that equates number right score to the IRT ability estimate.

One can ask therefore, given the emphasis in the literature over the past ten years for IRT over CTT, what are the advantages of IRT over CTT? In a practical sense, the only real difference offered by IRT would at first appear to be the apparent sample-free nature of the item parameter estimates, the ability to provide standard errors of the parameter estimates made during the estimation procedure and the ability to statistically test the fit of items to the model. One could, however, argue about the utility of the standard error estimates. Certainly those provided during this study by the BILOG implementation of IRT are very wide. Moreover in this study they (the standard errors of the estimates) allowed large, consistent, differences in parameter estimates to be non-significant.

The advantage provided by IRT lies perhaps in the fact that the estimation procedures commonly used can provide standard errors of the estimates (made by a test) conditional on the estimated ability level. This information, summarised in the Test Information Function (TIF), is perhaps the most often overlooked in using IRT to develop a test, but may be one of the few advantages of IRT over CTT. The CTT analogue of the TIF is the test reliability (whether estimated by coefficient alpha or KR-20) which provides only a single score as an indicator of the accuracy of the estimate of an examinee's ability. Therefore the test developer using IRT procedures can not only estimate an examinee's ability but also gain an indication of the accuracy of the estimate of the examinee's ability.

Although the difference in parameter estimates were not significantly different in the purely statistical sense, they were certainly large enough to cause concern over the IRT claims of the invariance of the item parameter estimates. One is forced to ask just how invariant is invariant?

Certainly the early adherents of IRT, especially proponents of the Rasch model, made quite sweeping claims for IRT. Wright and Panchapakesan (1969) claim that the outcome of an individual's attempt at a test item is "...the product of the ability of the person and the easiness of the item and nothing more!" (p23) which is a considerable simplification of the situation, particularly the assumptions underlying the model. Even Rasch (1966) gives scant recognition to the one assumption that is fundamental to any latent trait model, the statistical independence of the items, or in other words the dimensionality of the test.

The mathematical invariance of the item parameter estimates is predicated on the dimensionality of the latent space. Early IRT proponents took an axiomatic approach to the dimensionality issue; tests were unidimensional, items which did not fit the model were discarded. As Wright and Panchapakesan state "The model assumes that all the items used are measuring the same trait." (p25). This was the initial concern of the earlier critics of IRT but these were largely satisfied by McDonald (1981) when he used the common factor model to provide the basis for the dimensionality assumptions required of IRT.

The model as specified by Lord and Novick (1968), however, requires a **complete** latent space and requires this for **all populations of interest.** Even when this is satisfied, the item parameter estimates are sample-free only up a linear transformation of the item difficulty and the item discrimination (Stocking & Lord, 1982). Thus the dimensionality of the test must be satisfied for all possible populations of interest in the first instance, and even if this is so, the best one can hope for is a high correlation between the two sets of item parameter estimates.

This high correlation between the two sets of item parameter estimates certainly occurred in the data here, but only for the item difficulties. The behaviour of the discrimination parameters is something of a concern, the apparent lack of relationship between the parameter estimates over the two samples being contrary to what was expected under the model. Unfortunately the limitations of the present study restrict us to simply reporting the discrepancy and speculating on the stability of the discrimination parameter. One wonders whether this is not an indication of support for the Rasch model (or single parameter models) over the two parameter models?

What is perhaps most interesting is the similarity of the behaviour of the "comparable" CTT and IRT parameter estimates, that is the p-values and the item difficulty and the item reliabilities and the item discrimination. In both cases, the two sets of parameters behaved almost identically. If one of the major advantages of IRT over CTT is the sample-free (up to a linear transformation) item parameter estimates, yet the CTT parameters behave identically, how much of an advantage is provided by the IRT procedure? IRT requires fairly strong assumptions be made about the data to achieve linearly related item parameter estimates while CTT makes no such assumptions and, in this study at least, achieves the same result.

In terms of advantages of IRT over CTT, the quality of the information provided by CTT can also be improved. Given that the p-values are simply the proportion of examinees correctly responding to an item, one can obtain a sample variance for this and from this calculate a standard error of the proportion. Certainly, this requires that the calibration sample be largely representative of the target population, but IRT also requires this of its calibration sample, and as shown in this study this is no idle requirement. Also, as has been shown above, the item parameter estimates for the two models behaved very similarly across the different calibration samples.

6.2.3 Latent Traits versus True Scores

Despite the apparent differences in the concepts of latent traits and true scores, are they really that different? The main differences between the two are really to do with the different emphases placed on the trait being measured and an individual's standing on this trait.

The similarities between the true-score based CTT model and the latent-trait based IRT model were formalised by Lord and Novick (1968). They showed not only the fundamental relationship between true scores and latent traits but also that the p-value is directly related to the b parameter and that the a parameter is a known monotonic increasing function of the biserial correlation between the item and the latent trait, which is measured in CTT by the item reliability. The similarities observed between the results of the two analyses conducted here support this and the work of Douglass et al (1979).

6.3 Improvements to the Study

The concept of content and operation components in test item performance requires considerable further investigation. In this study only a limited range of secondary abilities were examined and the findings could benefit from being able to be generalised across a wider range of abilities. In particular only a limited range of fluid ability items were included in the study.

Having examined fluid ability as a secondary ability, and finding support for the contention that it may in fact be simply a manifestation of general ability (within a hierarchical model), the next step would be to test the hierarchical model at the item level. The aim in the present study was limited to determining whether it was necessary to include fluid ability in a test of general ability. Determining the validity of Gustafsson's (1984) unifying model, at the item level, would require a complete investigation, but the results of this study indicate that it may be a fruitful investigation.

In general, the limitations to the investigation of the factor structure of the model of intelligence are primarily in the range of items used and in the need to extend the study to fully examine the hierarchical model of intelligence.

The study provided an applied comparison of the utility of the CTT and IRT test development methodologies in a relatively common test development situation. Given the similarities found in the performance of the two methodologies in this setting, the next step would be to broaden the scope of the comparison of the methodologies to encompass both a wider range of abilities of examinee and a wider range of test types. Although the practical similarities between the two have been shown within the usual limits of test development, the more refined statistics available to the IRT methodology should provide for much better estimation at the more extreme limits of testing.

To adequately test the dimensionality issues raised, it would be necessary to compare tests of varying degrees of dimensionality. For example, a test could be developed with item loadings at, say, 0.9 or better, and comparing this with a test that met the "standard" factor analysis criteria, that is items loading 0.3 or better. These two tests would be compared in terms of their items' fit to the IRT model chosen and the stability of their item parameter estimates over calibration samples with different compositions. In this case one would hypothesise that the more unidimensional the test, the better the fit to the IRT model and the more stable the item parameter estimates.

Another possible problem that was highlighted in this study was the accuracy of the estimates provided by the IRT development procedure. The literature has shown that BILOG is one of the better implementations of the IRT (Mislevy and Stocking, 1989) procedure and that the sample sizes used were adequate for the task (Harwell & Janosky, 1991). Despite this, the standard errors provided in the estimation of the item parameter estimates were such that large, consistent, differences in the estimates were ultimately non-significant. This was particularly noticeable for the more extreme parameter estimates. Given that this information provides one of IRT's main advantages over CTT, it should be investigated.

Finally, one of the main findings of the study was the similarity of behaviour of the CTT and IRT parameter estimates across different samples. Given the potential importance of this finding for the practical test developer this is a finding that would be well worth replicating.

6.4 Contributions of the Study

The primary contribution of this study is to directly compare the IRT and CTT test development methodologies in an applied test development setting; this is something that has been lacking in the literature to date. It has shown that for normal test development purposes, there are few differences between the two models and in the results of the test they produce. Moreover, it has been shown that simple number right scores can provide a very close approximation of IRT ability estimates. Thus, a test could be developed using the precision of IRT, and then administered with a very simple scoring formula, thus reducing the complexity of administering the test while retaining the power of the IRT development procedure.

More general questions have been raised about the similarities between the two models. Some of the primary advantages of IRT over CTT have been shown to not occur in this applied setting. In particular the item parameter estimates behaved almost identically across two substantively different samples of examinees.

An examination of the factor structure of the set of items used for the test development has yielded some further support for the notion that fluid ability is a manifestation of general ability rather than being a secondary ability. This remains to be conclusively tested, but has provided a different way of looking at performance on a test item, and one that appears to be consistent with the two major paradigms of intelligence.

REFERENCES

Ackerman, T. (1987) The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence. ACT Research Report Series 87-14.

Andrich, D. (1988) Rasch Models for Measurement. Sage University Paper series on Quantative Applications in the Social Sciences, 07-001. Beverly Hills: Sage Publications.

Bachelor, P.A. (1989) Maximum likelihood confirmatory factor-analytic investigation of factors within Guilford's structure of intellect model. Journal of Applied Psychology, 74, 797-804.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Bock, R.D. and Lieberman, M. (1970) Fitting a response model for n dichotomously scored items. Psychometrika, 35, 179-197.

Boring, E.G. (1929) Intelligence as the tests test it. New Republic, 34, 33-37.

Burt, C. (1909) Experimental tests of general intelligence. British Journal of Psychology, 3, 94-177.

Burt, C. (1939) The relations of educational abilities. British Journal of Educational Psychology, 9, 45-71.

Burt, C. (1940) The Factors of the Mind. London: University of London Press.

Butcher, H.J. (1968) Human Intelligence: Its Nature and Assessment. London: Methuen.

Carpenter, P.A., Just, M. A. and Shell, P. (1990) What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. Psychological Review, 97, 404-431.

Carroll, J.B. (1981) Ability and task difficulty in cognitive psychology. Educational Researcher, 10, 11-21.

Carroll, J.B. (1991) Cognitive psychology's psychometric lawgiver. Contemporary Psychology, 36(7), 557-559.

Cattell, R.B. (1966) The scree test for the number of factors. Multivariate Behavioural Research, 1, 245-267.

Cattell, R.B. (1971) Abilities: Their Structure, Growth and Action. Boston: Houghton-Mifflin.

Christoffersson, A. (1975) Factor Analysis of Dichotomous Variables. Psychometrika, 40(1), pp 5-32.

Cronbach, L.J. (1970) Essentials of Psychological Testing. New York: Harper and Row.

Dawes, R.M. and Corrigan, B. (1974) Linear models in decision making. Psychological Bulletin. 81(2), 95-106.

Detterman, D.K. (1982) Does "g" exist? Intelligence, 6, 99-108.

Detterman, D.K. Basic Cognitive Processes Predict IQ. Paper presented at the Army Personnel Research Establishment Seminar on Intelligence and Psychometrics. Plymouth Polytechnic, September, 1987.

Dinero, T.E. and Haertel, E. (1977) Applicability of the Rasch model with varying item discriminations. Applied Psychological Measurement. 1(4), 581-592.

Douglass, F.M., Khavari, K.A., and Farber, P.D. (1979) A comparison of classical and latent trait item analysis procedures. Educational and Psychological Measurement. 39, 337-352.

Einhorn, H.J. and Hogartrh, R.M. (1975) Unit weighting schemes for decision making. Organizational Behaviour and Human Performance. 13, 171-192.

Embretson, S.E. (1984) A general latent trait model for response processes. Psychometrika. 49, 175-186.

Embretson, S.E. and Wetzel, C.D. (1987) Component latent trait models for paragraph comprehension tests. Applied Psychological Measurement. 11(2), 175-193.

Fraser, C. (1988) NOHARM Users Guide. University of New England.

Galton, F. (1888) Co-relations and their measurement, chiefly from anthropometric data. Proceedings of the Royal Society of London, XLV, 135-145.

Goldstein, H. (1980) Dimensionality, bias, independence and measurement scale problems in latent trait test score models. British Journal of Mathematical and Statistical Psychology. 33, 234-246.

Guilford, J.P. (1956) Psychometric Methods. New York: McGraw-Hill.

Guilford, J.P. (1967) The Nature of Human Intelligence. New York: McGraw-Hill.

Guilford, J.P. (1984) Varieties of divergent production. Journal of Creative Behaviour, 18, 1-10.

Guilford, J.P. (1985) The Structure of Intellect Model. In B.B. Wolman (Ed) Handbook of Intelligence. (pp. 225-266) New York: John Wiley and Sons.

Gulliksen, H., (1951) Theory of Mental Tests. New York: Wiley.

Gustafsson, J. (1984) A unifying model for the structure of intellectual abilities. Intelligence, 8, 179-203.

Hambleton, R.K. and Cook, L.L. (1983) Robustness of Item Response Models and Effects of Test Length and Sample Size on the Precision of Ability Estimates. In Weiss D.J. (Ed) New Horizons in Testing. (pp. 31-49) New York:Academic Press.

Hambleton, R.K. and Swaminathan, H. (1985) Item Response Theory: Principles and Applications. Boston: Kluwer-Nijhoff.

Harwell, M.R. and Janosky, J.E. (1991) An empirical study into the effects of small datasets and varying prior variances on item parameter estimation in BILOG. Applied Psychological Measurement. 15(3), 279-291.

Hogg, R.V. and Craig, A.T. (1978) Introduction to Mathematical Statistics. New York: Macmillan.

Horn, J.L. (1988) Thinking about human abilities. In J.R. Nesselroade and R.B. Cattell (Eds) Handbook of Multivariate Experimental Psychology (2nd ed., pp 645-685). New York: Plenum.

Horn, J.L. and Cattell, R.B. (1966) Refinement and test of the theory of fluid and crystallised intelligence. Journal of Educational Intelligence, 57, 253-270.

Horn, J.L. and Knapp, J.R. (1973) On the subjective character of the empirical base of Guilford's structure-of-intellect model. Psychological Bulletin, 80, 33-43.

Hunt, E. (1982) Towards new ways of assessing intelligence. Intelligence, 6, 231-240.

Irvine, S..H., Dann, P.L. and Anderson, J.D. (1990) Towards a theory of algorithm-determined cognitive test construction. British Journal of Psychology, 81, 173-195.

Jensen, A.R. (1982) Reaction Time and Psychometric g. In H.J. Eysenck (Ed) A Model for Intelligence. New York: Springer-Verlag.

Joreskog, K.G. (1969) A general approach to confirmatory maximum likelihood factor analysis Psychometrika, 34, 183-202.

Joreskog, K.G. (1970) A general method for analysis of covariance structures. Biometrika, 57, 239-251.

Kaiser, H.F. (1960) The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20, 141-151.

Kelderman, H., Mellenbergh, G.J., and Elshout, J.J. (1981) Guilford's facet theory of intelligence: An empirical comparison of models. Multivariate Behavioural Research, 16, 37-61.

Kline, P. (1986) A Handbook of Test Construction. London: Methuen.

Lawley, D.N. (1943) On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 61, 273-287.

Lord, F.M. (1983) Small N Justifies Rasch Model. In Weiss (Ed) New Horizons in Testing. (pp. 51-61) New York: Academic Press.

Lord, F.M. and Novick, M.R. (1968) Statistical Theories of Mental Test Scores. New York: Addison Wesley

Lumsden, J. (1976) Test Theory. In Rosenzweig, M.R. and Porter, L.W. (Eds) Annual Review of Psychology. Palo Alto: Annual Reviews Inc.

Marsh, H.W., Balla, J.R. and McDonald, R.P. (1988) Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. Psychological Bulletin, 103(3), 391-410.

Maybery, T. (1990) Sternberg's mixed model applied to indeterminate linear syllogisms: A mismatch. British Journal of Psychology, 81, 271-283.

McDonald, R.P. (1981) The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology. 34 100-117.

Mislevy, R.J. and Bock, R.D. (1990) Bilog 3: Item Analysis and Test Scoring with Binary Logistic Models. Mooresville: Scientific Software.

Mislevy, R.J. and Stocking, M.L. (1989) A consumer's guide to LOGIST and BILOG. Applied Psychological Measurement. 13(1), 57-75.

Nunnally, J.C. (1978) Psychometric Theory. New York: McGraw-Hill.

Pearson, K (1900) On the correlation of characters not quantitatively measurable. Royal Society Philosophical Transactions, Series A, 195, 1-47.

Pellegrino, J.W., and Glaser, R. (1980) Components of inductive reasoning. In R.E. Snow, P.A. Federico and W.A. Montague (Eds) Aptitude, learning and instruction: Cognitive process analysis of aptitude (Vol. 1) Hillsdale, NJ: Erlbaum.

Pellegrino, J.W. and Glaser, R. (1982) Analyzing aptitudes for learning: Inductive reasoning. In R. Glaser (Ed) Advances in instructional psychology (Vol. 2) Hillsdale, NJ: Erlbaum.

Pellegrino, J.W. and Kail, R. (1982) Process Analysis of Spatial Aptitude. In R.J. Sternberg (Ed) Advances in tne Psychology of Human Intelligence, Vol. 1. Hillsdale, NJ: Earlbaum.

Rasch, G. (1960) Probabilistic Models for some intelligence and attainment tests. Copenhagen: Nielson and Lydiche (for Danmarks Pedagogiske Institut).

Rasch, G. (1966) An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology. 19(1), 49-57.

Smith, P. (1986) Application of the Information Processing Approach to the Design of a Non-verbal Reasoning Test. British Journal of Educational Psychology, 56, 119-137.

Snow, R.E. (1979) Theory and method for research on aptitude processes. In R.J. Sternberg and D.K. Detterman (Eds). Human intelligence: Perspectives on its theory and measurement. Norwood, NJ: Ablex.

Spearman, C. (1904a) "General Intelligence," objectively determined and measured. The American Journal of Psychology, 15, 201-292.

Spearman, C. (1904b) The proof and measurement of association between two things. American Journal of Psychology, 15, 72-101.

Sternberg, R.J. (1977) Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities. Hillsdale, NJ: Erlbaum.

Sternberg, R.J. (1980) Sketch of a componential subtheory of human intelligence. Behavioural and Brain Sciences, 3, 573-614.

Sternberg, R.J. (1985) Beyond IQ: A Triarchic Theory of Human Intelligence. New York: Cambridge University Press.

Sternberg, R.J. (1985) Cognitive Approaches to Intelligence. In B.B. Wolman (Ed) Handbook of Intelligence. (pp. 59-118) New York: John Wiley & Sons.

Sternberg, R.J. and Gardner, M.K. (1982) A Componential Interpretation of the General Factor in Human Intelligence. In H.J. Eysenck (Ed) A Model for Intelligence. New York: Springer-Verlag.

Stocking, M.L. and Lord, F.M. (1982) Developing a Common Metric in Item Response Theory (Educational Testing Service Research Research Report RR-82-25-ONR). Princeton, N.J.: Educational Testing Service.

Tinkleman, S.N. (1971) Planning the Objective test. In R.L. Thorndike (Ed) Educational Measurement. (2nd Ed: pp. 46-80) Washington: American Council on Education.

Thurstone, L.L. (1938) Primary Mental Abilities. Psychometric Monographs, Number 1.

Thurstone, L.L. (1940) An experimental study of simple structure. Psychometrika, 5, 153-168.

Thurstone, L.L. (1947) Multiple Factor Analysis. Chicago: University of Chicago Press.

Thurstone, L.L. and Thurstone, T.G. (1941) Factorial studies of intelligence. Psychometric Monographs, Number 2.

Undheim, J.O. (1981a) On intelligence II: A neo-Spearman model to replace Cattell's theory of fluid and crystallized ability. Scandinavian Journal of Psychology, 22, 181-187.

Undheim, J.O. (1981b) On intelligence IV: Toward a restoration of general intelligence. Scandinavian Journal of Psychology, 22, 1251-265.

Undheim, J.O. and Gustafsson, J. (1987) The hierarchical organization of cognitive abilities: restoring general intelligence through the use of linear structural relations (LISREL). Multivariate Behavioral Research, 22, 149-171.

Undheim, J.O. and Horn, J.L. (1977) Critical evaluation of Guilford's structure-of-intellect theory. Intelligence, 1, 65-81.

Velicer, W.F. and Jackson, D.N. (1990) Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. Multivariate Behavioural Research, 25 (1), 1-28.

Vernon P.E. (1947) Research on personnel selection in the Royal Navy and the British Army. American Psychologist, 2, 35-51.

Vernon, P.E. (1950) The Structure of Human Abilities. London: Methuen.

Vernon, P.E. (1979) Intelligence: Heedity and Envionment. San Francisco: Freeman.

Wainer, H. (1976) Estimating coefficiennts in linear models: it don't make no nevermind. Psychological Bulletin. 83(2), 213-217.

Wesman, A.G. (1971) Writing the Test Item. In R.L. Thorndike (Ed) Educational Measurement. (2nd Ed: pp. 81-129) Washington: American Council on Education.

Wilks, S.S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. Psychometrika. 3(1), 23-40.

Wingersky, M.S., Barton, M.A., and Lord, F.M. (1982) LOGIST user's guide. Princeton, NJ: Educational Testing Service

Wright, B. and Panchapakesan, N. (1969) A proccedure for sample-free item analysis. Educational and Psychological Measurement. 29, 23-48.

ANNEX A

UNIQUE VARIANCES - DIFFERENT FACTOR MODELS

Models

Item	3	4	1	Reject
1	0.94	1	0.947	*
2	0.64	1	0.694	
3	0.93	0.99	0.933	*
4	0.8	1	0.849	*
5	0.88	0.99	0.908	*
6	0.92	0.92	0.926	*
7	0.71	0.61	0.743	
8	0.48	0.92	0.579	
9	0.59	0.75	0.64	
10	0.55	0.84	0.61	
11	0.93	0.92	0.95	*
12	0.64	0.71	0.69	
13	0.77	0.57	0.8	
14	0.88	0.56	0.9	*
15	0.84	0.6	0.88	*
16	0.7	0.93	0.76	
17	0.73	0.63	0.76	
18	0.8	0.74	0.83	*
19	0.59	0.88	0.64	
20	0.57	0.76	0.63	
21	0.97	0.76	0.98	*
22	0.9	0.76	0.92	*
23	0.64	0.81	0.68	
24	0.83	0.59	0.86	*
25	0.55	0.59	0.63	
26	0.44	0.97	0.51	
27	0.39	0.89	0.48	
28	0.51	0.63	0.6	
29	0.57	0.76	0.6	
30	0.71	0.51	0.75	
31	0.79	0.49	0.82	
32	0.95	0.5	0.95	*
33	0.57	0.57	0.64	
34	0.68	0.57	0.75	
35	0.6	0.71	0.69	
36	0.75	0.75	0.8	
37	0.41	0.97	0.46	
38	0.87	0.53	0.88	*
39	0.44	0.55	0.52	
40	0.91	0.45	0.92	*

Models

Item	3	4	1	Reject
41	0.55	0.79	0.62	
42	0.82	0.44	0.84	*
43	0.97	0.88	0.97	*
44	0.87	0.5	0.9	*
45	1	0.89	1	*
46	1	0.51	1	*
47	0.97	0.8	0.97	*
48	0.9	0.97	0.91	*
49	0.82	0.84	0.84	*
50	0.86	1	0.88	*
51	0.99	1	0.99	*
52	0.89	0.97	0.9	*

ANNEX B

CLASSICAL TEST THEORY ITEM STATISTICS - T0

CLASSICAL ITEM STATISTICS FOR NUMERIC SUBTEST

ITEM NAME	NUMBER TRIED	NUMBER RIGHT	P-VALUE	LOGIT/1.7	ITEM*TEST CORR BISERIAL
Q7	209.0	152.0	.727	.58	.431
Q9	209.0	174.0	.833	.94	.453
Q10	209.0	134.0	.641	.34	.596
Q18	209.0	179.0	.856	1.05	.472
Q19	209.0	91.0	.435	-.15	.589
Q20	209.0	137.0	.656	.38	.525
Q26	209.0	137.0	.656	.38	.576
Q29	209.0	192.0	.919	1.43	.379
Q30	209.0	47.0	.225	-.73	.507
Q38	209.0	49.0	.234	-.70	.343
Q39	209.0	50.0	.239	-.68	.636
Q40	209.0	29.0	.139	-1.07	.252
Q48	209.0	32.0	.153	-1.01	.299
Q49	209.0	53.0	.254	-.64	.395

CLASSICAL ITEM STATISTICS FOR SPATIAL SUBTEST

ITEM NAME	NUMBER TRIED	NUMBER RIGHT	P-VALUE	LOGIT/1.7	ITEM*TEST CORR BISERIAL
Q4	209.0	54.0	.258	-.62	.266
Q5	209.0	78.0	.373	-.30	.242
Q8	209.0	191.0	.914	1.39	.627
Q14	209.0	203.0	.971	2.07	.212
Q15	209.0	200.0	.957	1.82	.431
Q16	209.0	150.0	.718	.55	.394
Q24	209.0	197.0	.943	1.65	.423
Q25	209.0	204.0	.976	2.18	.643
Q28	209.0	182.0	.871	1.12	.559
Q34	209.0	160.0	.766	.70	.443
Q35	209.0	160.0	.766	.70	.524
Q36	209.0	163.0	.780	.74	.352
Q44	209.0	156.0	.746	.64	.365
Q45	209.0	5.0	.024	-2.18	.032
Q47	209.0	85.0	.407	-.22	.017

CLASSICAL ITEM STATISTICS FOR VERBAL SUBTEST

	NUMBER	NUMBER			ITEM*TEST CORR
ITEM NAME	TRIED	RIGHT	P-VALUE	LOGIT/1.7	BISERIAL
Q1	209.0	159.0	.761	.68	.225
Q2	209.0	156.0	.746	.64	.569
Q3	209.0	110.0	.526	.06	.143
Q6	209.0	137.0	.656	.38	.209
Q11	209.0	165.0	.789	.78	.281
Q12	209.0	139.0	.665	.40	.532
Q13	209.0	165.0	.789	.78	.397
Q17	209.0	76.0	.364	-.33	.388
Q21	209.0	167.0	.799	.81	.221
Q22	209.0	156.0	.746	.64	.271
Q23	209.0	191.0	.914	1.39	.516
Q27	209.0	96.0	.459	-.10	.596
Q31	209.0	134.0	.641	.34	.381
Q32	209.0	33.0	.158	-.98	.121
Q33	209.0	184.0	.880	1.17	.668
Q37	209.0	181.0	.866	1.10	.569
Q41	209.0	166.0	.794	.79	.618
Q42	209.0	94.0	.450	-.12	.309
Q43	209.0	45.0	.215	-.76	.142
Q46	209.0	98.0	.469	-.07	-.030
Q50	209.0	88.0	.421	-.19	.242
Q51	209.0	29.0	.139	-1.07	.011
Q52	209.0	101.0	.483	-.04	.313

ANNEX C

ITEM RESPONSE THEORY ITEM PARAMETER ESTIMATES - TO

SUBTEST NUMERIC : ITEM PARAMETERS AFTER CYCLE 12

ITEM	INTERCEPT	SLOPE	THRESHOLD	DISPERSN	ASYMP	STD
	S.E.	S.E.	S.E.	S.E.	S.E.	POST RESID
Q7	.744	.716	-1.040	1.397	.000	.820
	.136*	.197*	.247*	.384*	.000*	
Q9	1.217	.791	-1.539	1.264	.000	.537
	.187*	.236*	.341*	.378*	.000*	
Q10	.587	1.137	-.517	.880	.000	.983
	.163*	.302*	.125*	.234*	.000*	
Q18	1.372	.836	-1.642	1.196	.000	1.048
	.221*	.272*	.382*	.390*	.000*	
Q19	-.152	1.026	.148	.975	.000	1.515
	.128*	.277*	.138*	.263*	.000*	
Q20	.610	1.047	-.583	.955	.000	.732
	.150*	.243*	.134*	.221*	.000*	
Q26	.684	1.263	-.542	.792	.000	.424
	.172*	.309*	.119*	.194*	.000*	
Q29	1.685	.669	-2.518	1.494	.000	1.842
	.291*	.347*	.986*	.775*	.000*	
Q30	-.892	.698	1.277	1.432	.000	1.608
	.140*	.190*	.329*	.390*	.000*	
Q38	-.756	.395	1.916	2.533	.000	.887
	.114*	.123*	.598*	.786*	.000*	
Q39	-.960	1.064	.902	.940	.000	.405
	.180*	.277*	.210*	.244*	.000*	
Q40	-1.130	.303	3.732	3.304	.000	1.174
	.142*	.131*	1.557*	1.429*	.000*	
Q48	-1.076	.350	3.075	2.858	.000	.801
	.135*	.138*	1.121*	1.130*	.000*	
Q49	-.724	.525	1.380	1.905	.000	.387
	.124*	.145*	.391*	.525*	.000*	

* STANDARD ERROR

SUBTEST SPATIAL : ITEM PARAMETERS AFTER CYCLE 12

ITEM	INTERCEPT S.E.	SLOPE S.E.	THRESHOLD S.E.	DISPERSN S.E.	ASYMP S.E.	STD POST RESID
Q4	-.707 .131*	.476 .157*	1.486 .471*	2.100 .691*	.000 .000*	1.195
Q5	-.320 .100*	.393 .142*	.816 .366*	2.546 .923*	.000 .000*	1.339
Q8	2.240 .527*	1.297 .541*	-1.727 .400*	.771 .322*	.000 .000*	.524
Q14	2.161 .286*	.316 .309*	-6.835 6.523*	3.163 3.089*	.000 .000*	1.205
Q15	2.100 .295*	.535 .306*	-3.926 1.936*	1.870 1.070*	.000 .000*	.693
Q16	.682 .125*	.580 .191*	-1.176 .343*	1.725 .568*	.000 .000*	.562
Q24	1.954 .287*	.780 .289*	-2.505 .788*	1.282 .475*	.000 .000*	1.410
Q25	3.442 .749*	1.371 .525*	-2.511 .543*	.729 .280*	.000 .000*	.835
Q28	1.717 .333*	1.188 .419*	-1.445 .314*	.842 .297*	.000 .000*	.694
Q34	.924 .166*	.766 .243*	-1.206 .306*	1.305 .415*	.000 .000*	.818
Q35	1.251 .290*	1.323 .423*	-.946 .174*	.756 .241*	.000 .000*	.653
Q36	.875 .140*	.509 .217*	-1.719 .632*	1.966 .838*	.000 .000*	.944
Q44	.770 .140*	.630 .236*	-1.222 .387*	1.586 .594*	.000 .000*	.696
Q45	-2.180 .290*	.160 .596*	13.610 50.215*	6.243 23.207*	.000 .000*	2.561
Q47	-.217 .090*	.018 .124*	11.855 79.632*	54.646 369.532*	.000 .000*	.499

* STANDARD ERROR

```
SUBTEST VERBAL   :   ITEM PARAMETERS AFTER CYCLE  12

ITEM      INTER     SLOPE     THRESHOLD DISPERSN    ASYMP   CHISQ    DF
          S.E.      S.E.      S.E.      S.E.        S.E.    (PROB)
------------------------------------------------------------------------
Q1         .730      .339      -2.153     2.952      .000      2.9  5.0
           .102*     .128*       .793*    1.114*     .000*  ( .7173)

Q2         .949     1.040       -.912      .961      .000      2.2  2.0
           .148*     .218*       .150*     .202*     .000*  ( .3321)

Q3         .065      .169       -.383     5.930      .000     12.2  6.0
           .083*     .092*       .520*    3.240*     .000*  ( .0568)

Q6         .400      .279      -1.432     3.582      .000       .8  5.0
           .089*     .112*       .609*    1.437*     .000*  ( .9758)

Q11        .847      .388      -2.180     2.574      .000      3.0  4.0
           .109*     .137*       .736*     .909*     .000*  ( .5572)

Q12        .593      .938       -.632     1.066      .000       .6  3.0
           .127*     .222*       .132*     .252*     .000*  ( .8902)

Q13        .937      .625      -1.499     1.600      .000      6.0  3.0
           .136*     .170*       .327*     .435*     .000*  ( .1119)

Q17       -.374      .530        .706     1.887      .000      2.0  4.0
           .094*     .122*       .227*     .434*     .000*  ( .7311)

Q21        .839      .227      -3.697     4.408      .000      9.7  5.0
           .105*     .117*      1.868*    2.277*     .000*  ( .0828)

Q22        .692      .379      -1.828     2.642      .000      1.4  5.0
           .103*     .121*       .561*     .842*     .000*  ( .9208)

Q23       1.731      .772      -2.243     1.295      .000      1.8  2.0
           .223*     .226*       .503*     .380*     .000*  ( .4169)

Q27       -.095     1.257        .076      .796      .000      4.9  3.0
           .118*     .287*       .099*     .182*     .000*  ( .1813)

Q31        .392      .471       -.833     2.125      .000      9.0  5.0
           .092*     .123*       .272*     .556*     .000*  ( .1080)

Q32      -1.000      .168       5.936     5.937      .000      2.4  4.0
           .113*     .102*      3.592*    3.612*     .000*  ( .6700)

Q33       2.057     1.545      -1.331      .647      .000       .3  1.0
           .386*     .429*       .178*     .180*     .000*  ( .6088)

Q37       1.558     1.033      -1.508      .968      .000       .2  2.0
           .236*     .292*       .277*     .273*     .000*  ( .9148)
------------------------------------------------------------------------
```

```
SUBTEST VERBAL  :  ITEM PARAMETERS AFTER CYCLE  12

ITEM    INTER     SLOPE     THRESHOLD DISPERSN   ASYMP   CHISQ    DF
        S.E.      S.E.      S.E.      S.E.       S.E.    (PROB)
---------------------------------------------------------------------
Q41   |  1.330  |  1.325  |  -1.004  |    .755  |  .000  |      .0  1.0
      |   .250* |   .354* |    .145* |    .202* |  .000* | ( .8562)
      |         |         |          |          |        |
Q42   |  -.125  |   .390  |    .320  |   2.562  |  .000  |     4.0  4.0
      |   .087* |   .115* |    .239* |    .757* |  .000* | ( .4045)
      |         |         |          |          |        |
Q43   |  -.775  |   .182  |   4.259  |   5.496  |  .000  |     2.4  5.0
      |   .102* |   .116* |   2.679* |   3.505* |  .000* | ( .7887)
      |         |         |          |          |        |
Q46   |  -.073  |   .000  |1497.952  |********  |  .000  |     3.1  6.0
      |   .082* |   .093* |********* |********  |  .000* | ( .7982)
      |         |         |          |          |        |
Q50   |  -.203  |   .436  |    .465  |   2.294  |  .000  |     4.6  4.0
      |   .089* |   .115* |    .242* |    .607* |  .000* | ( .3338)
      |         |         |          |          |        |
Q51   | -1.079  |   .094  |  11.481  |  10.644  |  .000  |     2.0  4.0
      |   .119* |   .151* |  18.360* |  17.148* |  .000* | ( .7453)
      |         |         |          |          |        |
Q52   |  -.038  |   .430  |    .088  |   2.328  |  .000  |     4.7  4.0
      |   .088* |   .116* |    .208* |    .628* |  .000* | ( .3162)
---------------------------------------------------------------------
                                              * STANDARD ERROR

********* - means that the statistic could not be calculated.
```

ANNEX D

T1 ITEM-MODEL FIT STATISTICS: TWO PARAMETER AND RASCH MODELS

Scale	Item	2PL Model	Rasch Model
Numeric	7	0.820	1.160
	9	0.537	0.631
	10	0.983	2.525
	18	1.048	0.862
	20	0.732	0.762
	26	0.424	1.107
	39	0.405	2.077
Verbal	2	0.332	0.012
	12	0.890	0.009
	17	0.731	0.008
	23	0.417	0.583
	27	0.181	0.001
	37	0.915	0.074
	41	0.856	0.010
Spatial	8	0.524	1.918
	16	0.562	0.359
	25	0.835	1.301
	28	0.694	1.603
	34	0.813	1.233
	35	0.635	1.609

Note: Standardised residuals are presented for the Numeric and Spatial scales while the probabilities associated with the Chi-Square test are presented for the Verbal scale.

ANNEX E

ITEM PARAMETER ESTIMATES FOR BOTH SAMPLES

CTT Item Parameter Estimates

Item	T0 Sample		T1 Sample	
Numeric	p value	r_{bis}	p value	r_{bis}
7	0.727	0.436	0.814	0.280
10	0.641	0.564	0.811	0.551
19	0.435	0.517	0.544	0.562
20	0.656	0.582	0.798	0.563
26	0.656	0.617	0.795	0.468
30	0.225	0.490	0.461	0.333
39	0.239	0.558	0.334	0.344
	Alpha =	0.726	Alpha = 0.660	
Verbal	p value	r_{bis}	p value	r_{bis}
2	0.746	0.574	0.879	0.383
12	0.665	0.498	0.771	0.475
13	0.789	0.363	0.849	0.552
17	0.364	0.348	0.526	0.452
27	0.459	0.558	0.647	0.496
33	0.880	0.703	0.919	0.413
41	0.794	0.599	0.868	0.137
	Alpha =	0.719	Alpha = 0.580	

	CTT Item Parameter Estimates			
Item	T0 Sample		T1 Sample	
Spatial	p value	r_{bis}	p value	r_{bis}
4	0.258	0.392	0.388	0.257
5	0.373	0.253	0.582	0.388
16	0.718	0.381	0.841	0.474
34	0.766	0.419	0.846	0.576
35	0.766	0.462	0.787	0.484
36	0.780	0.181	0.776	0.199
	Alpha =	0.586	Alpha = 0.614	

p value - item facility r_{bis} - item-total biserial correlation

IRT Item Parameter Estimates

Item	T0 Sample		T1 Sample	
Numeric	a	b	a	b
7	0.467	-1.469	0.422	-2.432
	0.123	0.410	0.109	0.648
9	0.585	-2.032	0.826	-2.453
	0.181	0.670	0.256	0.913
10	0.772	-0.650	0.827	-1.659
	0.173	0.220	0.203	0.520
18	0.528	-2.409	0.263	-5.697
	0.166	0.790	0.150	3.274
20	0.847	-0.690	0.759	-1.624
	0.191	0.237	0.174	0.461
26	0.927	-0.665	0.571	-1.849
	0.219	0.251	0.1266	0.451
39	0.884	1.196	0.323	1.406
	0.246	0.431	0.086	0.376
Verbal	a	b	a	b
2	0.954	-1.063	0.457	-3.036
	0.260	0.375	0.113	0.773
12	0.656	-0.825	0.424	-1.993
	0.175	0.271	0.100	0.489
17	0.460	0.852	0.615	-0.122
	0.123	0.262	0.134	0.117
23	0.688	-2.643	0.498	-4.336
	0.205	0.854	0.160	1.434
27	0.923	0.160	0.799	-0.685
	0.236	0.169	0.201	0.239
37	0.795	-1.954	0.447	-3.280
	0.240	0.672	0.133	0.999
41	0.639	-1.625	0.233	-5.016
	0.173	0.486	0.102	2.200

IRT Item Parameter Estimates

Item	T0 Sample		T1 Sample	
Spatial	a	b	a	b
8	0.868	-2.404	0.586	-4.150
	0.334	1.094	0.338	2.518
16	0.371	-1.675	0.460	-2.588
	0.116	0.546	0.107	0.630
25	0.680	-4.115	0.482	-3.776
	0.283	1.854	0.143	1.162
28	0.538	-2.378	0.519	-3.454
	0.170	0.859	0.147	1.023
34	0.609	-1.507	1.066	-1.785
	0.162	0.453	0.321	0.802
35	1.709	-1.068	0.960	-1.430
	0.785	1.136	0.245	0.523

a - item discrimination parameter   b - item difficulty parameter

The figure immediately under the parameter estimate is the standard error for the estimate.

ANNEX F

DETAILS OF NOHARM ANALYSIS

Background

The problems normally associated with applying factor analysis to binary (or dichotomous) data stems from the use of either the phi coefficient or the tetrachoric correlation coefficient as the bivariate measure of association. Because the range of values that can be taken by the phi coefficient is dependent on the p-values of the individual items, it was originally thought that applying factor analysis to a matrix of phi coefficients would lead to the extraction of factors based solely on the difficulty of the items in the data sets (McDonald, 1985). The tetrachoric correlation coefficient, on the other hand, leads to problems because most factor analytic methods require that the correlation matrix be Grammian (Christoffersson, 1975), a situation theat does not always occur with a matrix of tetrachoric correlation coefficients.

McDonald (1967) showed, however, that while traditional linear factor analytic methods were unsuitable for application to binary data, nonlinear methods, particularly those derived from latent trait item analysis, were suited to this data. Christoffersson (1975) developed a method, based on the work of McDonald (1967), for fitting a normal ogive model to a set of item covariances. He did this by using generalised least squares to estimate a ten-term series approximation to the tetrachoric function which he used to express the item covariances (Balla & McDonald, 1985).

Fraser (1988) implemented essentially the same method for fitting the normal ogive model in his program NOHARM, but substituted ordinary least squares and only used a three-term approximation to the tetrachoric function (Balla & McDonald, 1985).

Using NOHARM

The user of NOHARM supplies a matrix of scored item responses (i.e., ones or zeros corresponding to either correct or incorrect responses respectively) along with various parameters of the model to be fitted. Depending on the model to be fitted, the user also supplies a variety of information concerning the data supplied and the output required, values representing the probability of examinees correctly guessing the answer to each item, and a pattern matrix to indicate which variables are to load on which factor.

NOHARM allows the user to constrain parameters of the pattern matrix to be equal to zero (i.e., the item does not load on the factor), to be equal to another parameter that is to be estimated (for example all items loading on the same factor can be estimated to have the same loading, and Fraser (1988) uses this method as an example of estimatin the rasch model), finally, parameters can be estimated independently of all other parameters (this is the technique used for estimated general factor analysis models). the NOHARM user thus has the flexibility to estimate a broad range of models simply on the basis of the pattern of coefficients supplied in the pattern matrix.

NOHARM supplies three indices of the fit of the model to the data: the residual inter-item covariances, the root mean square of the residual covariances (rmsr), and the unique variances of trhe individual items (the additive inverse of the communalities) after the model has been estimated. McDonald (1981) asserts that the dimensionality of a set of variables can be ascertained if the residual covariances, after fitting the appropriate model, are small. Thus Fraser (1988) provide an approximate value against which to test the obtained RMSR, namely four times the reciprocal of the sample size. He also states that the individual residual covariances can be examined and any patterns of high values taken as an indicator that there may be further factors to be extracted from the data. Finally, he used in the current study, the unique variances give some measure of how well individual items are represented by the hypothesized factor structure.

The Current Analysis

In the current analysis, a matrix of item responses was obtained and scored using an SPSS* program written by the author; these item scores were then analyses using NOHARM. Three NOHARM analyses were conducted: in the first, a single factor was hypothesised with all items loading on the factor and all of the loadings estimated independently of each othe. In the second analysis three factors were hypothesised; items identified as tapping the verbal, numeric and spatial domains were each loaded on to one of the factors and their loadings were then estimated independently of the other parameters. In the final analysis, four factors were hypothesised, the three above and the fourth representing the "fluid" ability domain. In this case items from the other three domains that were identified as tapping the fluid domain (e.g., number analogy itmes as opposed to arithmetic items) were loaded onto the fluid factor and then the loading estimated, again independently of other parameters. In no case was a complex solution estimated, that is, all itmes were estimated loading onto only one factor.

In the current analysis guessing parameters were set to the inverse of the number of response choices and the factors to be extracted were uncorrelated. This latter decision was based on the notion that the scales in the final test would be able to be used independently in a differential prediciton model. In hindsight, it would have been useful to also estimate the model allowing the factors to be correlated as this less restrictive model would probably have yielded a better fit to the data and the utility of unciorrelated scales could have been considered in light of a more complete set of data.

As indicated previously, both the RMSR values and the item unique variances were examined to determine the fit of the model. The RMSR values were all well below Fraser's (1988) suggested data, and the examination of the unique variances indicated that the three factor model fit the data best of all.

References

Balla, J.R. & McDonald, R.P. (1985). Latent trait item analysis and facet theory - a useful combination. Applied Psychological Measurement, 9(2), 191-198.

Christoffersson, A. (1975). Factor Analysis of dichotomised variables. Psychometrika, 40(1), 5-32.

Fraser, C. (1988). NOHARM User's Guide. University of New England.

McDonald, R.P. (1967). Nonlinear factor analysis. Psychometric Monographs. Number 15.

McDonald, R.P. (1981). The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology, 34, 100-117.

McDonald, R.P. (1985). Factor analysis and related methods. Hillsdale, NJ: Lawrence Earlbaum Associates.

ANNEX G

A NOTE ON RELIABILITIES

It is important to note that the reliability coefficients presented in Table 5-2 of this paper represent coefficients calculated in the reduced set of items after the conduct of the factor analysis (i.e., they are based on a total of 26 items across three [or four] scales). these reliabilities then, do not reflect the reliabilities of the final test. In fact the entire project was aimed at determining the methodology to be used for developing the actual test and did not attempt to develop the test as such.

The aim of presenting the Spearman-Brown values in this table was to be able to compare the internal consistency of the scales from the different models. Because the four-factor scales were subsets of the three-factor scales, it was necessary to adjust their reliabilities to account for these differences in scale length.

The final test is likely to have between 60 and 75 item as opposed to the 26 items discussed in this thesis. This would yield significant improvements in the reliabilities of the scales over those presented on table 5-2. If the final test has 75 items (i.e., three times the number of items used when the values in Table 5-2 were calculated), and given the exisiting reliabilities, the reliabilities for the three scales would show the increases given in Table B-1 below.

Table G-1: Estimated Reliabilities of Final Test

Factor	Reliability	
	Original	Final Test
Verbal	0.674	0.861
Spatial	0.610	0.826
Numerical	0.726	0.888

These Final Test reliabilities can be considered reasonable for three scales within a 75 item test and would provide a sound base for the type of personnel decisions made in the selection model used by the Army.