"COMPUTER-BASED SELECTION TESTS: PSYCHOLOGICAL AND MEASUREMENT IMPLICATIONS OF ADAPTIVE TESTING"

By

OTHMAN ALKHADHER, BA, MSc. Psychology

Thesis Submitted to The University of Nottingham For The Degree of Doctor of Philosophy, September, 1994

ABSTRACT

The aim of this thesis is to develop realistic expectations about the psychological and psychometric implications of using computerized adaptive tests (CAT). A review is carried out of literature on computerized-based testing (CBT) and CAT. A field study as well as four laboratory experiments were conducted to achieve that goal. The current research strongly suggested the equivalence between the paper-and-pencil (P&P) and CAT formats for the Abstract Reasoning (AR) and Mechanical Reasoning (MR) tests of the Differential Aptitude Tests (DAT), but failed to do so for the Numerical Ability (NA) test. Also, the CAT version of DAT can predict a performance variable as accurately as can the P&P format. Overall, testees' attitudes toward several aspects of computerized testing were positive. The results confirmed the negative relationship between computer experience and computer anxiety. Moreover, knowledge of CAT behaviour negatively affected subjects' performance, but did not increase the level of their state anxiety. This suggested that a form of feedback acts during the adaptive test which has a negative effect on testees' performance and response time. This assumption was confirmed. Subjects spend a shorter time on the subsequent item after negative feedback (wrong) on the previous item than after positive feedback (right). It has been found that although the response time for answering an individual item was higher for CAT format than for P&P format, the CAT version of DAT resulted in a 20% reduction in completion time of the test. Also, the difficulty level of the initial items has a significant effect on testees' overall scores. The findings of this thesis suggest that CAT has numerous advantages and potential for improving the efficiency and accuracy of testing, and has potential areas of future contribution within personnel selection and assessment. This potential can be realized if proper consideration is made in designing, developing, and implementing these testing systems, and if professional standards are maintained by developers and users.

To my mother and father

ACKNOWLEDGEMENTS

No one works around three years preparing for his/her thesis and writes it alone, it only seems that way at times. In addition to all the many friends, colleagues, and students who have provided advice and motivation, there are a number of people who deserve specific mention. I would like to thank David Clarke and Neil Anderson, my supervisors, for all their advice and guidance over several years. This work was financially supported by Kuwait University to whom I am very grateful for their assistance and encouragement. I wish to acknowledge with thanks Selby MillSmith Ltd., Saville & Holdsworth Ltd., and ASE-NFER-NELSON, for the readiness and cooperation they showed in providing and making the necessary modifications to their software. Also, I appreciate the aid of Eamonn Ferguson and Peter Bibby for their advice concerning statistical techniques and the interpretation of results. Thanks go to Kuwait Petroleum Corporation for their support and enthusiasm in providing access opportunities to collect data for the field study. I would like to thank Rainer Kurz, from Saville & Holdsworth Ltd., for his invaluable help. The research participants have been very helpful and cannot be personally acknowledged, yet the spirit with which they shared their time and insight made this work possible. Finally, I would like to thank my wife Huda, my children Talal and Hamoud, and all my family in Kuwait; although they have not read a word of it (yet), they have provided the solace and joy that made it possible for me to complete this thesis. It is dedicated to them. God ... thank you for your help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

Today, psychological testing plays an important role in our lives. The psychological test is one of our most crucial assistant whenever information about individuals is needed to help make decisions about them. Whether in school, college, a job, or clinic, our ability and personality are routinely assessed and monitored using psychological or educational tests for the purpose of selection, promotion, development, counselling, or diagnosis. Most of the other alternatives to psychological tests are either psychometrically less efficient, more expensive, or not widely applicable. There is now a growing awareness in society and organizations about the importance of the objectivity and validity of the measurement used for human assessment. The psychological test emerges as one of the strongest candidates that can fulfil these demands by measuring human ability and personality more objectively in the most valid and least discriminatory and expensive way.

Good use of occupational tests, one type of psychological test used in the world of work, helps reduce waste in training and maximize the chance of a good matching between an individual and his/her job; the result is a rise in productivity and individual satisfaction. Occupational tests are used for all types and levels of job selection: low-level employees, professional or mid-level jobs, and managerial testing. They comprise many types: cognitive ability, aptitude, interests, and personality tests.

Until twenty-five years or so ago, most psychological tests were administered using paper and pencil. Paper-and-pencil administration suffers from a number of shortcomings. Hakel (1986) summarized these shortcomings as: "*excessive administration time, poor differentiation among people of extreme ability, limited capacity for measuring some types of abilities (target identification and tracking), cumbersome and error-prone scoring, expensive and time-consuming replacement, and high vulnerability to theft and compromise"*. With paper-and-pencil administration, all testees are confronted with the same test items. Some of these items are too easy, others are too difficult, and few items match the testee's ability. Items which are too difficult encourage random guessing and omission, and increase testee's anxiety. Items which are too easy are boring and decease testee's motivation. Neither items which are too easy nor those which are too difficult are suitably informative about testee's ability.

The advent of cheap portable computers coupled with Item Response Theory has led to one of the most exciting areas of applied measurement, and made the solution to the above problems possible. Today, computers can carry out several functions in testing processes, such as selecting which test to administer, presenting instructions, administering and scoring tests, and the analysis and interpretation of results, beside storing data and controlling peripheral equipment such as videotapes. Computers reduce the cost and time of testing, enhance the standardization in test administration, conduct more complex scoring and analysis procedures, provide reports about testees' performance and personality, conduct tests with moving stimuli, record response latencies, and reduce the direct involvement of the examiner in the actual testing situation.

However, a computerized copy of paper-and-pencil tests does not improve efficiency, advance psychometric properties (Weiss & Vale, 1987), or solve the problem of poor differentiation among people of extreme ability. What is needed is a test which is able to tailor its difficulty to the testee's ability and presents for him/her only those items which enhance our knowledge about his/her ability. This method of testing, which came to be called computerized adaptive testing (CAT), was difficult to develop without computer technology. Unlike paper-and-pencil tests, CAT items are selected during rather than before administration. One of the most important advantages of using CAT is its ability to provide more precise ability estimates with fewer items than P&P tests. Uninformative items are eliminated, and only those which provide further information about the testee's ability are presented. The direct advantage is a reduction in test length of around 50% without sacrificing measurement quality. CAT based on IRT also measures well at any point on the trait continuum especially the ability levels of those of high or low ability. Other researchers found that CAT ensures more security for test results and items, and helps reduce test anxiety by presenting items which challenge but do not discourage the testee.

In spite of its promising role in occupational psychology, most of the work in the area of computerized (adaptive) testing has been done in fields such as counselling, educational, or clinical psychology. Therefore, the purpose of this thesis is to review developments in computerized (adaptive) testing particularly in personnel selection, and to highlight the important aspects and issues in the use of CATs for assessment and selection, benefiting from work which has been carried out in other areas but which is still relevant to the occupational field. The objective of the work described in the thesis is to assess the psychometric as well as the psychological benefits and limitations of adaptive testing using both real and simulated data, and to develop realistic expectations about the implications of using CAT. Specifically, the thesis investigates the issue of equivalence between P&P and CAT formats, the predicted validity for both formats, the time taken to complete both formats, the effect of feedback on time taken to answer an item, computer anxiety and testee's reactions and attitudes towards computer adaptive testing, the effects of previous computer experience and prior test experience on performance and ability to identify CAT, the effects of feedback and knowing about CAT behaviour on testee's scores and anxiety, and finally, the differential effects of item difficulty arrangement on performance in adaptive tests. To achieve these goals, two introductory literature reviews about computerized testing and adaptive testing were conducted. Also, one field study and four laboratory experiments were undertaken as empirical work for this thesis.

The thesis is organized into seven chapters. It begins with a review of recent developments of computer-based testing in research and practice in Chapter 1. Based on this introductory chapter, a paper by the author, Anderson,N. and Clarke D. has been accepted for publication in The European Work and Organizational Psychologist. Chapter 2 discusses adaptive testing and its applications for selection and assessment. The two chapters that follow, Chapters 3 and 4, address the issues of equivalence, the predictive validity of the paper-and-pencil and computerized adaptive formats, testing time and testees' reaction and anxiety to computerized adaptive testing for selection purposes. Chapter 5 assesses the effects of knowledge about adaptive tests on subjects' performance, and the effects of immediate knowledge of results on the testee's score, anxiety, and answering time. Based on this chapter, a paper by the author, Clarke D, and Anderson,N, has been presented to The British Psychological Society, Occupational Psychology Conference (1994). Chapter 6 deals with the differential effects of item difficulty arrangement on performance. The thesis concludes with Chapter 7, which summarizes the key findings and discusses the implications of the results.

Edward Tolman once said that in the end, the only sure criterion is to have fun. Writing this thesis has been a labour of love, and I hope that the thrill of research discovery and justification I have experienced in my work is reflected in the following chapters.

CHAPTER 1

CHAPTER 1

Computer-Based Testing: A Review of Recent Developments in Research and Practice

1.1 Introduction

Over the last 30 years, computer technology has developed at an exponential rate. There are new advances in computer hardware and software virtually every week, and while computers have become progressively more powerful, they have become disproportionately less expensive. This revolution in computer technology has brought about rapid and important advances to many fields of applied psychology, including occupational, clinical, educational, experimental, counselling, and other areas of applied psychology. Perhaps one of the areas where most rapid advances have been made is in personnel selection and assessment (Alkhadher et al., in press), with several of the major test suppliers and occupational psychology consultancies developing computer-based systems (Krug, 1988; Stoloff & Couch, 1992; Trapp & Hammond, 1991). More generally, computers have been used in selection scenarios for conducting interviews, keeping applicant records, presenting check lists and rating scales, for self and psychological monitoring, and most widely, for test administration, scoring and analysis (Anderson and Shackleton, 1986; Schoenfeldt & Mendoza, 1991; Kratochwill et al., 1991). Aiken (1988) mentioned many factors which have contributed to this growth including the need for more effective methods of selection, placing and diagnosing of individuals, improvements in hardware, and the spread of social services.

Primarily, computers have been used to develop automated versions of paper-and-pencil (P&P) tests to assist in presenting and scoring them. Today, computers can carry out several functions in testing processes, such as selecting which test to administer, presenting instructions, administering and scoring tests, and the analysis and interpretation of results, beside storing data and controlling peripheral equipment such as videotapes (Baker, 1989; Bartram, 1989a). In fact, the growing use of computers in psychological testing represents only one aspect of the increasing use of computers in measurements in all types, such as for measuring physical strength and endurance (Murphy & Davidshofer, 1994).

In spite of the rapidly expanding application of computer-based tests (CBTs) in organisational selection, with some exceptions (Bartram & Bayliss, 1984; Bartram, 1994; Burke, 1993), most of the major reviews of developments in this area originate from sub-disciplines other than occupational psychology, such as counselling, educational, or clinical psychology (e.g. Bloom, 1992; Butcher, 1987; Bunderson et al., 1989; Hedlund & Vieweg, 1988; Nurius, 1990). It is therefore timely and appropriate to review advances in this area from an explicitly occupational-organisational perspective. Thus, the purpose of this chapter is to review developments in computer-based testing (CBT) particularly in personnel selection, and to highlight the important aspects and issues in the use of CBTs for assessment and selection, benefiting from work which has been carried out in other areas but which is still relevant to the occupational field. Seven crucial issues are examined:

(1) A brief historical review

(2) The potential advantages and limitations of CBTs

(3) The reliability and validity of CBTs

(4) The equivalence of conventional and computerized tests

(5) The acceptability of computer assessment

(6) The innovative use of computers

(7) Ethical issues concerning using CBTs.

1.2 A Brief Historical Review

The history of automated psychological tests began in the late 1960's with the onset of optical scanning for scoring examinees' responses. This was considered a significant advance in reducing both scoring time and errors (Kovac, 1989). The earliest attempts did not use the normal monitor and keyboard. Space (1981) noted that many of the early trials to automate the traditional paper-and-pencil tests were done by non-computer systems, applying different levels of automation. Gedye and Miller (1969) developed an automated system from a teaching machine called the *ts512* using a pictorial paired associate learning test for assessing their subjects. Brinton and Rouleau (1969) automated the Hidden and Embedded Figure tests, but their device was not fully automated, as it required a clinician as the supervisor to check the subject's response accuracy and to control other attached peripherals. The advantages that they noticed lay in the accuracy control and flexibility in manipulating a wide range of parameters.

Early attempts were also initiated by Elwood (1969, 1972) who automated the Wechsler Adult Intelligence Scale (WAIS) by designing multimedia presentations. In comparing the new results with results collected from traditional face-to-face administration using test-retest methods, Elwood and Griffin (1972) reported that generally high reliabilities were found between these modes of administration. Furthermore, Elwood (1972a) found a 50% reduction in administration costs. Unfortunately, this attempt failed to automate the scoring function.

Using a PDP-8 computer, Elithorn and Telford (1969) tried to assess intelligence by automating a multiple choice non-verbal route-finding perceptual maze task. Their attempt was to understand the perceptual process that different examinees adopt to solve the maze task using an oscilloscope and teletypewriter. In another attempt, Brierley (1971) automated Anstey's Dominoes which was designed as a parallel test to Raven's Matrices. He reported little, if anything, was lost in terms of reliability and validity, and that subject motivation was enhanced by this automated procedure, though he noted the unpleasant feeling that some subjects confront as a result of the relative degree of isolation and lack of proper feedback about test results, and the difficulties in observing and sparse information about anomalies in performance.

Attempts have also been made to automate the scoring and interpretation of MMPI by Swenson (1960) and Pearson et al. (1965). A number of investigations have since attempted to automate the same test. For example Dunn, Lushene, and O'Neil (1972) examined the response latencies of 77 college students using an automated version of MMPI. Also, Hansen, Johnston and Williams (1977) developed an on-line management information system for assessing mental health patients which is capable of utilizing the MMPI and a number of other psychological tests such as the Shipley-Hartford Test.

Overton and Scott (1972) have automated the Peabody Picture Vocabulary Test and found a high correlation with the paper-and-pencil version. However, higher rates of initial failures have been found, which they believed could be eliminated by proper instruction and training. Knights et al. (1973) compared these findings with the results obtained from Coloured Progressive Matrices and concluded that automated tests took more time but were faster to score.

Countless attempts have been made to develop automated computerized versions of existing P&P tests, mostly in areas other than the occupational field. For example, the perceptual Maze Test (Elithorn et al., 1963, 1982); The Modestly Automated Psychological Screening (MAPS) (Acker, 1980); The Eysenck Personality Inventory (Katz & Dalby, 1981); The Mill Hill Vocability Scale (Beaumont, 1980, 1987); These efforts show that the more feasible computerized psychological tests have a promising future.

These attempts have formed the basis for developing new computerized tests devoted to selection purposes. The application of CBT's in job selection have benefited from the commutative research carried out for assessing, for example, testee's acceptance of this kind of testing, the equivalence of computerized tests to its paper-and-pencil formats, the innovative uses of computers in assessment, and the numerous advantages gained by this method. For example, the MICROPAT system (Bartram & Dale, 1983) which was developed for the selection of Army Air Corps helicopter pilots contains seven tests designed to assess a range of psychological attributes such as perceptual-motor coordination and decision-making. Illiana Aviation Systems and Technical Solutions Inc. has also developed the Portabat to measure a range of information processing abilities and personality characteristics considered important in the selection of candidates for flight training (Telfer, 1985).

Computerized versions are now available for the most well-known tests (Krug, 1988), and most of the products of the leading American and European occupational consultancy companies have been computerized to aid in employee selection and assessment. In the U.K market today, many computerized tests have been developed by leading test development companies for selection and assessment, such as Saville & Holdworths Ltd. (SHL), NFER-NELSON and The Psychological Corporation, to assess different aspects of candidates' and employees' traits at work.

1.3 Potential Advantages and Limitations of CBTs

1.3.1 Advantages of CBT

As with most new technologies, there are advantages and limitations associated with computerized psychological tests. Undoubtedly, automated testing offers a number of advantages over P&P testing, although some of these deserve careful evaluation. Many reports have emphasised the advantages of computers, such as their ability to use adaptive strategies, ease of recording response latencies, savings in cost and time, providing immediate feedback about the examinees' performance, and the use of innovative forms of items. Computers also perform routine administration and scoring tasks objectively and accurately without the need for supervision by a trained psychologist. The following is a brief description of some of the advantages claimed by the developers of systems, testees, and respondents.

Microcomputers may offer special benefits by reducing the time of the assessment session compared to that of the traditional P&P method, hence making the selection process faster (Bunderson et al., 1989; Olsen et al., 1986). The saving in time has been found with both normal CBTs and computerized adaptive tests (CATs). The speed is attributed mainly to the speed of the automated scoring, interpretation, and report writing. Time can be reduced by around 50% using CATs without any loss on measurement precision. Thus, increased speed of administration and scoring of tests is combined with the elimination of clerical errors.

Computerized testing can be as cost-effective as traditional methods. Once a test is installed, it can be used repeatedly at little extra cost (Space, 1981; Traver, 1986). The saving in cost comes from replacing hand scoring and conventional materials used with P&P (such as booklets, pencils, answer sheets, erasers, and watches). Some researchers have found a reduction in cost by up to 50% for computerized testing compared to P&P assessment (Johnson & Williams, 1980). The costs of obtaining the equipment needed for computerized testing alone are still relatively high, but as microcomputers are used for additional management computing purposes such as word-processing and accounting, the marginal costs of testing by computer should become more favourable in the future. Using secretarial applicants for job selection purposes, Schmitt et al. (1993) showed that computerized testing is practical, cost-effective, and psychometrically sound. The costs of the machines are dropping sharply, and in any case, most computerized testing software available today does not require a high hardware specification to run.

Computerized testing offers promising opportunities for research , which was previously very difficult to conduct. For instance, it becomes easier to keep track of many aspects of the subject's behaviour, rather than simply listing the respondent's answers. For example, CBT permits the accurate recording of response latencies, either for an entire testing session or for a specific item. Other aspects of subject's test behaviour, such as the order in which items are answered, changing of answers and skipping of items, are easily recorded (Wise & Plake, 1989). This has useful implications in certain situations. Tests have been found to relate significantly to some aspects of behavioural patterns (Stout, 1981). For instance, if a long time was spent over a particular part of the instructions this could mean that the instructions were not clear and should be simplified (Tylor, 1983). If a delay frequently occurs on a certain type of item, there may be some psychological implications about the candidate's personality or the clarity of the items (Temple & Geisinger, 1990). Also, an extreme latency may indicate an invalid response or that a particular item had some emotional importance for the examinee. These implications could be used as possible measures for test validity, fatigue, or anxiety (Butcher et al., 1985), and also to provide details about both cognitive and non-cognitive aspects of performance (Weinman, 1982). Hakel (1986) believes that this feature can help us to a sharper conceptualisation of our measures.

It is also possible to reduce unintentional errors associated with coded responses. Computers can easily match between the question number and the answer space, eliminating any possibility of unintentional mismatches. There is no possibility of failing to completely erase or fill in an answer, skipping a question in the test booklet but not on the answer sheet, or misuse of optical scanners (Bunderson et al., 1989).

As computers are able to instruct the testee as well as score and analyse the respondent's results, the direct involvement of the examiner in the actual testing situation can be reduced. Many testees feel reluctant about asking the tester for help during the test but feel easier with responding to computer instructions. The computer is usually programmed not to move to the next piece of instruction or start the actual test until it ensures that the testee understands fully the instructions and responds correctly to the examples. This has two benefits: freeing the psychologists from routine tasks and enabling them to devote more time to the more complex issues involved in the assessment process; and adding more reliability and objectivity to the testing process (Butcher, 1987).

Computerized tests thus give flexibility and efficiency in manipulating items and responses and can provide immediate reports about testee performance during and after the test session. This is important for some situations, such as in college placement testing and for assessing large numbers of candidates. Some computer systems have been equipped to provide immediate interpretative reports about subject performance. Computer-based test interpretation (CBTI) reports, which are used mostly for personality questionnaires, can be seen as an aid for psychologists in their decision making and should be used in conjunction with professional judgements (Rolls & Harris, 1993; Gutkin & Wise, 1991). However, the validity of such reports is the area which is most attacked by critics (Fowler, 1985; Matarazzo, 1983). Skinner and Pakula (1986) point out that there have been no predictive validity studies published to justify their use in making clinical, educational or personnel decisions. However, recent study by Bartram (1994c) showed that the ICES personality test report have good discriminative validity. He pointed out that there are two main issues to consider when evaluating a CBTI system: first, the validity of the information generated by it, and the ways in which that information is likely to be used.

Other practical advances in CBT have recently been made. For example, Saville & Holdsworth Ltd. have developed a computerized assessment system which allows candidates to be assessed in their native tongue, and provides examiners with interpretative reports in any of the twelve European languages. Bartram (1989) reported that, in France, applicants are able to communicate through a MINITEL terminal system to fill in the application forms of certain companies. This has been found to increase both the response rate and the speed of the process of obtaining and pre-screening applications.

A number of other potential benefits have been noted; for instance, greater standardisation of the testing process, ensuring that every question that should be answered will be answered; considerably reducing the workload placed upon the tester so that the sample size can be increased, eliminating the need for collecting the test material (e.g. answer sheets or test booklets) and hence reducing the chance of test material loss or abuse, and ease of updating test materials. Other important advantages such as the use of adaptive methods, dynamic graphic material, and the provision of feedback are discussed later in the section on the innovative use of CBTs.

To summarise, the application of computer technology to personnel selection is a logical step because of the advantages mentioned. Existing evidence suggests that the advantages of CBT include its cost effectiveness, practicality, and increased reliability. Given these potential benefits, computers seem to have a promising role in behavioural assessment for selection and placement.

1.3.2 Limitations of CBT

As well as the potential benefits described above, computerized testing suffers from some important limitations. For P&P tests, large group administration is time and cost effective, whereas conducting computerized tests for a group of testees requires a larger number of microcomputers or terminals to be available with an operating system able to read the CBTs software. Unless there is a large number of testees and frequent use of these machines for assessment or any other function such as word-processing, this can become a costly process. However, small numbers of candidates can be scheduled over a number of days using one microcomputer to reduce potential cost ineffectiveness in CBT.

Concerns over testee privacy and the confidentiality of information stored in computers are frequently mentioned in discussions of ethical problems related to computer assessment (Meier & Geiger, 1986). The ability of the computer to manipulate, store, and retrieve information, magnifies the potential for invasion of privacy and abuse (Sampson, 1983). This issue will be discussed in more detail later in this chapter.

One final point is that taking a test via computer may be seen as less natural than using paper materials. In most cases, it is difficult to see more than one question on a computer screen, and to scan the whole test to determine the preferred strategy for answering or to get a feeling for the kind of questions to be answered. Physical complaints such as tiried eyes, headache, neck, back and hand pain, have also been reported because of sitting in front of a computer for long periods of time.

1.4 Reliability and Validity of CBTs

As with P&P tests, it is important that CBT maintain good reliability and validity. Wise and Plake (1990) found that the most common result is that the reliabilities of CBT and P&P tests are very similar. Greaud and Green (1986) found that tests administered on a computer were at least as reliable as conventionally administered tests. Beaumont and French

(1987) reported good reliabilities on the Mill Hill Vocational Scale, the Standard Progressive Matrices and all scales of the Eysenck Personality Questionnaire with the exception of the P scale, which seemed to be due entirely to the poor reliability of the data collected. Reliability on the Money Road Map Test was acceptable, but was poor in relation to the Differential Aptitude Tests. Administering both formats to 83 naval pilots and flight officers, Federico (1991) found that the CBT and P&P measures were not significantly different in reliability. This indicates that, in general, scores obtained from computerized tests appear to be at least as consistent as those obtained from P&P tests.

Researchers have found that computerized tests have an acceptable level of predictive validity. Schmitt et al. (1993) carried out a pilot study into computer-based testing procedures for the selection of secretarial applicants. They concluded that the CBT represented a more job-relevant and more face-valid approach to assessing clerical skills than traditional P&P tests or typing examinations. McHenry et al. (1990) administered a predictive battery of cognitive ability, perceptual-psychomotor ability, temperament/personality, interest, and job outcome preference measure, to 4039 enlisted soldiers in nine Army jobs. They found that the computerized cognitive testing battery predicted general soldiering proficiency as well as the conventional general cognitive ability composites. Burke (1984), using a sample of 217 clerical employees, found a multiple correlation of .63 between a computerized reasoning ability test and an overall job performance rating criterion for the general clerk job family. Also, Silver and Bennett (1987) using 34 secretaries, reported a correlation of .62 between the Minnesota Clerical Test and interactive word processor tasks, compared with a .55 correlation for the P&P format of the same tests. From these and other studies (e.g. Marshall-Meis et al., 1983) it appears that CBTs provide a predictive validity comparable to that obtained from P&P tests. The study of Federico (1991) showed that the relative discriminative validity of CBT and P&P measures was dependent on the specific statistical criteria selected. That is, the discriminant coefficients, F ratios, and corresponding means indicated that the validities of CBT and P&P measures were about the same for distinguishing groups above or below mean curriculum grade, but according to the pooled within-groups correlations between the discriminant function and CBT and P&P measures, the former had better validity than the latter. Perceptual/psychomotor and cognitive tasks in a computer-aided aptitude test were studied by Park and Lee (1992) to predict the success of a trainee in flight training. They concluded that a computer-aided battery of tests can provide a multidimensional methods for efficiently predicting the performance of pilot candidates.

Taking all these advantages and limitations into consideration, one can foresee that computer-based testing still possesses sufficient advantages to secure its future in personnel selection and assessment.

1.5 Equivalence of Conventional Form And CBT's

Due to the fact that most of the computerized tests currently available have been adapted from earlier P&P formats, the issue of whether results obtained from computerized tests are equivalent to those obtained from their traditional counterparts has become an important one. Investigation of test equivalence enables us to understand those factors which affect performance when conventional tests are converted to an automated format. It also enables norms, cutting scores and validity data from the traditional test form to be generalised to the computerized version (Hofer & Green, 1985). There is no guarantee that results obtained from these two forms will be parallel, even when the item content remains unchanged. Situational and technical factors which may affect equivalence are discussed later. The APA guidelines (1986) specified two conditions to establish the equivalence between two formats: *'(a) the rank orders of scores of individuals tested in alternative modes closely approximate each other, and (b) the means, dispersion, and shapes of the scores distributions are approximately the same, or have been made approximately the same by rescalling the scores from the computer mode'.* Burke (1993) pointed out that the issue of mean equivalence between test formats is not very important for most of the ability tests used by personnel psychologists, as long as the two modes assess the same construct. He argues that ensuring no change in testees' ranks between the two forms provides greater evidence for equivalence.

Research into the equivalence of two forms of the same test seems to be inconsistent (Federico, 1991). Concerning ability tests, some studies comparing results obtained from computerized tests with their P&P versions have found no significant differences (Beaumont, 1985a; Harrel et al., 1987; Huba, 1988; Rock & Nolen,1982; Wilson et al., 1982). For example, Rock and Nolen (1982) found that the two versions of the Raven's Coloured Progressive Matrices test do not differ significantly on the dimensions tested. Comparing the effects of a computerized administration procedure on the reliability and validity of verbal scales of the Multidimensional Aptitude Battery with standard P&P administration, Harrell et al., (1987), using 80 undergraduate subjects, found no significant differences in scores or anxiety across formats. However, subjects showed more positive feelings toward the computerized format. High degrees of equivalence between different modes of presentation are expected for unspeeded tests, textual items, tests with unchanged test length or item format, and for tests requiring some form of multiple-choice or forced-choice response (Burke & Normand, 1987).

On the other hand, some researchers reported significant differences, although the difference are quite small. For example, Beaumont and French (1987) administered eight different psychological tests and found that some tests were less amenable to computerisation. Kubinger et al. (1991) tested the equivalence of the conventional version of the Standard Progressive Matrices -German version- with its computerized format using various subject samples including federal employees. They found that an item bias appeared between the two forms and that the computerized form led to score differences of up to 13 IQ points lower. Neubauer et al. (1991) assessed the German version of Raven's Advanced Progressive Matrices. They found scores in the computer format were significantly lower than for the standard version. This was partly because the test was completed more rapidly, and several items were more difficult than in the conventional format. They found that the correlations between forms ranged from .70 to .82. Lee et al. (1986) used the Arithmetic Reasoning Subtest of the ASVAB and found a significant main effect between the two modes. Kovac (1989) used 121 clerical job applicants (98 females and 23 males) to study the effects of administration mode and test type on both test scores and response times. Half of the sample were given P&P tests and the other half computerized tests (numerical and verbal reasoning). It was found that computerized administration slowed applicants down, resulting in lower scores for the computer applicants, although percentage correct did not differ significantly across the modes. Mazzeo and Harvey (1988) discussed results obtained from different equivalence studies and concluded that equivalence must not be assumed for these two modes of presentation. As Lee et al. (1986) noticed, the inconsistency of these results may be due to differences in methodology, test content, population tested, or the design of the study.

Concerning personality tests, most studies of automation have not reported significant differences between the two forms (e.g. Katz & Dalby, 1981; Gitzinger, 1990; Ridgway et al, 1982; Schoonman, 1989). For example, Gitzinger (1990) found no differences between the different versions on a German-version defence mechanisms test. Also, Fekken & Holden

(1989), using the Personality Research Form, found that the P&P form was comparable to the computerized version. Honaker et al. (1988) examined the equivalence of the Microtest computer format of the MMPI with its P&P format using 80 subjects. They found no significant differences between the two formats in terms of means and standard deviations. Also, the rank-order of scores and the corresponding values reliability for the computerized format were similar to the test-retest correlations for the P&P format. On the other hand, Schuldberg (1990) found that subjects, when they moved from P&P to CBT formats, tended to change their responses on the MMPI to 'cannot say', and to shift from false to true on second administration. Davis and Cowles (1989) found that subjects responding to a computerized questionnaire measuring anxiety and locus of control seemed to give responses indicating lower levels of test anxiety than they gave to P&P format of the test (fake good). Allred (1986) faced similar problem and found a difference in the scores of the two modes.

A number of factors have been identified which may influence test performance when administered by microcomputer systems. Some studies have shown that previous computer experience significantly influences computerized test performance (Lee, 1986; Cornwell et al., 1993). Llabre et al. (1987) conducted a study to determine the effect of computer administered tests on test anxiety and performance, and concluded that computer administered tests can potentially increase test anxiety and depress test performance for examinees who are relatively unfamiliar with computers. Another noteworthy finding was that an individual's attitude towards computers is highly influenced by previous experience with computers (Burke et al., 1987). However, Dimock & Cormier (1991), using DAT, found no evidence that difference in performance was influenced by either the testee's level of computer experience or anxiety.

The advantages of past experience may introduce a bias against subjects who have had limited access to computers and have not developed the required skills. Hofer (1985) warned that "because the advantages of computer technology are distributed unevenly, a modern version of cultural bias may be alleged, that groups lacking in computer experience will be disadvantaged if forced to take tests on computer". One study by Johnson and Mihal (1973) reported that black people performed relatively better on one computerized intelligence test (Co-operative School and College Ability Test) than on its conventional version, compared with white individuals. Available evidence suggests that prior training in computer applications can reduce the anxiety caused by computers and hence improve the testee's performance. Organisations should ensure that sufficient training and examples are provided before the actual testing session commences.

Other factors which may affect test performance relate more directly to the test used. It has been found that respondents to computerized tests take less time to complete the session than the traditional versions (Henly et al., 1989). This difference could be attributed to the ease with which testees can respond to computerized items, for example by pressing the appropriate key or clicking the mouse button, compared with finding and blacking the right box on the answer sheet using a pencil. Greaud and Green (1986) used simple arithmetic questions and reported large differences in total scores on speeded tests between the two modes (3:2 respectively). Agapitou (1993) found no strong evidence of equivalence between the computerized and P&P versions of the CP7.1 test (speed and accuracy in checking non-contextual material). It seems that the problems arise from speed rather than power tests. This suggests that speed test developers or users should consider specifying shorter times for administering computerized tests than their P&P versions in order to avoid any advantages that the extra time may yield.

Another explanation for the differences between the two modes could be attributed to the format differences in recording responses. It is likely that allowing testees to skip and review previous items on P&P tests and make changes to their answers, while restricting these features with computer-based tests, will affect test scores. Lunz et. al. (1992) used a computerized adaptive test in their study, and found that approximately 32% of the testees improved their estimates after review, but did not change their pass/fail status. They concluded that the importance of disallowing item review was not supported by their study. Similar findings have been reported by Harvey (1987), however, he found no statistically significant differences. Moreover, computers usually present items individually, preventing the testee from making a quick scan over the whole test to choose a preferred response strategy unless a specific button is pushed each time a new item is presented, which could take a considerable time. Finally, it should be noted that APA guidelines (1986) encourage developers and users to allow testees to review their answers.

Other studies have found that the use of different response models may lead to different performances (Beaumont, 1985a; Levy & Barowsky, 1986). Beaumont (1985b) recorded response latencies during a continuous performance task using four standard microcomputer response media (a keyboard, a keypad, a light-pen, and a touch-screen), and reported significant differences between them. The fastest response was made with the touch-screen, while the slowest was with the light-pen. He also found that the keypad was superior to the full keyboard, and that the physical arrangement of the keys was shown to be the factor which most affected the speed of response. Studies are required to investigate how different groups of people interact with different response devices. Moreover, lack of typing skill may be an important factor which affects scores, especially for tests which require more than selecting and blacking the appropriate answer box.

To summarise, having reviewed this literature, it seems important that unless evidence of equivalence is empirically established, comparability should not be assumed.

1.6 The Acceptability of Computer-Based Testing

The acceptability to both subjects and professionals of computer-administered tests has been addressed by many researchers (e.g. Bresolin, 1984; Erdman, Klein & Greist, 1985; Burke et al., 1987). Both positive and negative findings have been reported (Weinberg & English, 1983; Rosen et al., 1987; Moe & Johnson, 1988; Garrison & Baumgarten, 1986). It is difficult to apply advanced technology to different uses in psychological assessment without gaining the acceptance of subjects and testers, since negative reactions may influence the data obtained from subjects and hence the validity of the measurement used (Harvey, 1987).

Meir and Lambert (1991) reviewed the research in this domain and concluded that negative psychological reactions slow both the acceptance and the useful application of computers (Hofer & Green, 1985; Rosen et al., 1987). Martin and Nagao (1989) found negative reactions to computerized interviewing among applicants applying for high status managerial jobs.

Other research indicates that subjects tend to express positive reactions toward computer based administration (e.g. Gitzinger, 1990; Calvert & Waterfall, 1982; Burke et al., 1987; Skinner & Allen, 1983). Lukin et al. (1985) asked their subjects to indicate a preference between computer and traditional administration formats, and found that 84% of the subjects preferred computer administration, because the computer was "more fun", "different", "faster", or because the P&P test was "too much like school work". Mathisen et al. (1985) also reported several studies where subjects reported a positive feeling toward computers. Another study reported that candidates showed higher test motivation with computerized tests than with P&P tests (Arvey et al., 1990). Davis and Cowles (1989) reported a strong tendency for individuals using the computerized format to give responses related to low trait anxiety and "internality". The general conclusion is that most testees like being tested by a computer and provide more or as much personal information as in a P&P test.

Interestingly, other research suggested that practitioners have been more resistant to the use of CBT than have their testees (Burkead & Sampson, 1985; Dutro, 1983; Byrnes & Johnson, 1981; Hedlund et al., 1980). Johnson and Williams (1980) found testees' responses to be strongly favourable to computer administration, but staff felt more neutral and even somewhat negative. Fowler (1985) reviewed a number of studies and concluded that staff unfamiliar with computers tended to believe that their subjects would not agree to interact with the computer. Sampson (1983) and Byrnes and Johnson (1981) attributed this staff rejection to the lack of organisational readiness and planned strategy for change, as well as the lack of an approach to overcoming staff resistance. Other possible reasons are fear of being replaced by a computer, or perceiving the computer as a threat to staff control over selection decision making, and avoiding legal responsibility for using this new technology

(Nurius, 1990). Some suggestions for solving this problem have been offered. For example, after staff had attended a two and a half day seminar on computer use, Klonoff and Clark (1975) reported positive attitudes toward computers. Education and preparation, as well as implementing a planned change strategy to reduce staff threat and uncertainty toward computers, may be useful for successful change toward accepting computer testing systems (Byrnes, & Johnson, 1981).

Individual differences appear to be related to computer acceptance. Nurius (1990) reviewed several studies and found that unfamiliarity with computers is likely to be associated with certain groups such as women, ethnic minorities, or older, disabled, or economically disadvantaged individuals. Brosnan and Davidson (1994) reviewed the research literature and concluded that there is a strong evidence supporting differences in relation to the phenomenon known as computer-phobia, where females are generally more likely to posses higher levels of computer anxiety than males and to experience more negative attitudes towards computers. Wagman (1983) found that men had more favourable attitudes toward computers than women. Jay (1985), perhaps not surprisingly, noted that testees with more experience and those who had physically touched a computer held more positive attitudes. Others have reported significantly higher levels of anxiety among elderly testees and among students with certain academic majors (Rosen et al., 1987). Pocius (1991) concluded in his review that introversion-extroversion and traits characterising introversion-extroversion are related to many aspects of human-computer interaction.

There is a need for a better understanding of the attitudes of testees and users in the development of more friendly and efficient software for psychological assessment. Schmitt et al.(1993) measured the reactions of 47 applicants for secretarial jobs toward the new computerized selection procedure. They concluded that applicants will respond much more favourably when they believe they are being evaluated using materials that appear relevant to the job for which they are applying. However, perhaps concerns over testees' acceptance of this method of assessment will fade, as new generations become gradually more exposed to computers.

1.7 Innovative Uses of Computers in Assessment

The computer provides an opportunity for the development of entirely new forms of testing methods (Butcher et al., 1985). Its unique capabilities include the use of adaptive methods, dynamic graphic material, and providing feedback to testees both during and after the test session.

1.7.1 Adaptive Testing

The advent of computers coupled with applications of the Item Response Theory

(Lord, 1980) has made dramatic changes in test administration feasible. With few exceptions, administering conventional tests necessitates that all items should be administered to each testee, irrespective of whether they are too easy or too difficult for them. However, in the adaptive ('tailored' or 'branched') test, testees are presented with different test items according to their level of ability as indicated by their performance so far. Clearly, this can enhance the information obtained from a test. In practice, adaptive testing usually proceeds as follows. Unless some estimation of the testee's ability level is available, an item of moderate difficulty is presented first. If the answer is correct, the next item will be more difficult, but if the answer is wrong, an easier item will be presented. Progressively, the presentation of additional items will be based on the response of the testee to the previous items. This tailored testing strategy continues so that only items contributing meaningful information about the testee will be presented until a certain level of criterion is obtained, which could be a certain level of measurement precision, a desired number of items, or a defined time (Weiss, 1985; Weiss & Vale, 1987).

Substantial advantages have been obtained from implementing adaptive strategies instead of conventional, non-adaptive tests. Testing time can be reduced by 25% to 75% without any loss in measurement precision (Olsen, 1990; Weiss, 1985; Moreno et al., 1984). As a result of eliminating items which are too difficult or too easy, discouragement and boredom can be reduced (Rocklin & O'Donnel, 1987). Moreover, Weiss (1985) reported studies by Kiely et al. (1983) and McBride and Martin (1983), where they compared the reliability and validity of different types of tailored tests with their traditional counterparts and found that the reduction in test length produced by administering adaptive test does not sacrifice either the reliability or validity of the test. (See also Cudeck, 1985; Kent & Albanese, 1987; Koch et al., 1990).

In this view, Weiss (1973) developed a type of adaptive test called 'stradaptive testing' (stratified adaptive testing), in which items are divided into sequences of subsets based on their difficulty level. According to the responses, the testee branches up or down through the strata until they reach a ceiling stratum, where a certain test length or fixed time is reached.

New attempts have also been made to allow the testee to adapt the level of test difficulty in the way he/she prefers. Studies found that the 'self-adapted' test leads to higher ability scores and estimates and minimises the effects of test anxiety, without any overall loss of measurement precision compared with computerized adaptive and traditional tests (Rocklin & O'Donnell, 1987; Wise et al., 1992). Because of the promising role of CAT in human assessment, the next chapter will be devoted mainly to more details and discussion about this innovative technique.

1.7.2 Dynamic Material

The P&P form of a test tends to be static, whereas the computer has the ability to produce dynamic and more realistic material involve movement, color, speech, sound, and interactive graphics. More complex graphic displays can thus be presented using a video disk, or a film controlled by the computer. For a mechanical comprehension test item, for example, a clear and simple dynamic motion can be depicted to help the testee to understand the particular property of the motion involved. This limits the possibility of confounding the understanding of the item with other examinee skills such as reading ability. This should improve the validity of the test (Wise & Plake, 1989).

Hunt et al. (1988) compared subjects' performance on a battery of 10 computerized tests of spatial ability requiring reasoning about static or dynamic spatial displays. His study indicated that ability to reason about dynamic visual motion was distinct from the ability to reason about static displays. Also, computer-administered static reasoning tasks can be used to replace current P&P procedures for assessing spatial ability. He called for research to investigate the predictive validity of such dynamic tests in real-life settings. A practical example of this strategy is the Micropat system (Bartram & Dale, 1983) which contains dynamic test items tapping cognitive skills which are critical for the selection of helicopter pilots. Illiana Aviation Systems and Technical Solutions Inc. has also developed the Portabat to measure a range of information processing abilities and personality characteristics considered important in the selection of candidates for flight training (Telfer, 1985). Burke (1993) discussed the potential capability of computers to develop new item types and dynamic test stimuli for use in job selection.

1.7.3 Providing Feedback

Computers can also provide immediate feedback to both the user and testee regarding performance. Computers make it possible to employ different assessment strategies for feedback, by building an intelligent operating system, via well researched principles of Artificial Intelligence (AI) into the computer person interface. For instance, feedback can be given during the test session itself about the correctness of each item response, about the total number of items remaining, or about how much time is remaining. However, although providing feedback to testees is easy to do, its usefulness to them remains unclear. Wise and Plake (1989) reviewed research on the effect of item feedback on examinees and found inconclusive results. For example, Morris and Fulmer (1976) and Rocklin and Thompson (1985) found positive effects through lower test anxiety and higher test scores. Conversely, other studies found an increase in anxiety level and a decrease in test scores (Strang & Rust, 1973; Wise et al., 1986). By dividing subjects into high and low ability groups, Betz (1977) found that knowledge of results during a session improved the performance of a high ability group, but no significant differences were found for the control group. Although providing feedback to testees is an easy task to do, its usefulness to them remains unclear.

Such innovative strategies in candidate assessment would have been difficult, if not impossible, without computers. Bartram (1994) reports a number of new computerized tests as examples of novel uses of this technology in selection and assessment. More advanced hardware and software are expected to be produced in the near future, which hold out a great deal of promise for applicant assessment and placement. However, psychological measurement theories need to be reviewed and updated in order to take full advantage of these sophisticated machines.

1.8 Ethical Issues Concerning Computer-Based Tests

Serious concerns have been expressed that various professional groups are using computerized testing without adequate training. For example, Eyde & Kowal (1987) and Matarazzo (1983) addressed the possibility that computerized tests were being used by unqualified professionals who may be not aware of the limitations of such technology. In part, it is the responsibility of publishers to ensure that computerized psychological tests are not sold to unqualified buyers. However, sufficient training in CBTs (for example, measurement principles, test limitations, basic computer literacy, test experience with using specific applications in CBT for selection and assessment, and test interpretation limitations) may be useful to reduce the blind acceptance of software and to limit such problems (Meier & Gieger, 1986).

There is also a concern over the confidentiality and privacy of testee data when a test is conducted by computer. Technological advances make it cost-effective to keep larger amounts of data about individuals for longer periods of time compared with conventional documentary records (Gambrell & Sandfield, 1979). This increases the possibility of data abuse against identifiable individuals. Some computer systems even make it possible to gather physiological data without the testee's permission while he or she is interacting with the computer, and this is clearly a matter which may violate the ethical principles of psychological research (Sampson & Pyle, 1983). In the UK, the Data Protection Act (1984) was enacted to protect individuals against possible misuse of information about them kept in a computer, and to reduce the threat to individual privacy. Accordingly, a number of British employers have introduced policies and codes of practice designed to enhance the security of employee data stored in computers (e.g. British Petroleum Company plc. and Birmingham Metropolitan District Council).

As a step to safeguard employee privacy, Evans (1984) reports that IBM "has discontinued the use of personality and intelligence tests because these may constitute an unwarranted intrusion into an applicant's emotional or private life. Aptitude tests are still used for certain types of occupation". However, there is a belief that the confidentiality of testee results has less chance of being violated if the information is stored in a secured computer file (Burke & Normand, 1987) and a password is utilised to limit access to confidential information (Bunderson et al., 1989; Walker & Myrick, 1985).

As with P&P assessment, the psychometric properties of computerized tests, such as test reliability and validity, normative group data, test population, scoring procedures, decision rules, and interpretative statements, should be evaluated and made available to users to give them the full chance to select those tests which are psychometrically and ethically sound.

1.9 Conclusion

This review of recent developments in CBT has shown the numerous advantages and potential areas of future contribution for CBT within personnel selection and assessment. CBT has the potential for improving the efficiency and accuracy of testing, while simultaneously decreasing the costs of selection. Some of the limitations mentioned are likely to disappear in the near future with the advent of cheaper, more advanced and sophisticated machines. These benefits are sufficient to ensure an important role for computers in assessment and selection. More sophisticated and efficient adaptive tests, dynamic forms of presentation, high resolution graphical displays, speech analyses, valid interpretative computerized reports and free response items are expected in the near future. In addition, Artificial Intelligence is expected to contribute more to the area of testing; for example, in designing new testing instruments and items. It is perhaps this area of innovation which holds the greatest promise for future improvements in CBT procedures (O'Neil & Baker, 1991). It is therefore ironic that at the present stage of development, most CBT applications are parallel-form versions of existing P&P tests. Future research should surely address the as yet under-exploited potential of computers for presenting dynamic and high resolution graphics in an adaptive manner to testees.

In conclusion, CBTs are here to stay. The professional community of Work and Organisational psychology is charged with the responsibility of ensuring that CBT applications, as well as being valid and reliable, are used by selection practitioners in industry in ways which are fair and justifiable and do not encroach upon the personal privacy of applicants, nor contravene acceptable standards of ethical usage in other ways. Because of the promising role of computerized adaptive testing in selection and assessment, the remaining part of this thesis will explore in more depth this method of testing and its psychological and measurement implications, to form more realistic expectations about CAT.

CHAPTER 2

CHAPTER 2

Adaptive Testing and its Applications for Selection and Assessment

2.1 <u>Introduction</u>

For a long time, the 'traditional test theory' (Gullicksen, 1950) used with P&P tests was the dominant theory in the human cognitive assessment literature. Under this theory, all testees have two scores, a 'true score', which cannot be measured directly, and an 'observed score', which is usually contaminated with some amount of error. The theory requires that all testees take the same items under identical time limits and conditions. In a situation which requires the measurement of a wide range of ability levels, such as job selection, the test should contain as broad a range of item difficulties as the ability range of the subjects to be tested.

In this situation, the test developer usually has to choose between two options, or to combine the two to some degree (Weiss, 1985; Weiss & Yoes, 1991). The first option is to select items concentrated around the same level of difficulty, usually in the middle of the trait level of the population selected. This kind of test is called a 'Peaked Conventional Test'. A few very easy and a few very difficult items are left on both sides of the trait continuum. Because low ability testees will find most of the test items very difficult, and may become frustrated, and high ability testees will find most of them very easy, and probably become bored, this type of test provides limited information about their trait levels. Also, as the accuracy with which a test assesses at any point on the ability continuum is (roughly) proportional to the number of items with difficulties matching that level (Wainer, 1990), this kind of test will measure most precisely those individuals whose trait levels are at or near that difficulty level, but it will measure poorly for those with trait levels further away. The situation becomes more problematic when the test is presented to a population with trait levels concentrated at different point, such as the presentation of a relatively easy test to high ability candidates.

The second option for the test developer is to select several items at all levels of the trait to be measured. This kind of test is called a 'Rectangular Conventional Test'. In this case, only a few of the items in each test will be suitable for testees at a given ability level. The test will be able to differentiate between testees with different ability levels, but with an overall lower level of precision (Weiss, 1985; Weiss & Yoes, 1991) provide a good description of the limitations of traditional test theory). The ideal solution, then, is to present those items with difficulty levels matching the testee's own ability level, that is peaked about his/her ability level.

Binet (1909) designed the first standardized test, later known as the Stanford-Binet intelligence test (Terman & Merrill, 1960). According to this test, the tester, based on relevant information about the testee, usually his/her age, selects the starting difficulty level. Each item is scored immediately after being administered. The tester tries to identify the level at which all items are answered correctly (the basal level) and the level at which all items are answered incorrectly (the ceiling level). The assumption is that all items below the basal level would have been answered correctly if administered, and similarly all items above the ceiling level would have been answered incorrectly if administered. Once the basal and ceiling levels are identified the test is terminated. It is then scored by adding a specified number of months for each correct item to the year designation for the basal level. This method of adaptive testing requires a tester to administer the test for each individual testee, which may be costly and unpractical when large numbers of people are to be tested. Moreover, the items at the ceiling and basal levels are too difficult or too easy, and therefore they are not very informative.

For these reasons, there was a need to adapt the test to the examinee on a larger scale, rather than individually. The first attempt at this was made after almost fifty years by Fred Lord (1980), who began a long and comprehensive research program in the late 1960s to deal with these problems. Since then there has been increased interest in adaptive testing (see Wainer, 1990; Weiss, 1983), with much of the work being carried out under the sponsorship of military research organizations (Wiskoff & Schratz, 1989).

The advent of computers coupled with Item Response Theory (IRT) (Hambleton, 1989; Hulin et al, 1985) made the solution to the above problems possible. Computers try to do what a wise tester would do. The most appropriate items for the test (those with a fifty percent chance of a correct response) are presented, whereas those providing limited information about the testee's ability are eliminated. Far too difficult and far too easy items are generally avoided, and only items that match the difficulty level of the testee are presented. In other words, the test tries to adapt (or peak) its difficulty to the testee's level of ability, without any intervention from the examiner. This method of testing, which is based on IRT, has been referred to as Computerized Adaptive Testing (CAT). CAT can provide measurements of equal precision at all points on the ability continuum. It also provides more information at the ability extremes than P&P tests, with adequate information at average ability (Hambleton, 1991).

A number of large scale testing programs now use adaptive testing (McBride et al., 1987; Hsu & Shermis, 1987), and a number of commercial software systems are now available for developing adaptive tests (e.g. MicroCAT: Assessment Systems Corporation, 1988). CAT versions are now available for many well-known P&P tests, such as the Armed Services Vocational Aptitude Battery (ASVAB), Scholastic Aptitude Tests (SAT), Differential Aptitude Tests (DAT), California Achievement Tests (CAT), College Board Adaptive Placement Tests, Stanford Achievement Tests, the Woodcock-Johnson Psycho-Educational Battery, and the Army's Computerized Adaptive Screening Test (CAST).

2.2 Adaptive Strategies

Although adaptive testing generally refers to IRT-based CAT, other kinds of adaptive tests are still used. They can be classified into three categories according to the adaptive strategy employed: item-by-item adaptive strategy; subtest-by-subtest adaptive strategy; and model adaptive strategy. Vale (1981) and Weiss (1985) provide good descriptions of each of these strategies. The following sections briefly describe each strategy.

2.2.1 Item-by-Item Adaptive Strategy

The item pool is structured so that each item is tied with two other items, one more difficult and one easier than that item. Correct response lead to the pre-specified more difficult item, and incorrect response to pre-specified easier one. Examples of this type are Pyramidal (Figure 2.1) and Robbins-Monro (Figure 2.2) strategies. As shown in these two figures (from Vale (1981)) there are a number of items at each difficulty level, and all testees start with the same medium difficulty item.

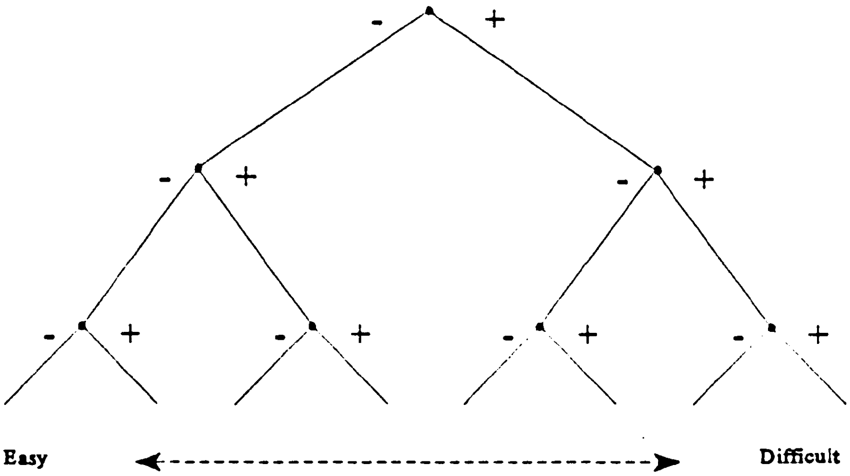

Figure 2.1. A pyramidal testing strategy.

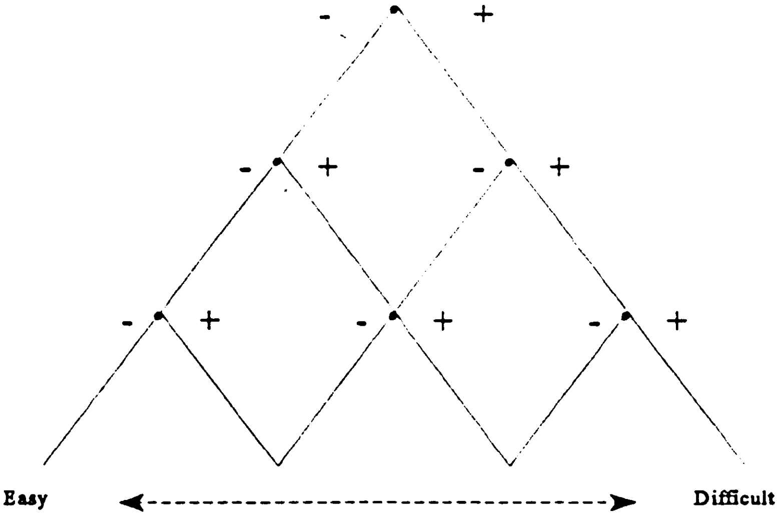

Figure 2.2. A Robbins-Monro process.

2.2.2 Subtest-by-Subtest Adaptive Strategy

Instead of branching to one item, this strategy branches to sets of items, called subtests. The test starts with a routing or locator subtest, which is usually of medium difficulty. Depending upon the testee's performance on this subtest, he/she is branched to another more or less difficult one. Once a testee has branched to another subtest, he/she cannot return to any previously administered subtest. If the test consists of two subtests (the locator and the subsequent subtest), it is said to be using a two-stage strategy (Figure 2.3 represents three subtests, from Murphy & Davidshofer, 1994). If it consists of more than two subtests (the locator and more than one subtest), it is referred to as a multi-stage strategy.

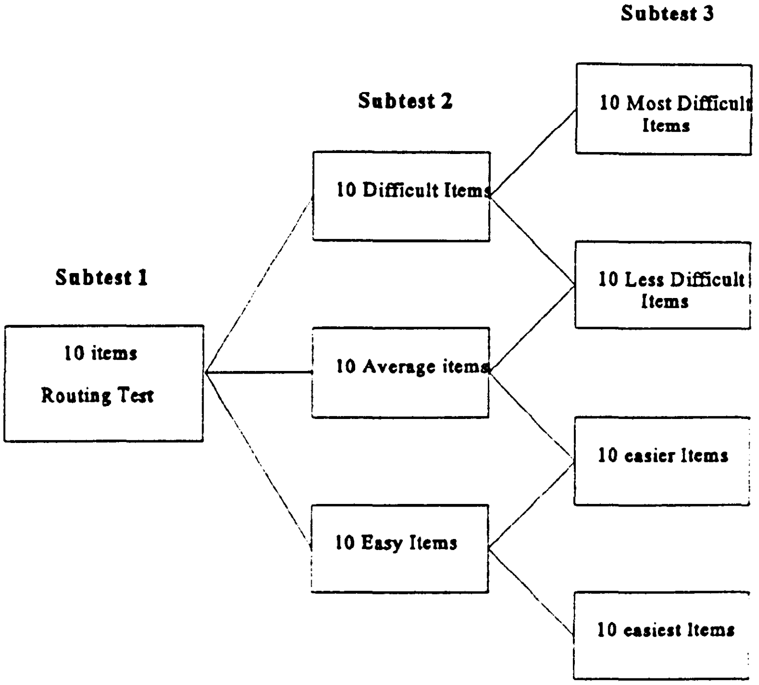

Figure 2.3. A three-stage testing strategy.

A revised version of this type is the strategy where the testee's ability is re-evaluated after each response to decide whether to remain in the same subtest or be shifted to another one. If required, the testee can return to a previous subtest, but will be presented with unadministered items. An example of this is the stradaptive (stratified adaptive) strategy (Weiss, 1973) mentioned in the previous chapter.

Another simpler form of this type is called the flexilevel strategy (Lord, 1971) (Figure 2.4). The test items are arranged on a continuum of equally spaced difficulty from very easy to very hard so that there is only one item for each difficulty level. Starting with an item in the middle of the continuum, the testee is branched to a more or less difficult item not previously administered, depending on his/her answer. The test stops after a predefined number of items.

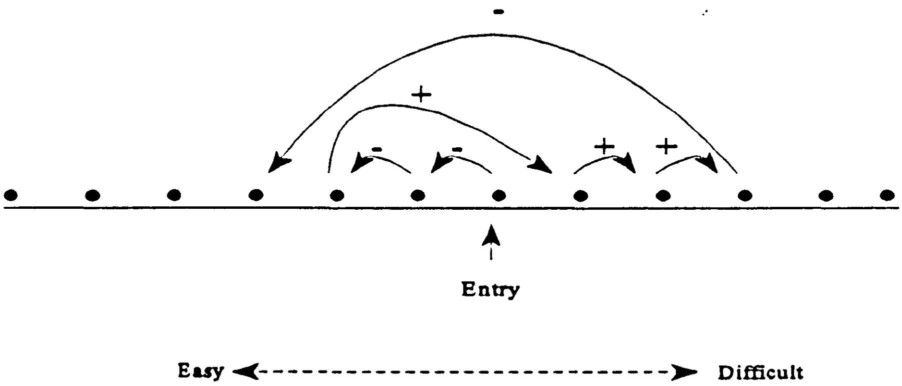

Figure 2.4. A flexilevel testing strategy.

The main limitation of the item-by-item and subtest-by-subtest adaptive strategies is that they use only item difficulty index in structuring the item pool, ignoring other useful item characteristics which are item discrimination and susceptibility to guessing indexes. The exception is the stradaptive test, which uses item discrimination indices beside item difficulty. Also, the scoring methods used for these strategies are somewhat arbitrary, without a theoretical or empirical base (Weiss, 1985).

2.2.3 Model Adaptive Strategy

The adaptability of this kind of test stems from the fact that only those items which are most informative about the testee's position on the ability or latent (unobservable) continuum are selected for administration, from an item pool with known item parameters. This strategy, which is usually based on IRT, assumes that the probability of getting an item correct is related to the testee's position on the latent trait or ability continuum. The test starts by presenting an item of average difficulty. After each answer, the trait estimate is updated and an item which is able to improve the estimate is selected (Figure 2.5 shows the structure of CAT, from Thissen & Mislevy (1990)).

Adaptive Test Logic

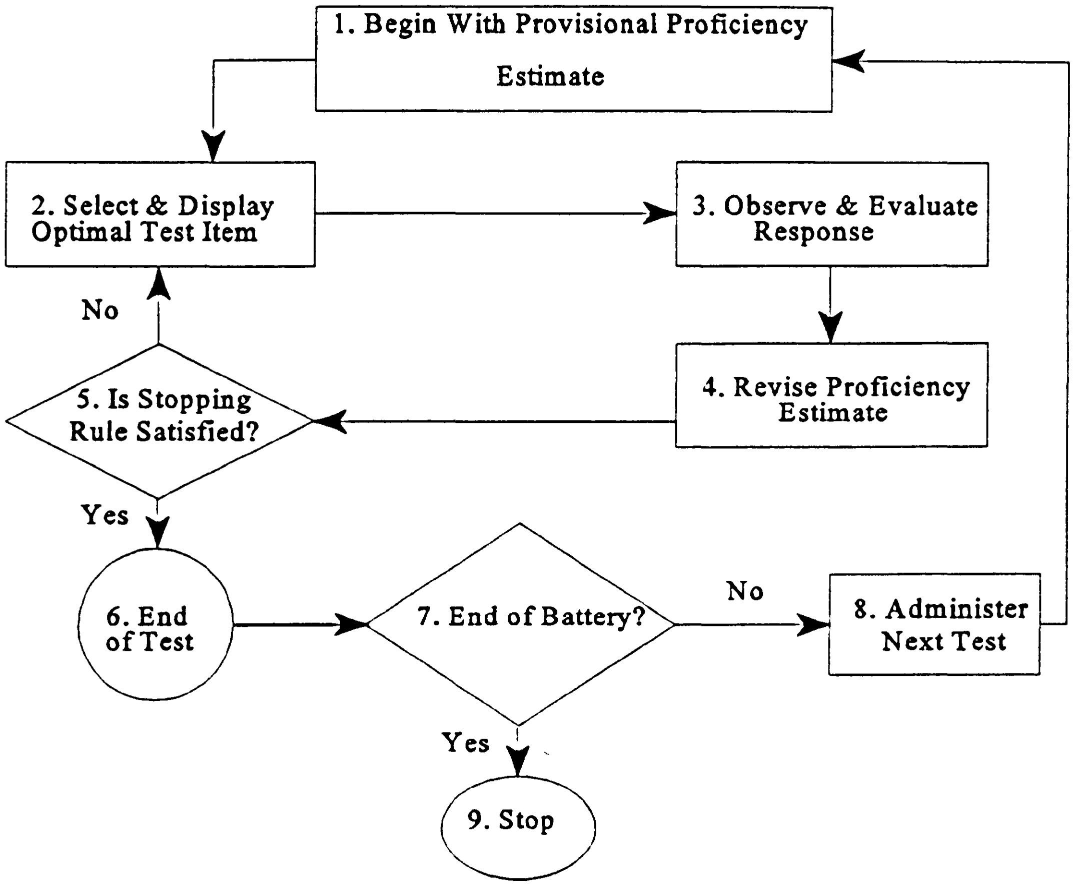
1. Begin With Provisional Proficiency Estimate
2. Select & Display Optimal Test Item
3. Observe & Evaluate Response
4. Revise Proficiency Estimate
5. Is Stopping Rule Satisfied?
No
Yes
6. End of Test
7. End of Battery?
No
Yes
8. Administer Next Test
9. Stop

Figure 2.5. A flowchart describing an adaptive test.

The test terminates when a predefined level of precision has been reached or when a predefined number of items have been administered. The testee's score is the last trait estimate he/she reaches. Other strategies which reflect this model of adaptability, but differ slightly, are maximum likelihood (Lord, 1980) and Bayesian (Owen, 1975) strategies. Because the understanding of these methods requires a fairly sophisticated background in mathematics, their technical details will not described here. Weiss (1985), Vale (1981), and Hambleton et al. (1991) provide good descriptions of these procedures.

Simple adaptive strategies can be administered using P&P format, whereas more complicated ones, like the IRT-based CAT, need to be computerized. Because of its promising role, the remainder of this chapter will be devoted mainly to the IRT-based CAT.

2.3 Item Response Theory

IRT (Hambleton, 1989; Hulin et al., 1985; Drasgow & Hulin, 1990), which is sometimes referred to as latent trait theory or item characteristic curve theory, is a mathematical model which can be used to estimate the testee's trait level based on his/her responses to a set of items with known characteristics, as well as to assess the error of measurement. The theory hypothesizes that the underlying trait for the variable being measured is unidimensional. This means that the test items assess only one variable, which could be ability, a personality trait, or knowledge. The probability of getting an item correct is related to the testee's position on that latent trait.

The item in IRT is considered to be the unit of measurement, where each item in a given test covers a specific region on the trait's continuum, depend on its difficulty level (*b*). That is, each item provides information about certain range of the ability. Answering an item correctly assumes that the testee has the proficiency (theta) necessary to pass that item. Therefore, the task is to present only those items which provide accurate information about the testee's position on the underlying trait continuum.

A model which uses an item difficulty index (*b*) as the only parameter to characterize each item is called a one-parameter logistic or Rasch model. The two-parametric model uses the item difficulty index (*b*) as well as the item discrimination index (*a*). The three- parametric model uses the guessing parameter (*c*) in addition to (*b*) and (*a*) (Birnbaum, 1968). Any parameter excluded from the model is considered to be a fixed index. However, we should notice that although conventional test theory (CTT) and IRT use both item difficulty and item discrimination parameters, they tend to define them differently.

These three parameters (*a*, *b*, *c*) are independent of the sample of individuals on which they are estimated. They can be plotted on one curve, called the 'item characteristic curve' (ICC) or 'item response function' (IRF). This estimates the probability that a person with a certain ability responds correctly to a particular item (Figure 2.6, from Wainer & Mislevy, 1990).

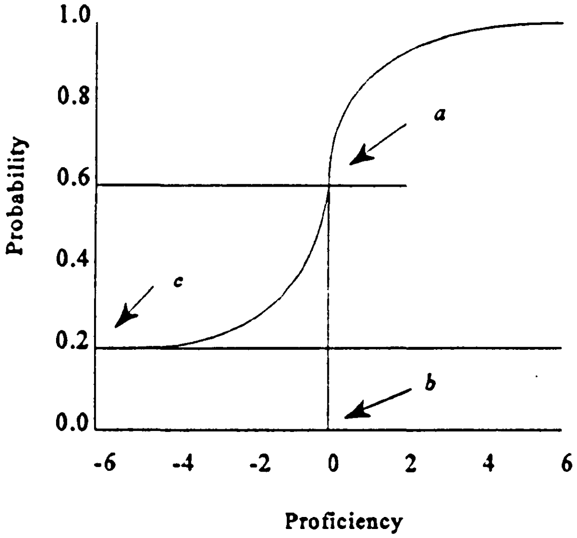

Figure 2.6. Typical ICC for the 3 parametric model.

The point on a latent trait continuum (proficiency) where there is a probability of 50% of a correct answer (or 60% assuming guessing) specifies the item's difficulty index (*b*). This point is considered to be the centre of the curve. The slope of the curve is determined by the item discrimination index, whereas the probability associated with the lower left hand end of the curve is determined by the item guessing index.

There is also a growing consensus among testing experts that measurements based on IRT are among the best measures of test bias (Murphy & Davidshofer, 1994; Drasgow &

Hulin, 1990). This can be achieved by examining the ICC of different group taking a test, for example males and females to check whether an item is more difficult or less discriminatory for high and low ability within the group.

The IRF can also be converted to another curve to provide what is called the 'item information function' (IIF) or 'item information index' (Figure 2.7, From Wainer & Mislevy, 1990). The height of the curve's peak and the spread of the curve along the latent trait continuum are determined by the item discrimination index. This indicates the measurement precision of that particular item at any point on the continuum. The location of the curve is determined by its difficulty level. An item with a high susceptibility to random guessing causes asymmetry in the item curve and some shift in the curve location and lowering of the peak.

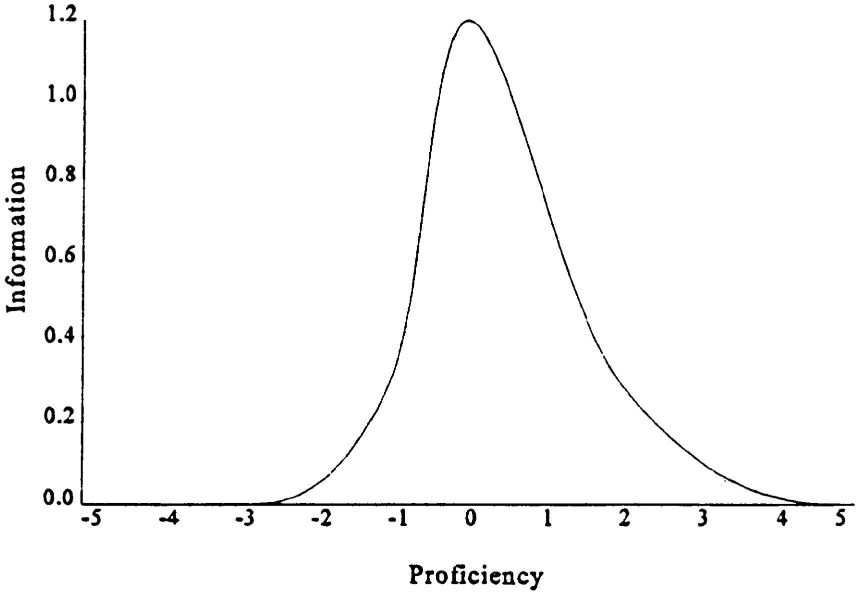

Figure 2.7. Typical IIF for the 3 parametric model.

After each estimation of the testee's ability (or after an estimation of average ability at the start of the test), the item which yields the most information about examinee ability, and which has not already been presented, is selected by the computer. The higher the information gain from an item at a given level, the more accurate the measurement will be at that level. The rule is that a correct answer raises the trait level estimate, and an incorrect answer lowers it.

To minimize any delay in presenting the next item, the computer usually makes two alternative calculations based on the two possible answers, correct and incorrect, while the current item is being presented. The test is terminated when, for example, a specific level of measurement error is reached.

Currently, IRT is being used for test construction and development and for equating and linking the scores of two formats of test, as well as for constructing adaptive tests (Weiss & Yoes, 1991). Drasgow and Hulin (1990) provide a comprehensive review of the applications of IRT to important measurement and substantive problems faced by industrial and organizational psychologist, such as how IRT allows a rigorous evaluation of measurement bias.

2.4 Requirements of CAT

Assuming that the appropriate IRT model has been chosen, CAT requires five essential components to develop a practical adaptive test: a pool of items to select appropriate items from; a starting point; a method for selecting the items; a method for scoring the test; and termination rules. These requirements are discussed below.

2.4 .1 Item Bank

A large bank of valid items is important in order to be able to form optimal tests which differ widely in terms of the ability being measured. These items need to be of high quality for many different levels of proficiency, and highly discriminatory to produce short tests. Weiss (1985) suggested that a minimum of 100 items is required to form an item bank. In most cases no more than 20 items are needed for presentation to testees. Although CATs use fewer items than P&P tests, they require a larger number of items covering a wide range of abilities to be stored in an item pool. This become more important if a parallel form of the test needs to be used. It is important that the item characteristics satisfy the demands of the IRT model selected. IRT item statistics need to be available, and these can be obtained using different groups with large sample sizes.

2.4 .2 Starting Point

Selecting an appropriate starting point helps to reduce the number of items administered to a testee, especially when the starting point matches the testee's ability level. Normally, information about the testee's age and education or other self-report information is used to select the proper starting point. In the absence of such information, an item or a set of items of average difficulty is presented. However, items of average difficulty could be very difficult for low ability testees or very easy for high ability testees. This may affect the overall performance.

2.4.3 Item Selection Procedure

Most adaptive tests follow one of two test strategies: two-stage strategies or multistage strategies, but generally they follow the same rules. A correct answer is followed by a more difficult item, and an incorrect answer is followed by an easier item. With the two-stage strategies (Lord, 1980) there are two sets of items. The first set is called the routing test, which all testees start with. Depending on the score on these items, the testee is then given one of a number of different tests, which have been designed to cover all difficulty levels. The second test is usually longer than the routing test and is aimed at the testee's level of ability. No computer is necessary to implement this strategy because it can be scored by hand, using a conventional P&P format to administer it to a group of examinees. However, a computer can speed up the process of administering and scoring the test.

With multistage strategies, all examinees start with an item of average difficulty and branch to another item or set of items after each response. There are two models of multistage strategy, the first of which is called the fixed-branching model. All testees use the same structure, but each takes a unique route reflecting his/her ability level. The only item statistic used in designing the structure is the item difficulty. When there is only one item at each difficulty level, the test is called 'flexilevel', but when there is set of items at each difficulty level the test is called 'stratified-adaptive' or 'stradaptive' (Weiss, 1973).

The second model for multistage strategy is called 'variable-branching'. In this model, items are selected from a pool with known item statistics. The rule is that an item which increases the examinee's ability estimate is selected from the item pool and presented to him/her.

2.4.4 Scoring Procedure

Using adaptive tests based on IRT, different testees get different items at different levels of difficulty, and also, in some cases, different numbers of items. Because most testees get around 50% of their answers correct (assuming the use of a certain level of measurement precision as a condition for termination) regardless of the difficulty level of the items attempted, the scoring system used in classical test theory (one point for each correct answer) does not work for adaptive tests. In CAT, this system is replaced by an inference from an IRT model about the difficulty level of the items that testees answered correctly. The relationship between proficiency and proportion of correct responses is combined with the data to produce the test score (Wainer et al., 1990). Whatever the IRT scoring method for ability estimation used (maximum likehood or Bayesian procedures), most systems provide two theta estimates: point estimate (the estimate of the testee's location on the theta continuum) and precision estimate (index of accuracy of theta measurement) (Bloxom, 1989). With multistage fixed-branching models more simple methods may be used, like the difficulty of the last item administered, or the average difficulty of all items administered during the test (Vale, 1981).

2.4.5 Terminating Rules

Regardless of the testing strategy used, adaptive tests use one or a combination of stopping rules. A first rule is to stop when an acceptable level of measurement error (SEM) is achieved. The number of items for each testee may vary, but the levels of measurement precision obtained from each testee are the same. However, the measurement precision can be adjusted so that different acceptable levels of precision are obtained for different ability levels (e.g. at the middle levels of ability and at the extremes). A second rule is to stop testing after a fixed number of items have been presented. In this case, the testees may be measured with varying degrees of precision. A third rule is to stop after a pre-specified amount of time has elapsed. This helps in setting a realistic period of time for the test, especially for very slow testees. Again, this could be at the cost of the precision of the proficiency estimate.

The fourth rule is used for classification purposes, such as in job selection, when the decision is to accept or reject candidates. The test may be set to terminate when the testee's confidence interval no longer overlaps with the specified cutoff score. In this case, those with an estimated ability near the pass/fail point will be presented with more items, and more precise estimates will be obtained for their abilities. Those with ability estimates far from the pass/fail point (very high confidence in the decision) will take fewer items, and will be measured less precisely. Bergstrom and Lunz (1991) compared the levels of confidence in pass/fail decision, obtained with CATs and P&P tests. The subjects (645 medical technology students) took a variable length computer adaptive test and two fixed length P&P tests. The CAT was set to terminate when the subject's error of measurement was 1.3 times the error of measurement either above or below the pass/fail point (one-tailed 90% confidence interval), or when a maximum test length was reached. They found a higher confidence in the accuracy of pass/fail decisions when the CAT implements a 90% confidence stopping rule than with P&P tests of comparable length. They concluded that using a confidence interval terminating rule with CAT helps each subject to take a minimum number of items and ensures the pass/fail decisions are made with a high level of confidence.

Finally, developers will have to decide wether or not the testees will be informed of the terminating criteria. Hiding the termination rule will increase the testee's uncertainty.

2.5 Reliability of CAT

The reliability of a test is concerned with the extent to which a test consistently measures what it is intended to measure. With P&P tests there are three main types of reliability: test-retest; alternate form; and internal consistency. In CAT, reliability is considered in terms of the precision of testee ability estimates, which can be predetermined as a stopping rule. Samejima (1977) and Bartram (1994) argued that the standard error of measurement is more useful in score interpretation than the conventional reliability coefficient.

The use of the conventional reliability coefficient is based on the false assumption that error variance is the same for all scores regardless of testee ability, although conventional tests measure poorly at the extremes. However, the estimated standard error of measurement for CAT is constant over a wide range of abilities, even when it is not used as a stopping rule.

The reliability coefficients used in CAT are marginal reliability, alternate form reliability, and test-retest reliability (Green et al., 1984). The construction of marginal reliability for IRT scores parallels the construction of internal consistency estimates of reliability for P&P scores (Thissen, 1990). For comparison purposes, the vital question is whether the precision of measurement obtained using CAT differs from that obtained with P&P tests. Divig (1988) carried out a computer simulation study and demonstrated that a CAT format of the ASVAB had a higher reliability than a P&P format for five subtests (general science, arithmetic reasoning, word knowledge, paragraph comprehension, and mathematics knowledge).

Alternate form reliability for CAT requires a separate item pool for each form of the test, to avoid presenting the same items in the second version. Thus, the correlation between the two forms assesses whether the two item pools are equally related to the same underlying psychological construct. Item sampling may cause variation which could affect the correlation between the two forms. McBride & Martin (1993) indicated the possibility of administering the two parallel forms of the test simultaneously, using software designed to separate the items into two tests.

Test-retest reliability is concerned with the stability of the test over time. Early studies in which adaptive tests were administered to real testees were concerned mainly with test-retest reliability. These studies showed higher test-retest reliability for adaptive tests in comparison with conventional tests (e.g. Betz & Weiss, 1975; Larkin & Weiss, 1974; Vale & Weiss, 1975). As is the case with P&P tests, CAT test-retest reliability is affected by memory and motivation factors. These factors could cause the item selection procedure to produce different items on the second administration when the same test is needed. In conclusion, Bloxom (1989) analysed recent studies reporting real-data results and concluded that CATs provide greater efficiency and reliability than P&P tests.

2.6 Validity of CAT

Concerns have been raised about CAT's content-related validity. Steinberg (1990) argues that item selection procedures may not ensure content representativeness, since items are selected on the basis of high discrimination at a specific level of ability. This may be particularly important when establishing the equivalent P&P form, for which content validity is considered to be relatively easy to control. Green (1988) points out that balanced content may not always be easy to achieve. Rechase (1989), on the other hand, thinks that current item selection procedures are adequate. Burke (1993), distinguishes between controlling for content representativeness of items at each level of ability, and ensuring that test performance is a representative sample of job performance or job-required knowledge.

Item selection procedures may also cause context effects, when an item presented acts as a clue for the answers to subsequent items. This could affect the subsequent item difficulty. Although this problem can also be found with conventional P&P tests, the situation with the CAT is more serious, since different items (equivalent tests) are presented to different testees according to their performance. This could distribute the disadvantages unfairly among testees. Of course, this can be avoided by ensuring the independence of each item from the others when developing the items. However, it is very difficult to scrutinise all possible pairs in a large item pool. Wainer and Kiely (1987) and Wainer (1990) suggest using pre-clustered sets of items (testlets) to reduce the problems of item order, context effects, and content balancing.

In their effort to assess the convergent validities of both adaptive and P&P tests using arithmetic reasoning, vocabulary, and paragraph comprehension tests given to 356 Marine recruits, Moreno et al. (1984) concluded that a CAT as long as 15 items can have the same convergent and discriminant validities as a P&P test which is twice as long. McBride and Martin (1983) using verbal ability test reached a similar conclusion. Bloxom (1989) criticized the two studies because they do not show the relative precision of adaptive and P&P tests as a function of the testee's level of theta. They indicated the importance of not assuming that a very short adaptive test will necessarily provide a high level of reliability or convergent validity. Moreover, for a given test, the move from P&P to a computerized mode could affect the equivalence between the two formats, and hence the construct validity of the test (Steinberg, 1990). In his study to compare the effect of mode of administration of test items (CAT versus P&P), Rackse (1986) reported that some items were found to operate differently when administered on a computer screen compared to a P&P administration. He called for further research to determine the cause of the differences in item performance. The effect of switching from P&P to CAT format is thought to have three aspects: a) an overall mean shift, in which all items may be easier or harder; b) an item-mode interaction, in which a few items may be altered and others not; and c) the nature of the task itself, which may be changed by CAT administration (Federico, 1991). Mode effects were discussed in the first chapter.

One important point, particularly in the selection and placement process, is the issue of the criterion-related validity of CAT. What is available so far suggests that CAT yields similar predictive validity compared to its P&P counterpart (e.g. Cudeck, 1985; Kent & Albanese, 1987; Koch et al., 1990; Sympson et al., 1982, 1984; McBride, 1980; McBride & Martin, 1983; Moreno et al., 1984; Sand & Gade, 1983; Moreno et al., 1985). Weiss (1985) reviewed the studies by Kiely et al. (1983) and McBride and Martin (1983), where they compared the reliability and validity of different types of tailored tests with their conventional P&P counterparts. They found that the reduction in test length produced by administering the adaptive test does not sacrifice either reliability or validity. Sands and Gade (1991) evaluated CAST (Computerized Adaptive Screening Test) as an automated replacement for EST (Enlisted Screening Test) and to develop a prediction model for using CAST to forecast AFQT (Armed Forces Qualification Test) scores. They reported that CAST predicts AFQT as accurately as does the EST, and that CAST is considerably more efficient. McBride (1980) used a military recruit population and reported higher concurrent validities (correlations with a 50 item criterion test) for adaptive tests at test lengths up to 10 items, but equal or slightly higher validities for conventional tests from 15 to 30 items in length. Thompson and Weiss (1980) correlated scores on adaptive and P&P tests with grade point average (GPA) for groups of college students and found significantly higher correlations for some adaptive tests in comparison with P&P tests. Moreover, although they did not find any statistically significant increases in criterion-related validity due to adaptive testing, Sympson and Wiess (1982) reported the feasibility of adaptive testing in a military testing environment.

Their data showed that adaptive tests could provide levels of measurement precision obtainable only with much longer ASVAB tests, and that the adaptive tests were one-third to one-half the length of P&P format. However, as is the case with the P&P tests, the difficulty in assessing the predictive validity of CAT concentrates on problems related to the criterion used, restriction of range, and the ethical and legal problems of item bias (or what is called differential validity).

2.7 CAT and Speeded Tests

Speeded tests contain very easy items of matched difficulty; the testee would certainly answer all of them correctly if there were sufficient time. The question here is not simply whether the testee gets the items right, but also how quickly he/she can do so. The earlier discussion in this chapter applies mainly to power tests, where there are a number of items varied in terms of their difficulty levels to measure the testee's proficiency without regard to how long it takes him/her to respond to the items. Because speeded tests measure at least two dimensions of traits, the ability being measured and the response speed, they violate the assumption of unidimensionality imposed by IRT. Speeded tests do not lend themselves to individual tailoring in the way that power tests do (Henly et al., 1989), and the difficulty of the items cannot be used as a factor for item selection procedure. Therefore, the P&P items can simply be administered by a computer. For this reason, the speeded tests of ASVAB and DAT have not been adapted. Mead and Drasgow (1993) conducted a meta-analysis to investigate the equivalence of computerized and paper and pencil tests. They analysed scores from 115 tests and found a high levels of equivalence (*r* around .90) for power tests, but lower equivalence (*r* around .60) for highly speeded tests.

There is no need for initial item selection strategies, provisional estimates, or complicated item selection functions when dealing with speeded tests. However, a stopping rule must still be specified (Thissen & Mislevy, 1990). The stopping rule may be based on either a fixed number of items or a fixed time limit. In these cases, the time to complete the test or the number of items completed can be used to calculate the test score. Greaud & Green (1986) compared the different scoring strategies used with speeded tests. They found that the average number of correct responses per minute was a more reliable measure than was a number-correct score. However, it is possible to get a high number of correct responses per minute by pressing the computer keys randomly without even reading the items. Computers can easily be programmed to calculate the response time for each item, as well as to control the exposure time of an item.

Speeded tests are sensitive to mode of presentation and to hardware and software changes. It should also be mentioned that the order effects mentioned earlier are magnified when tests are speeded (Hambleton, 1986). The only reliability indices available for speeded tests are those based on alternate forms or test-retest (Thissen, 1990).

2.8 Adaptive Personality and Attitude Questionnaires

All the discussion so far has concentrated on the application of IRT for dichotomous item responses (correct or incorrect). Disregarding a few applications of IRT models to the analysis of multichotomous responses, such as Likert-type response scales found in survey or personality and interest questionnaires (e.g. Hulin et al., 1982; Steinberg, 1986; Thissen & Steinberg, 1988), the power of computerized adaptive assessment in this domain has remain largely unexplored (Wainer et al., 1990). As with any application, the design of an adaptive system for the measurement of personality and attitudes needs an item bank, an item selection procedure, an IRT model providing item characteristics and a scoring procedure, and a termination rule. The item characteristics, or scale values, can be estimated from the responses of a group of individuals similar to those who are to be measured. The result is known as an 'item category response function' for each possible response, which then can be used in maximum likelihood procedures to estimate testee trait levels (Weiss & Yoes, 1991).

Some IRT models for multiple-response alternatives are available, e.g. the graded response model (Samejima, 1969), the rating scale model (Andrich, 1978), the partial credit model (Masters, 1985), and the nominal response model (Bock, 1972). Such models generally provide a distinct trace line for each possible response, to be used for estimating values of the latent attitude or personality trait. Koch et al. (1990) used the rating scale model to conduct a study measuring undergraduate students' attitudes toward alcohol. They found a correlation coefficient of .81 between the CAT and the P&P versions of the questionnaire, and concluded that the CAT procedure performed very well. The comparability of computerized adaptive and conventional questionnaire versions of the MMPI-2 has been assessed by Roper et al. (1991). They used 155 college students and found that profiles across both formats showed a high degree of similarity, and that substantial item savings were found with the adaptive version.

Questionnaires can also be designed so that items which are not relevant to the examinee are eliminated. For example, only those items which are related to the aspects of personality required for success in a specific job (e.g. persuasion, innovation, forward planning) may be presented. Those which are irrelevant are eliminated. Also biographical data items can also be reduced by, for instance, eliminating marital questions from presentation to unmarried individuals. This can dramatically reduce the number of items usually administered in P&P formats, and also the boredom experienced with long personality and attitude questionnaires. Again, adaptive systems for personality and attitude measurement can be more sensitive to the effects of item ordering and context (Bock et al., 1988).

2.9 System Design

All early attempts at developing computerized tests used large mainframe computers. However, for reasons of cost, poor display, absence of a suitable operating system and shortage of techniques for setting up adaptive testing systems, the early projects often failed (Hambleton, 1991). Today, almost all these problems have been resolved by the use of advanced technology. The hardware and software requirements for implementing simple computerized tests were discussed in the first chapter.

To enhance proficiency, CATs need sufficient memory (Random Access Memory, RAM) to store both the item pool characteristics and the program requirements so that they are accessible directly by the central processing unit (CPU). Sufficient Hard Disk (HD) storage is also needed to maintain a large item pool, instructions, and demonstration items. The necessary capacity of the storage depends on the size of the test(s) and the type of items. For example, text type items require less bytes of storage than pictorial type items. The Differential Aptitude Tests (DAT) require 256K of memory and 700K bytes of disk space. The ASVAB, on the other hand, requires around 4MB of disk space.

Moreover, CATs need sufficient speed of disk retrieval and calculation to minimize the waiting time between individual items. Usually, less than 2 seconds between items is acceptable. Anticipating the testee's responses and calculating an estimated level of ability for each possible answer while the testee is attempting to answer the current item, helps in displaying the next item faster. With these constraints in mind, most of the PC machines available today can serve as units to be used for CAT. However, although most systems are capable of administering computerized versions of P&P tests, it is a hard choice for users and organisations to decide which hardware standard must be chosen. The rapid developments in the computer industry make the choice of the best standard of hardware for computerized administration a difficult task, leading APA (1986) to consider a test which has been implemented on two different hardware combinations as two separate tests. Consequently, equivalence studies might become limited to the specific hardware on which the study was done, and cannot be generalised to other hardware setups (Schoonman, 1989).

On the other hand, software developed to implement CAT needs to be designed to undertake the following tasks: a) recording of the testee's biographical data; b) instructing the testees in how to take the test; c) easy updating by the test user of any items in the item pool; d) controlling the presentation of items according to the rules of IRT or any other specified algorithm; e) terminating the test on the basis of the testing strategy used; f) scoring the test and producing narrative scores; g) presenting the test results on the screen or on paper; h) recording any other desired data such as response time or norms; and i) storing of all test data. The model needs to be user-friendly, either menu driven or with a graphical user interface (GUI), and able to give remedial instructions when improper actions are taken. Moreover, the software needs to be crash-proof, and secure enough to prevent theft of the item pool or unauthorised access to the data. The developer can use any major programming language to develop CAT (e g. C+, PASCAL, FORTRAN, BASIC, ...etc.).

One of the well known testing systems which integrates all these hardware and software requirements is the MicroCAT (Assessment Systems Corporation, 1989), which has been in use since 1980. MicroCAT is menu-driven and provides all the facilities for implementing CATs and conventional tests, whether they are based on IRT or classical test theory. Also, it is able to integrate colour graphics as well as text items. MicroCAT was developed for personal computers so that occupational psychologists, among others can take full advantage of it to improve the efficiency of occupational tests. Hsu & Yu (1989) and Stone (1989) have reviewed many other packages available to support CATs.

2.10 CAT's Advantages

Many of the advantages which were mentioned in the first section concerning CBT can be generalized to CAT. These are: savings in time as a result of quick test scoring and reporting; the ability to use moving stimuli; the recording of response latencies; greater standardisation; a reduction of the direct involvement of the examiner in the actual testing situation; the ability to control the exposure time; and ease of updating the material .

One of the most important advantages of using CAT is its ability to provide more precise ability estimates with fewer items than P&P tests. Uninformative items are eliminated, and only those which provide further information about the testee's ability are presented. The direct advantage is a reduction in test length of around a 50% (Weiss, 1985). This is particularly helpful in situations where the time of testing is limited, or where a number of abilities are to be assessed, or with groups who are differentially sensitive to test length and exposure to many items which are above their ability level (Offerman & Gowing, 1993). However, the reduction in the number of test items does not sacrifice measurement quality. Bejar et al.(1977) in their live test administration study comparing an adaptive achievement test to a P&P classroom test and using information as the evaluative criteria, found that the use of adaptive tests results in scores which are less likely to be confounded by errors of measurement, and a reduction in the number of test items administered.

As mentioned earlier, most P&P tests measure most precisely for those testees with trait levels at or near the difficulty level of the test, but they are worse at measuring those with trait levels far from that particular level. In some situations where a comparison is being made between two or more groups who differ in their initial trait level (e.g. different age groups) or when treatment effects are being measured, the P&P test will measure more precisely and show greater changes for the group with a trait level at or near the test difficulty level. A CAT based on IRT avoids this problem by measuring well at any point on the trait continuum specially the ability levels of those of high or low ability (Anastasi, 1988), thus making more accurate comparisons between groups who differ in their initial ability level (Embreston, 1990). However, the author has not found evidence to support this claim.

Ensuring more security for test results and items is another advantage. In P&P tests the same items are presented to all testees, maximising the chance of cheating or discussion of the test items with others who are to be tested in later sessions. With CAT, different items are presented to testees at different ability levels, and different initial items at the same difficulty level are presented to each testee. Also, with CAT, there are no test booklets to be stolen. This minimises any chance of cheating.

In addition, CATs help reduce test anxiety by presenting items which challenge but do not discourage the testee. This helps to maintain a constant level of motivation in answering the test items (Betz, 1977; Betz & Weiss, 1975; Lord, 1980). Presenting far too difficult items for low ability testees cause frustration, blind guessing, and increase test anxiety. Conversely, easy items presented to high ability testees may make the test session a boring experience and lead to carelessness or ease suspicion about the correct answer of the item. Both cause unwanted error which may affects test reliability and validity. Given the projected differences in worker ability levels and the projected number of low-skilled applicants, CAT may prove particularly useful (Offerman & Gowing, 1993).

Some CATs are designed without time limits, which means the testee can work at his/her own pace. In this way the testee's acceptance of the test is enhanced. However, a system program is necessary to control the test time for those who are not trying to answer or are taking an unusually long time. A small clock displaying the time remaining for the present item or for the whole test in one corner of the screen may be useful, but it also could cause unwanted time pressure.

2.11 Limitations of CAT

Whether the transformation from P&P format to the computer is to develop simple computerized or CAT versions of a test, some common difficulties can be expected in both situations. These difficulties have already been discussed in the first chapter, but they can be summarised as follows: possibly higher anxiety from computers; limited numbers of testees in one session due to the financial cost of purchasing computers; threats to client privacy and confidentiality; the effects of experience in using computers; and poor screen resolution for pictorial items.

However, other possible limitations are found only with CAT versions. CATs based on IRT demand more attention and experience in developing test items, and the items need to be unidimensional. Other issues relate to improper item selection procedures; selecting items for their level of difficulty and discrimination may not satisfy the need for content and context balances or proper item ordering, which in turn may threaten the construct validity of the test as discussed previously.

Concerning item ordering, most CAT tests start by presenting an item of average difficulty. For reasons of test security, CATs avoid presenting the same initial moderately difficult and highly discriminating item for all testees (Drasgow & Hulin, 1990). However, moderately difficult item may be very difficult or very easy for those at the extremes of the ability continuum. Presenting a difficult item at the start of a test may increase the testee's anxiety and frustration, whereas an easy one could be boring and may reduce the testee's motivation. In both cases, the test length must be increased to reach the predefined level of measurement precision. CATs can be designed to start with easy items, but this would be at the cost of reducing the efficiency gains from adaptive tests. This issue will be further discussed in Chapter 6.

One of the headaches of CAT development, but an important issues nonetheless, is the issue of calibration of the test items (Gialluca, 1988). CAT requires a large number of people to assess the item parameter estimates before they can be used. This step is vital when the CAT format of a test needs to be equated with its P&P version, or when different CAT versions are intended to be equivalent. In practice, this is a complicated, time consuming, and very technical problem (Wainer & Mislevy, 1990). The issue of equivalence between the P&P and CAT formats of a test will be discussed in more detail in Chapters 3 and 4.

With P&P tests, most paragraph comprehension items require a relatively large paragraph of text, followed by multiple answers. Because of the limited size of the computer screen and the high interdependence between the multiple answers for one paragraph, items in this form may not be suitable in CATs. Some developers used shorter paragraphs and multiple answers on one screen (Wainer & Kiely, 1987). Green (1988) noticed that these short paragraph comprehension items are more similar to conventional word knowledge items than they are to previous paragraph comprehension tests, which may threaten the construct being measured (Bunderson et al., 1989).

Concern has also been expressed about the fairness of CAT. In practice, testees receive different items, depending upon their ability, but each test is intended to be randomly individualized and parallel (Lord & Novick, 1968). This may seem as if testees do not take the same test or have the same chance of success. Indeed, the notion of adaptability is not new in psychology. In psychophysical scaling, one of the examiner's jobs is to find the threshold value for the examinee receiving an auditory stimulus. This value is the point where the examinee has a 50% chance of detecting the stimulus. There is obviously no point in presenting certain levels of stimuli when they are clearly out of range. Put another way, the high jumper who fails to exceed the 1.5 metres mark need not attempt to jump 1.8 metres. Similarly, those who fail to answer a simple mathematical question presumably are unable to answer a more complex one (assuming no guessing). However, it is important that a CAT's items are ordered carefully in terms of difficulty along one dimension to ensure fairness.

Another limitation, from the testee's point of view, concerns the inability to omit, skip, or review items (Murphy & Davidshofer, 1994). Because the selection of the next item from the item pool depends upon the testee's answer to the present item, CAT requires that each item be answered before the next item is presented. From the developers' point of view, reviewing and altering item responses may change the estimate of testee ability in such a way that items will be poorly targeted and precision will be lost (Lunz et al., 1992). This seems annoying and unfair to those who would like to give more attention later on to difficult items, or who hope to find a helpful clue from subsequent questions, or who want to change an incorrect answer. The above assumption is not supported by two studies carried out by Lunz et al. (1992) and Wise et al. (1989). However, the developer and user may see this constraint as a security to prevent the omission of any item that the testee should answer.

Another more practical issue emerges when the test, for one reason or another, is unfinished. Some testees may not attempt to answer very difficult items at the end of a test to hide their true ability. Some low ability testees may be aware that CATs start with an average estimate of ability, and may not attempt to answer any items in the first place. What should be done in these circumstances? What sort of penalty should be given? No clear answer is available. CAT-ASVAB, for example, assumes that the testee would have answered the incomplete items at random, by guessing. For some testees, that could be just what they want; a guess is better than nothing. Even taking an estimate of ability after the last answer, when the test is not finished, is satisfactory for those who choose not to answer hard items in the final stages of the test. Similar practical problems are discussed by Noonan & Saravela (1991). Obviously, more studies are needed to find the most suitable way to handle this issue.

Finally, what happens when testees know that the test they are going to take is adaptive? That is, the difficulty of the next item(s) depends on their performance on the present item(s). CATs developers and users do not seem to be trying to hide the nature of the test from their testees. The orientation booklet of DAT indicates clearly on its cover paper the nature of the test . The question here is whether knowledge of the basic function of adaptive tests affects a testee's performance. This issue will be further discussed in Chapter 5.

2.12 Conclusion

CATs can be classified into three categories: adaptive item presentation, based on IRT parameters; adaptive item presentation times, based on previous response times; and adaptive content or composition of items, based on previous choice (Bunderson et al., 1989). Most of the studies so far are concentrated in the first category. The theoretical as well as the simulated and live-testing evidence provided by many studies suggests that the adaptive test does work, and its increased use should be expected (e.g. Garrison & Baumgarten, 1986; Moreno et al., 1984; Ward et al, 1986; Weiss, 1985). CAT promises more precise measurement at the extremes of the ability continuum than the equivalent P&P tests, and adequate measurement at the middle of the continuum. Most of the adaptive tests available today are versions of existing P&P tests. Also, there is limited but growing application of IRT to personality and interest questionnaires.

A new generation of adaptive tests is on the way. Rather than having to calibrate the test items, adaptive tests can be designed so that the test algorithms automatically generate the test items and control their psychometric characteristics. This is both more economical and more efficient (Bejar, 1986). Attempts have also been made to allow the testee to adapt the level of test difficulty in the way he/she prefers. Rocklin & O'Donnell (1987) conducted an experiment which contrasted a variant of computerized adaptive testing, self-adapted testing, with two more traditional tests (one relatively easy and one relatively difficult), which all shared the same bank of verbal ability items. With self-adapted testing, they allowed the testees rather than a computer algorithm to choose the difficulty of the next item. They found that the self-adapted test led to higher ability estimates and minimized the effects of test anxiety without any overall loss of measurement precision.

Wise et al.(992) also compared the relative effects of computerized adaptive testing and self-adapted testing, and they found that testees taking the self-adapted test obtained significantly higher ability scores and reported significantly lower post-test state anxiety than those taking computerized tests. They concluded that the self-adapted test is a desirable format for computer-based testing.

Other aspects of adaptability can also benefit from a computer's capabilities. Software can be designed to allow testees to adjust the size of the text letters and the brightness and colour of the screen, to find the combination they are most comfortable with. This gives the testee more control over the test environment.

Although computers allow the production of items with moving objects and audio capabilities, the fear of violating the assumption of unidimensionality of IRT-based items leads developers not to use such facilities. Such a violation is also likely to occur in arithmetic word problems which require verbal as well as numerical skills. However, some attempts have been made to measure two or more proficiencies simultaneously (Reckase, 1985; Whitely, 1980). For example, Reckase (1985) developed multidimensional IRT models which assume that a testee's response is dependent on his/her position on several latent traits. In a situation requiring the measurement of two abilities for example, the item response function is graphically depicted in three dimensions; theta 1, theta 2, and the probability of a correct response.

A number of other important issues need further investigation, such as the starting difficulty level, the item selection algorithm, and proper stopping rules. More studies are needed to limit the effects of context, item ordering, and content balancing. No studies have so far dealt with the possible effect of knowledge about adaptive tests on testees' performance. The success of CAT depends upon having a large bank of pre-calibrated test items to measure a unidimentional trait. Finally, the promising potentials of CAT has led some researchers to be more optimistic about its role in human assessment. Hakel (1986), for example, reached the conclusion that " change is coming, and computerized adaptive testing is going to force us to update our working knowledge of psychological measurement". Bartram (1989) also concluded that " We are likely to see an increase use of CAT based on such item-banks in selection and placement (initially within the military and Civil service and subsequently spreading out into industry and commerce through the graduate selection process)". To what extent their optimism is supported by this thesis will be revealed in the next chapters.

CHAPTER 3

CHAPTER 3

Experiment 1

The Equivalence and the Predictive Validity of the Paper-and-Pencil and Computerised Adaptive Formats.

3.1 Introduction

Considered in isolation, the advantages mentioned in the first and second Chapters associated with using the CAT format cannot justify switching from the P&P format of the same test. There must be proper study checking the equivalence between the two modes, if norms, cutting scores, and validity data from the traditional test format are to be generalized to the computerized version. As some of these CATs are alternative versions of existing P&P tests, it becomes important to establish the equivalence between the two formats before assuming that they can be used interchangeably.

A number of studies have been carried out to investigate the equivalence between the two formats. Henly et aL (1989) compared the conventional and the adaptive versions of DAT, and reported that the CAT version is an adequate representation of the conventional test except for the speeded test in the battery. Cudeck (1985) compared the two versions of the ASVAB and also reported favourable correspondence between them.

Mead and Drasgow (1993) conducted a meta-analysis to investigate the equivalence of computerised and paper and pencil tests. They analysed scores from 115 tests and found a high levels of equivalence (*r* around .90) for power tests, but lower equivalence (*r* around .60) for highly speeded tests. Other studies reported similar findings (Weiss & McBride, 1984; Maurelii & Weiss, 1981). On the other hand, the CAT version of Raven's Standard Progressive Matrices was tested by Kubinger et al. (1991) in order to assess its equivalence with the P&P format. They proved the homogeneity of the items, but an item bias was found between the two forms, where the computerized form led to a score difference of up to 13 IQ points on average. It should be noted, however, that these results are obtained mainly from studies conducted in academic and military organizations, and with American subjects only. The APA guidelines (1986) specified two conditions to establish the equivalence between two formats: "*(a) the rank orders of scores of individuals tested in alternative modes closely approximate each other, and (b) the means, dispersion, and shapes of the scores distributions are approximately the same, or have been made approximately the same by rescaling the scores from the computer mode*".

While many of the results obtained from previous validity studies (discussed in Chapter 2) are supportive of the use of adaptive tests, their generality with regard to practical applications in industrial and organizational environments is somewhat limited. The evidence is very restricted, and concentrates mostly on academic and military institutions. Also, most of these studies have been conducted in the United States, and mainly with English speakers. Therefore, studies are needed regarding the criterion-related validity of adaptive tests in comparison with P&P versions.

This study looks at the validity of both P&P and CAT formats of the Differential Aptitude Tests (DAT) for predicting the success of refinery operator trainees at the Kuwait Petroleum Corporation[1]. The selection of candidates who subsequently demonstrate a high probability of satisfactory performance at work is a goal of most selection committees. Admission decisions are serious ones, both for organization and candidates. The rejection of candidates who would actually succeed or the acceptance of those who fail results in bad experiences for the prospective employees, and interferes with the development of a successful career. Similarly, organizations suffer when an inaccurate selection decision is made, by losing a potentially productive employee or by choosing an individual who does not belong in the organization's environment.

At the Kuwait Petroleum Corporation, where this study was conducted, candidates are selected as refinery operators according to high school ratio (HSR)[2], age, and nationality. These criteria may not be enough for accurate selection, especially when we know that the validity and reliability of high school tests are open to question (Newstead & Dennis, 1994). This highlights the need for a study to find other possible

[1] The author would like to thank Kuwait Petroleum Corporation for their support and enthusiasm in providing access opportunities to collect data for this study.

[2] High school ratio (HSR) can be computed by dividing the total grades earned in the all final year high school courses by the maximum mark (sum of the full marks in all courses) and multiplying the results by 100.

variables which could be used in selecting candidates for similar training courses. The underlying assumption is that an investigation of the variables influencing trainee performance in the training programme could probably be used to set quantifiable criteria for admission.

Therefore, the first aim of this experiment was to investigate whether the scores obtained using a CAT format of the DAT would differ significantly from the scores obtained using the conventional P&P format, in order to determine whether the two versions can be regarded as equivalent. Unlike previous attempts, this study uses Kuwaiti subjects. The second aim of the study was to assess the predictive validity of the P&P and CAT formats of the DAT in predicting training performance, and to investigate whether the CAT format yields any better predictive validity.

3.1.1 Refinery Operators' Training Programme

The training programme was conducted by the International Human Resources Development Corporation (IHRDC-Boston, USA) to qualify new trainees at the Kuwait Petroleum Corporation (KPC) to work as refinery operators. The programme has been tailored to be job related, satisfy the needs of the future refinery operator, and follows on from an earlier job analysis conducted by the KPC (not given to the researcher). Although classroom teaching is the main mode of instructional delivery, due to the nature of the programme, discussion sessions, assignments, workshops, videos, and visits to refineries are also included. The programme is divided into three units which fall under the following categories: Unit 1, academic fundamentals (8 weeks); Unit 2, technical training (12 weeks); and Unit 3, refinery technology (12 weeks). The training programme subjects include Organic Chemistry (OC), Process Chemistry (PC), Process Physics (PP), Engineering Science (ES), Mathematics (Math), Engineering Drawing (ED). All trainees receive the same material, and all instructors cover the same ground. Once the trainees have completed 8 months training at the KPC Training Department, they are transferred to one of the refineries for the on-the-job training segment of this integrated training programme. A monthly evaluation, in the form of examinations and quizzes, is conducted to measure trainees' performance. In addition, other factors such as attendance, tardiness, class participation, assignments and discipline are all accounted for in these monthly evaluations.

3.1.2 Paper-And-Pencil Testing

The standard version of the Differential Aptitude Tests, first published in 1947 by the Psychological Corporation, was used in this study. Forms L and M were published in 1965, followed by Forms S and T in 1972. The first British edition of the DAT was published in 1979 (Forms S and T) followed by Forms V and W in 1981 (The Psychological Corporation, 1986). The DAT is a series of eight aptitude tests (Verbal

Reasoning, Numerical Ability, Abstract Reasoning, Clerical Speed and Accuracy, Mechanical Reasoning, Space Relations, Spelling and Language Usage). The test is an integrated assessment procedure widely used for educational and vocational counselling, and also in business and industry for selection and career planning decisions. The DAT is a well-developed and well-documented test which has been regularly updated (see *Review of psychological tests for assessment in vocational training*, 1992).

Three of the ability tests (Numerical Ability, Abstract Reasoning, and Mechanical Reasoning) were used in this experiment. The choice of the three tests used with the field study was not carried out according to job analysis study, as should normally be the case. The Kuwait Petroleum Corporation claimed that job analysis study for refinery job has previously been done, but was reluctant to make it available to the researcher on the ground that it is confidential. Therefore, the selection of these three tests was done by quick scanning of the training curriculum and what might appear relevant to the job. The items in the NA test were designed to test understanding of numerical relationships and facility in handling numerical concepts. Numerical ability is required for success in school and college courses such as mathematics, physics and engineering. These skills are also important for occupations such as accountancy, computing, bookkeeping, and statistical and clerical work. The AR test involves the ability to perceive relationships in abstract figure patterns. The test is relevant in situations where the curriculum, profession or vocation requires perception of relationships among things, rather than among words or numbers. The MA test is useful in making decisions as to the suitability of students for those occupations or curricula that requiring an appreciation of the principles of commonly-faced physical forces. It is useful for selection in occupations such as carpentry, mechanics or assembly work.

3.1.3 Computerized Adaptive Testing

The CAT version of the DAT (Form V), which was developed in 1986, consists of the same eight aptitude tests mentioned above, as well as a Career Planning Questionnaire. The system has the ability to score and record the results, and to print an individual profile report as well as a Career Planning Report. In addition, the user can tailor the required sets of the test battery, and can interrupt a testing session at any time and resume testing later. The tests used in this experiment were the NA, AR, and MA.

The Rasch model (1966) was used to compute the initial item statistics. On the basis of the subject's response and item parameter values, the ability estimate is updated using the Bayesian technique (Owen, 1975). Given this updated ability estimate, the most informative of the remaining items is selected, after which a new ability estimate is calculated. The process continues until twenty items have been administered. Equipercentile equating (Braun & Holland, 1982) is used to convert the ability estimate into raw scores (Henly et al. ,1989), and these raw scores are used in this study for analysis purposes.

3.2 Method

3.2.1 Examinees

A total of 122 refinery operator trainees took part in the study. All the trainees were Kuwaiti males who had graduated from the Science division (as apposed to the Arts division) at high schoolL They had passed the pre-qualification requirements stipulated by the Kuwait National Petroleum Corporation (KNPC) prior to joining the training course. They were selected according to age (under 21 years), nationality (all Kuwaitis), high school ratio (above 55%), health (no disease and good general build), and were not employees at any other company. The average age of the sample was 19.2 years (range between 18 to 20 years). The first language for all of them was Arabic. However, the trainees received English language training conducted by KPC throughout the eight months of the training programme. Most of the trainees had no prior work experience, and were selected from various high schools in Kuwait.

3.2.2 Materials

Differential Aptitude Tests (DAT)

For this study, the NA, AR, and MR tests of the Differential Aptitude Tests

(DAT), in both the P&P (Bennett, Seashore & Wesman, 1986) and the adaptive formats (McBride, 1986) were used. The CAT was developed from the items of Form V of the DAT. There were 40 items in the P&P format and 20 items in the computerized adaptive format (CAT). The test manual reports that the split-half reliability coefficients for the DAT lie between .89 and .95 for males and .85 to .95 for females (KR20 reliability ranged from .85 to .94 for boys and .79 to .94 girls). The test Manual reported good validity studies.

Items were presented either in a standard booklet format for the P&P version or on a computer screen for the adaptive version. The tests were presented to trainees in English language . The test used five-option multiple-choice questions. "Yes" and "No" keys used to enter the answers were labelled with green circle stickers. Each of the tests had a set of instructions with three sample problems, and the examinees were instructed to work as quickly and as accurately as possible.

3.2.3 **Procedure**

The 122 trainees were randomly assigned to the two groups: paper-and-pencil first group (P&PFG), and computerized adaptive test first group (CATFG). All trainees received the NA, AR, then MR tests respectively in each format. The P&PFG (n=60) received all the P&P format tests first followed by the CAT format, while the CATFG (n=62) received them in the reverse order (Appendix D). The interval between the two test sessions was between 7 to 10 days, with a 10 minute break between each test. The study represents a predictive validation study. The researcher waited until the end of the training programme to get all the trainees's performance data on the training courses.

All tests were administered according to the instructions in their manuals. The use of calculators was not allowed. All trainees received general test instructions, followed by directions for recording their answers. For the CAT format, answer revisions were not permitted. Scrap paper, two pencils and an eraser were provided. The experiment was conducted in the KPC Training Department. 20 IBM compatible machines with VGA screen and a standard QWERTY keyboard were rented to administer the CAT tests. After analysis of the data, a report was given to the coordinator of the training programme, followed by a discussion of the key findings. Subjects were thanked for taking part in the experiment and debriefed on the aims of the study and given feedback about their results.

3.2.4 Collection and Analysis of The Data

The following information was collected for each trainee: High School Ratio (HSR), English language level (EL), scores of the subjects during the last year at high school (English, Maths, Physics, Chemistry, Biology, and the total score for all these subjects), and score in the training programme courses of the three Units. Also noted were number of days absent, number of times late, and an overall evaluation. All this information was obtained from the trainee's file at the KPC training department. In addition to these data, scores in the NA, AR, and MR tests for both formats of the DAT were also collected. The confidentiality of all this information was assured.

3.3 Results

The criterion measure used in the validity analysis section of this study was the training course grade, number of days absent, and number of times late. Criterion scores were available for all the trainees who had been tested on the ability tests and who subsequently completed the training. The performance examinations required each trainee to demonstrate proficiency with respect to specific job-related tasks spelled out in the course objectives. Trainee's proficiency was rated from D-- to A+ (13 points with equal interval between grades). Due to inevitable problems of measurement error when trying to estimate a trainee's "true" achievement level in the training course, it cannot be assumed that the available criterion scores were highly reliable (Alken, 1988; Newstead & Dennis, 1994). The dependent variable for this study was performance on the training courses. The independent variables were the NA, AR, and MR test scores and trainee performance at high schooL Correlation coefficients were calculated between variables to test their predictive strength.

A multiple regression was also carried out. Multiple regression is suitable for ascertaining the extent to which criterion variance can be accounted for by a set of predictor variables, and it also enables estimates to be made about each predictor's contribution to the explained variance in the criterion variable. Moreover, t-tests, and analysis of covariance (ANOVA) were also performed, to assess the difference between groups. The tests of normality and homogeneity of variance for each format of the three tests showed that the scores skew of the scores do not differ significantly from the normal distribution, and the variability of scores in each format is approximately the same. A significance level of .05 was used in all analyses. All these tests were carried out using the Statistical Package for the Social Sciences (SPSSPC+).

3.3.1 Equivalence Between The P&P and CAT of DAT

Pearson Correlation coefficients were calculated to assess the correlation between the formats for each test used (Table 3.1). The correlation coefficient between the two formats of NA test was found to be .90 (p<.001). This coefficient is lower but not significantly so than the reliability reported in the DAT Manual for the P&P format (r=.92) (using Fisher's r-to-z transformation). The inter-correlation coefficients between the test and the other tests (P&P format only) are also lower than those reported in the Manual (NA & AR r=.70; NA & MR r=.52). Table 3.1 also shows the correlation of each test with the other tests in both formats. The correlations between formats for the same test ranged approximately from .87 to .90, whereas correlations between different tests ranged approximately from .22 to .48. The shrinkage of these validity coefficients is expected on cross-validation due to restriction of range, sampling error, and criterion reliability.

Table 3.1. Correlations between the P&P and the CAT formats for each test.

Correlations:	NAP	ARP	MRP	NAC	ARC	MRC
NAP		.3909**	.2994**	.8999**	.3763**	.2566*
ARP			.4788**	.2895**	.8738**	.3838**
MRP				.2179*	.4177**	.9028**
NAC					.3494**	.2162*
ARC						.3630**
MRC						

1-tailed Signif: * - .01 ** - .001
NAP=Numerical Ability Test-P&P, ARP=Abstract Reasoning Test-P&P, MRP=Mechanical Reasoning Test-P&P, NAC=Numerical Ability Test-CAT, ARC=Abstract Reasoning Test-CAT, MRC=Mechanical Reasoning Test-CAT.

3.3.1.1 *Numerical Ability Test*

Table 3.2 shows the mean, standard deviation, and number of subjects for each group for NA test. For the first administration, those who took the P&P test first (P&PFG) had almost the same mean as those who took the CAT first (CATFG). T-tests revealed no significant differences between the two groups (t=.20, df= 115.80, p= .843).

For the second administration, the P&PFG showed a higher mean than the CATFG. However, this difference was not significant (t=1.27, df=114.28, p= .208).

Table 3.2. Means, standard deviation, and number of subjects for both groups on the NA test.

	Group	Format	X	SD	N[3]
First Administration	P&PFG	(P&P)	15.6786	5.114	56
	CATFG	(CAT)	15.8710	5.434	62
Second Administration	P&PFG	(CAT)	17.3750	5.252	56
	CATFG	(P&P)	16.1613	5.148	62

P&PFG=Paper & pencil first group; CATFG=Computerised adaptive test first group

ANOVA (Table 3.3) on the raw scores revealed that the main effect of group was not significant (F (1,116)=.29, p=.590). However, the main effect of test format was significant (F(1,116)=12.58, p<.001), where the CAT version produced higher mean scores. Also, a significant results was found for the interaction between group and test format (F(1,116)=.25.11, p<.001). The format effect was bigger for those who took the P&P version first (P&PFG) (Figure 3.1). The lines, in the figure join the first and second administrations of a test format across groups.

Investigating the differences between first and second administration within each group, t-test revealed significant difference for P&PFG (t=5.86, df=55, p<.001), where

[3] The variation in the number of subjects in each cell is because a few subjects did not take all three tests.

the second administration was higher. However, no significant difference for CATFG (t=1.07, df=61, p=.289) was found. The correlation coefficient between the two formats was found to be .91 (p<.001) for P&PFG, and .91 (p<.001) for CATFG, which are high correlations. There were no differences between the two groups on either P&P or CAT formats (t=.51, df=116, p=.61; t=1.53, df=116, p=.130, respectively).

Table 3.3. ANOVA for effects of groups and format on the NA test's scores.

Source of Variation	SS	DF	MS	F	Sig of F
NA					
G	15.35	1	15.35	.29	.590
Format	29.09	1	29.09	12.58	.001 **
G BY Format	58.07	1	58.07	25.11	.000 **

1-tailed Signif: ** - .001; NA=Numerical Ability; G=Group

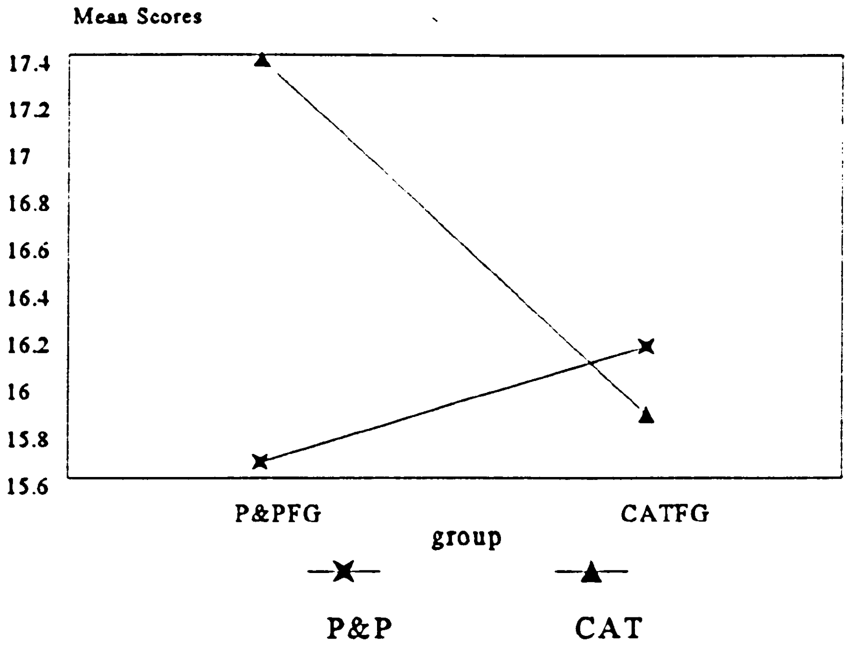

Figure 3.1. Mean scores on NA test on both formats for both groups.

3.3.1.2 *Abstract Reasoning Test*

The correlation coefficient between the two formats of AR test (Table 3.1) was found to be .87 (p<.001). This coefficient is significantly lower (using Fisher's r-to-z transformation) than the reliability reported in the DAT Manual for the P&P format (r=.94). The inter-correlation coefficients between the test and the other tests (P&P format only) are also lower than those reported in the Manual (NA & AR r=.70; AR & MR r=.64).

In the first administration of the test (Table 3.4), those who took the P&P version first (P&PFG) scored lower on average than those who took the CAT version first (CATFG). However, t-tests revealed no significant differences between the two groups (t=1.62, df=108.37, p=.108). For the second administration, the P&PFG again had a lower mean than the CATFG, though the difference was not significant (t=-1.93, df=119.00, p=.056). Investigating the difference between the first and second administrations within each group, t-tests revealed nonsignificant differences for both groups (P&PFG, t=.62, df=120, p=.535; CATFG, t=.62, df=120, p=.535). In both groups, the means for the second administration were higher than for the first administration. The correlation coefficient between the two formats was found to be .90 (p<.001) for the P&PFG and .90 (p<.001) for the CATFG.

Table 3.4. Means, standard deviation, and number of subjects for both groups on the AR test.

	Group	Format	X	SD	N
First Administration	P&PFG	(P&P)	29.2881	5.414	59
	CATFG	(CAT)	31.2742	7.893	62
Second Administration	P&PFG	(CAT)	30.3898	6.438	59
	CATFG	(P&P)	32.7097	6.803	62

P&PFG=Paper & pencil first group; CATFG=Computerised adaptive test first group

ANOVA (Table 3.5) showed no significant effects either for group (F(1,119)=3.31, p=.072) or format (F(1,119)=.31, p=.577). However, the group by format interaction was significant (F(1,119)=25.11, p<.001), where the format effect was bigger for those who took the CAT version first (CATFG) (Figure 3.2). The lines, in the figure, join the first and second administrations of a test format across groups.

Table 3.5. ANOVA for effects of groups and format on the AR test's scores.

Source of Variation	SS	DF	MS	F	Sig of F
AR					
G	280.26	1	280.26	3.31	.072
Format	1.68	1	1.68	.31	.577
G BY Format	97.30	1	97.30	18.11	.000 **

1-tailed Signif: ** - .001; AR=Abstract Reasoing; G=Group

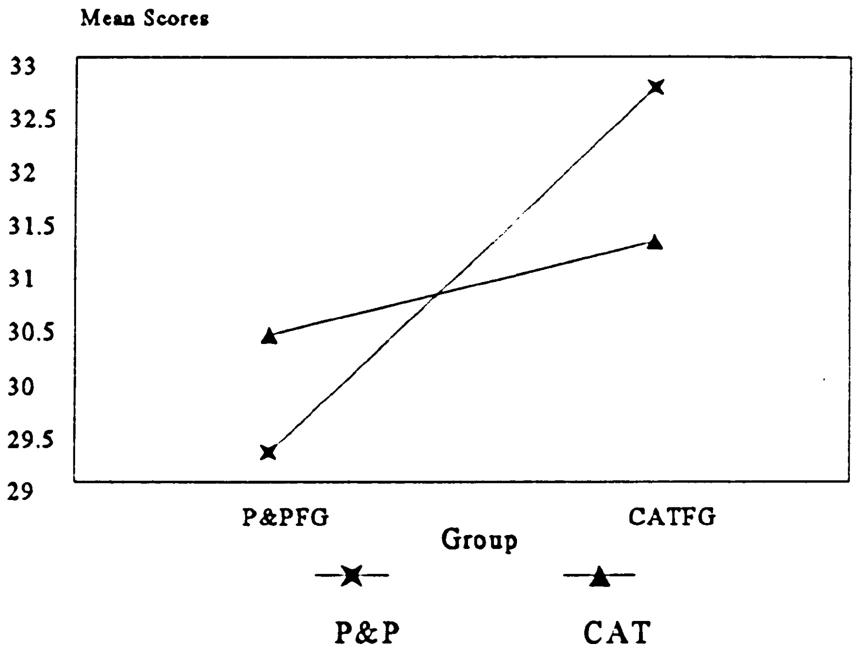

Figure 3.2. Mean scores on AR test on both formats for both groups.

3.3.1.3 *Mechanical Reasoning Test*

The correlation coefficient between the two formats of MR test was found to be .90 (p<.001). This coefficient is higher than the reliability reported in the DAT Manual for the P&P format (r= .89). However, no significant differences were found between the two correlations. The inter-correlation coefficients between the test and the other tests (P&P format only) are also lower than those reported in the Manual (NA & MR r=.52; AR & MR r=.64).

Table 3.6 shows the means, standard deviations, and numbers of subjects for each group. For the first administration, those who took the P&P first (P&PFG) produced a higher mean than those who took the CAT first (CATFG). However, t-tests revealed no significant differences between the groups (t=.50, df=116.57, p=.619). For the second administration, the same result was found (t=.67, df=116.67, p=.506). Investigating the differences between first and second administration, within each group, t-tests revealed nonsignificant differences for both groups (P&PFG, t=.43, df=118, p=.669; CATFG, t=.43, df=118, p=.669). In both groups, the means for second administration were higher than for the first administration. The correlation coefficient between the two formats was found to be .91 (p<.001) for the P&PFG and .91 (p<.001) for the CATFG.

Table 3.6. Means, standard deviation, and number of subjects for both groups on the MR test.

	Group	Format	X	SD	N
First Administration	P&PFG	(P&P)	43.3559	7.586	59
	CATFG	(CAT)	42.6333	8.199	60
Second Administration	P&PFG	(CAT)	44.7627	8.569	59
	CATFG	(P&P)	43.7333	8.262	60

P&PFG=Paper & pencil first group; CATFG=Computerised adaptive test first group

As with the AR test, a ANOVA (Table 3.7) revealed no significant effects either for group (F(1,117)=.36, p<.551) or format (F(1,117)=.24, p=.626). However, the interaction between them was significant (F(1,117)=15.97, p<.001). The format effect was bigger for those who took the P&P version first (P&PFG) (Figure 3.3). The lines, in the figure, join the first and second administrations of a test format across groups.

Table 3.7. ANOVA for effects of groups and format on the MR test's scores.

Source of Variation	SS	DF	MS	F	Sig of F
MR					
G	45.65	1	45.65	.36	.551
Format	1.40	1	1.40	.24	.626
G BY Format	93.47	1	93.47	15.97	.000 **

1-tailed Signif: ** - .001; MR=Mechanical Reasoning; G=Group

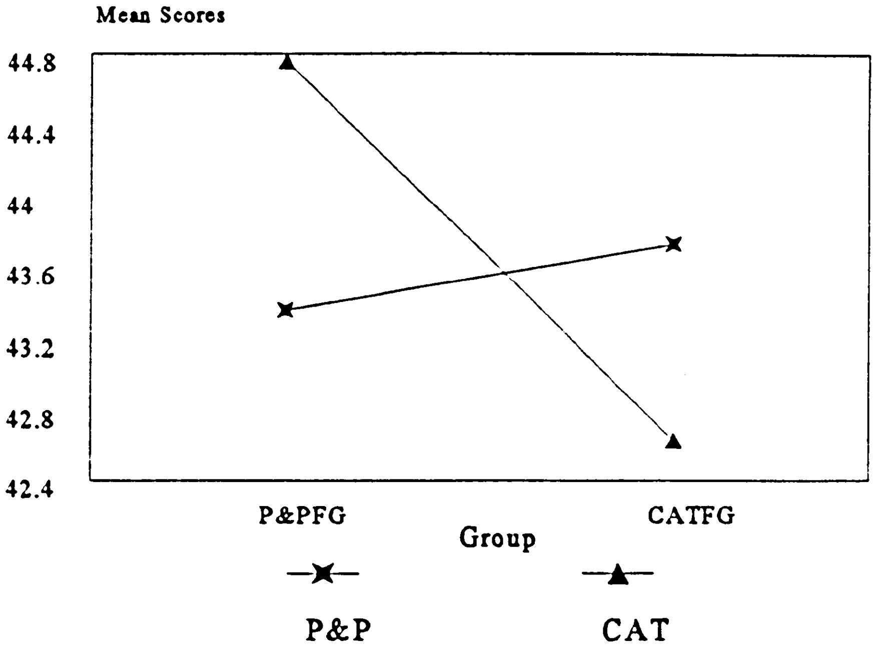

Figure 3.3. Mean scores on MR test on both formats for both groups.

3.3.2 Consistency of Evaluations

Comparing the trainees' performance in the mid and final exams, the performance of trainees was significantly improved for all subjects except in Process Chemistry (Table 3.8). This could be because of drop-out or expulsion of unsatisfactory trainees during the course. Absence also increased significantly, while arriving late for class decreased.

Table 3.8. Mean, standard deviation, number of subjects, standard error, and t-test results for mid and final evaluations for all variables.

	N	M	SD	df	t	p
OE2	118	5.7320	2.767	96	3.12	.002 **
OE1	118	5.1134	2.750			
Org. Ch.2	119	4.9583	2.722	95	3.54	.001 ***
Org. Ch.1	119	3.7083	2.722			
Pro Ch2	117	5.8646	2.940	95	-6.11	.000 ***
Pro Ch1	116	7.8125	2.914			
Eng Sc2	118	6.8125	3.398	95	4.99	.000 ***
Eng Sc1	117	5.1667	3.355			
Pro Ph2	119	5.5729	3.022	95	5.28	.000 ***
Pro Ph1	119	4.0104	2.812			
Math2	119	7.1368	3.817	95	5.02	.000 ***
Math1	119	5.4000	3.760			
Eng Dr2	118	6.5638	3.729	95	2.28	.025 *
Eng Dr1	118	5.5532	3.980			
Absence2	118	1.2259	1.774	95	4.78	.000 ***
Absence1	117	.3814	1.113			
Late 2	119	.6082	1.016	96	-2.08	.040 *
Late 1	119	.9381	1.749			

1-tailed Signif: * - .05 ** - .01 *** -.001; 1&2= 1st. & 2nd. evaluations; Org.Ch=Organic Chemistry; Pro.Ch=Process Chemistry; Eng.Sc=Engineering Science; Pro.Ph=Process Physics; Math=Mathematics, En.g.Dr=Engineering Drawing; Late= lateness; OE=Overall Evaluation.

Table 3.9 shows the correlation between the mid and final evaluations of the training courses. All correlations are positive and significant at the .001 level. Those trainees who obtained a high score on the first evaluation tended also to do so on the second evaluation, and vice versa. The highest correlation was found between the mid and final evaluations.

Table 3.9. Correlation between evaluations in the mid and final evaluatioms.

	Org. Ch.F	Pro. ChF	Eng. ScF	Pro. PhF	MathF	Eng. DrF	AbsenceF	LateF	OEF
Org. Ch.S	.3431**								
Pro. ChS		.4307**							
Eng. ScS			.5421**						
Pro. PhS				.5085**					
MathS					.6033**				
Eng. DrS						.3794**			
AbsenceS							.3442**		
Late S								.4669**	
OES									.7502**

1-tailed Signif: * - .05 ** - .01; F & S= 1st. & 2nd. evaluations. Org.Ch=Organic Chemistry; Pro.Ch=Process Chemistry; Eng.Sc=Engineering Science; Pro.Ph=Process Physics; Math=Mathematics, En g.Dr=Engineering Drawing; Late= lateness; OE=Overall Evaluation.

3.3.3 The Correlation Between High School Performance and Training Performance

Table 3.10 presents correlations between performance at high school and performance on the training programme. The Table shows the correlation coefficients for the five courses taken in the final year at high school (English, Mathematics, Physics, Chemistry, and Biology), the total of these scores, and the high school ratio, with the grades of the courses taken on the training programme. Arabic Language and Islamic Religion courses have been excluded from the analysis. Except for English, the other predictors show positive and significant correlations with Organic Chemistry, Process Chemistry, Engineering Science, Process Physics, Mathematics and Overall Evaluation. All the predictors show poor relationships with Engineering Drawing, absence, and lateness figures. The only exception to this is the correlation between Biology and Engineering Drawing. Engineering Drawing requires new skills which were not taught at high school. The highest correlations were found between total of high school courses and Engineering Science ($r=.607$, $p<.001$).

Table 3.10. Correlations between performance in high school subjects and training courses.

Correlations:	Org. Ch	Pro. Ch	Eng. Sc	Pro. Ph	Math	Eng. Dr	Absence	Late	OE
E	.1281	.1223	.2663*	.1933	.1007	.0594	.1350	.2008	.1942
M	.2690*	.2870*	.4663**	.2870*	.5322**	.1043	.1369	.1580	.3737**
Ph	.3348**	.2300*	.4619**	.3844**	.4364**	.1908	-.0176	.0401	.3924**
Ch	.4320**	.3875**	.4810**	.4168**	.3646**	.1087	-.1129	-.0311	.4376**
B	.4689**	.3840**	.5763**	.4734**	.4566**	.2700*	-.1742	-.0289	.5785**
T	.4262**	.3322**	.6068**	.4601**	.5452**	.1877	.0233	.1079	.5249**
HSR	.3367**	.2609*	.4756**	.3359**	.4122**	.1074	.1060	.1303	.4078**

1-tailed Signif: * - .05 ** - .01; Org.Ch=Organic Chemistry; Pro.Ch=Process Chemistry; Eng.Sc=Engineering Science; Pro.Ph=Process Physics; Math=Mathematics, Eng.Dr=Engineering Drawing; OE=Overall Evaluation; E=English, M=Mathematics; Ph=Physics, Ch=Chemistry; B=Biology; T=Total of E,M,Ph,Ch,B; HSR=High school ratio.

A principal factor analysis of the high school subjects and training courses with varimax rotation was conducted. In determining the number of factors, factors were considered only if they had eigenvalue greater than 1. Variables were considered only when their loadings were greater than .50. The process produced four factors which accounted for 74.8% of the total variance. The factor loadings coefficients are presented in Table 3.11. Factor 1 reflects foundation science (Mathematics, Physics, Chemistry, Biology, High school ratio). This factor accounts for 46% of the variance. Factor 2 reflects Process science and overall evaluation (Organic Chemistry, Process Chemistry, Engineering Science, Process Physics, and Overall Evaluation), and accounts for 15.6% of the variance. Factor 3 involves Engineering Drawing, and accounts for 6.9% of the variance. Factor 4 reflects lack of interest (lateness, absence, and English language). The factor accounts for 6.3% of the variance.

Table 3.11. Factor loadings of the high school subjects and training courses.

	FACTOR1	FACTOR2	FACTOR3	FACTOR4
High school ratio (HSR)	.88			
Physics (PH)	.86			
Mathematics (M)	.84			
Chemistry (CH)	.74			
Biology (B)	.63			
Mathematics in training course (MATH)	.52			
Organic Chemistry (OC)		.81		
Process Chemistry (PC)		.83		
Engineering Science (ES)		.78		
Overall Evaluation (OAE)		.78		
Process Physics (PP)		.76		
Engineering Drawing (ED)			.92	
English (E)				.78
Lateness (LATE)				.69
Absence (AB)				.51

3.3.4 The Correlation Between Psychometric Instruments and Trainees' Performance

Since the DAT was presented to Arabic speakers in its original form, without any translation from English to Arabic, Table 3.12 shows the relationship between English language ability (assessed previously by the Training Department) and performance in each ability test on both formats. The only significant correlations found were those with

Mechanical Reasoning test (P&P format) (r=.27, p<.05), and Abstract Reasoning (CAT format) (r=.22, p<.05).

Table 3.12. Correlations between English language level and performance on ability tests.

Correlations:	NAP	ARP	MRP	NAC	ARC	MRC
EG	.1359	.1977	.2666*	.1652	.2169*	.2029

1-tailed Signif: * - .05 ** - .01; EG= English Language; NAP=Numerical Ability-Paper-and-pencil format; ARP=Abstract Reasoning-Paper-and-penci formatl; MRP=Mechanical Reasoning-format; NAC=Numerical Ability- CAT format; ARC=Abstract Reasoning-CAT format; MRC=Mechanical Reasoning-CAT format.

The average test scores achieved by all trainees in the three ability tests (both formats) are presented in Table 3.13. The mean score of the CAT format of the NA test was significantly higher than that of the P&P format of the same test (t=-3, df=117, p=.003). For the other tests no significant differences were found (AR, t=.62, df=120, p=.535; MR, t=-.43, df=118, p=.669). Comparing the trainees' mean scores on these tests with those of fifth year English boys reported in the DAT manual (the oldest group of non-selected students, and therefore, the most suitable norm found), the trainees compared only moderately well against the comparison group. It is important to note that English is the second language for the trainees, and they do not actually represent Kuwaiti high school leavers, but only those who have been prevented from continuing their higher education because of low grades at high school. Therefore, the comparison results need to be interpreted with care.

Table 3.13. Mean and standard deviation for all trainees on ability tests in both formats.

Variable	Mean	Std Dev	%ile
NAP	15.93	5.12	45%ile
NAC	16.58	5.38	45%ile
ARP	31.04	6.37	30%ile
ARC	30.84	7.20	30%ile
MRP	43.55	7.90	30%ile
MRC	43.69	8.42	30%ile

NAP=Numerical Ability Test-P&P, ARP=Abstract Reasoning Test-P&P, MRP=Mechnical Reasoning Test-P&P, NAC=Numerical Ability Test-CAT, ARC=Abstract Reasoning Test-CAT, MRC=Mechanical Reasoning Test-CAT. %ile= Comparing the trainees' mean scores on these tests with those of fifth year English boys reported in the DAT manual

The key results of the study, i.e. the correlations between training programme performance and the psychometric instruments, are shown below in Table 3.14. The final scores for each course were used for the analysis. As far as the ability tests were concerned, the three tests showed a number of significant correlations with training courses. Of the three tests, the NA test in both formats correlated significantly with all the training courses and with overall evaluation. However, the correlation coefficients for the CAT format were higher than those of the P&P format, except for Engineering Drawing.

The AR test correlated poorly with overall evaluation and with all training courses except Engineering Drawing, where the correlation coefficient of the CAT format was lower than that of the P&P. The MR test in both formats significantly correlated with

Engineering Drawing (both formats got almost the same correlation coefficient) and with overall evaluation (the CAT format correlated more highly than the P&P format), but this test correlated with Process Physics in the CAT format only. No predictor correlated significantly with either absence or lateness. It should be noted that these correlations have not been corrected for attenuation in the criterion, nor for restriction of range, because the reliability of the criterion and the standard deviation of all applicants to the training programme were not available.

Table 3.14. Correlations between performance on ability tests and training courses.

Correlations:	Org. Ch	Pro. Ch	Eng. Sc	Pro. Ph	Math	Eng. Dr	Absence	Late	OE
NAP	.2799*	.2719*	.3120**	.2234*	.3907**	.3007**	.1702	.0917	.3672**
NAC	.3095**	.3204**	.3766**	.2529*	.4066**	.2927**	.1345	.1027	.4098**
ARP	.1028	-.0374	.0883	.1014	.1601	.3886**	.0716	.1464	.1917
ARC	.1227	-.0001	.1489	.1228	.1846	.3777**	.0416	.1476	.2119
MRP	.0594	.0376	.1836	.1916	.1639	.4061**	.0536	.0356	.2695*
MRC	.1068	.1091	.1981	.2681*	.2059	.4020**	-.0062	-.0915	.3109**

1-tailed Signif: * - .05 ** - .01; Org.Ch=Organic Chemistry; Pro.Ch=Process Chemistry; Eng.Sc=Engineering Science; Pro.Ph=Process Physics; Math=Mathematics, Eng.Dr=Engineering Drawing; OE=Overall Evaluation; NAP=Numerical Ability Test-P&P, ARP=Abstract Reasoning Test-P&P, MRP=Mechanical Reasoning Test-P&P, NAC=Numerical Ability Test-CAT, ARC=Abstract Reasoning Test-CAT, MRC=Mechanical Reasoning Test-CAT.

3.3.5 Multiple Regression

Forced entry multiple regression (Tables 3.15 & 3.16) was used to examine differences in the slopes and intercepts of regression lines, in order to determine the effects of HSR, and the NA, AR, and MR tests for both formats on the relationship between the performance predictor variables and the overall performance criterion (OAE) in the training course. As shown by the regression coefficients (B) and the standardized regression coefficients (Beta) weight, for both formats, the largest contribution to the relationship was made by HSR, followed by NA, MR and lastly AR. Except for AR for both format, taking account of all the other variables, all variables contributed significantly to the equation (p<.01).

Table 3.15. Final regression equation using CAT and P&P formats.

Variable	CAT		P&P	
Entered	B	Beta	B	Beta
HSR	.266	.334	.272	.342
NA	.284	.288	.256	.246
MR	.149	.237	.149	.222
AR	-.032	-.043	-.065	.097
	constant = -16.501		constant = -15.208	

HSR= High School Ratio; NA=Numerical Ability; MR=Mechanical Reasoning; AR=Abstract Reasoning

For the CAT format (Table 3.16), the overall multiple correlation coefficient of .561 is statistically significant at the .001 level, and the final equation for the above variables indicates that 31.5% of the OAE variance is accounted for by the equation. Elimination of the AR test does not significantly lower the R^2 , while all the remaining variables made significant contributions to R^2. An adjusted R^2 value was used to determine the reduction in R^2 which would be expected if the partial regression weights were used with different samples. The adjusted R^2 produced by the equation for all variables shrank slightly (from .315 to .290). If P&P format of the same tests were entered in the final equation instead of the CAT formats we will find that R^2 dropped to .279 (Table 3.16).

Table 3.16 Summary of multiple regression using CAT and P&P formats.

Regression Analysis	Variables Entered		R	R^2	Omitted Variables	% Increase in R^2 when omitted variable is added
No. 1	MR+HSR+NA+AR	*CAT*	.561	.315		
		P&P	.528	.279		
No. 2	MR+HSR+NA	*CAT*	.559	.313	AR	.15
		P&P	.524	.275	AR	.42
No. 3	HSR+NA+AR	*CAT*	.517	.267	MR	4.79
		P&P	.492	.242	MR	3.67
No. 4	MR+HSR+AR	*CAT*	.494	.244	NA	7.11
		P&P	.479	.229	NA	5.01
No. 5	MR+NA+AR	*CAT*	.457	.209	HSR	10.62
		P&P	.411	.169	HSR	11.03

HSR= High School Ratio; NA=Numerical Ability; MR=Mechanical Reasoning; AR=Abstract Reasoning

Lastly, to estimate the confidence intervals for a particular predicted grade using the CAT format, it was found that the maximum predicted OAE: Max. POAE = 11.373 + 1.085 (POAE), and the minimum is Min. POAE= 10.414 + .881 (POAE). This interval is the 95% confidence interval. This step helps selectors to avoid possible risks by selecting those candidates whose minimum POAE is above the minimum satisfactory performance (in this case a C grade).

3.4 Discussion and Conclusion

This study has evaluated two main issues associated with the use of CAT for selection and assessment. The first issue is the equivalence of P&P and CAT formats of the DAT. The study showed equivalence between the two modes for the AR and MR tests, but failed to do so for the NA test. In order for the CAT version of the NA test to be equivalent with its P&P version, then one set of the scores (either P&P or CAT scores) needs to be rescaled to be comparable with scores from the other test. It could be easier to rescale the CAT scores to maintain the P&P norms. The study confirms Henly et al. (1989) demonstration of essentially equivalence between the two modes of testing for the AR and MR but not for NA test, showing an evidence for the structure similarity of both versions.

With all tests, the mean score for the second administration was higher than for the first, suggesting that practice effects were taking place. This effects may occur as a result of remembering the same item used, instructions, and strategy used in the first administration. Job selection decision makers should be aware of such an effect, to avoid any bias in their selection process. Anastasi (1988) suggested that test publisher should indicate in their test manual the expected gains on a retest with a parallel form of the test.

The second aim of this study was to assess the predictive usefulness of using CAT to forecast training programme performance. Significant and positive relationships were found between the NA and MR tests and the overall evaluation of trainee performance. However, this result was not found with the AR test. The CAT format displays a marginally higher correlation with trainee performance than its P&P counterpart. The increase in correlation coeffientient did not exceed .05 in all cases. High school ratio showed a stronger relationship with trainee performance (r=.56) than the tests used. The study supports other findings (e.g. Cudeck, 1985; Kent & Albanese, 1987; Koch et al., 1990; Sympson et al., 1982, 1984; McBride & Martin, 1983; Moreno et al., 1984; Sand & Gade, 1983; Moreno et al., 1985) suggesting that CAT yields a similar predictive validity to its P&P counterpart.

The study shows that high school ratio is a valid and important predictor of training performance. HSR accounted for most of the predictable variability in the overall evaluation of trainee performance. Adding the CAT format of the MR and NA tests to the analysis significantly improved the prediction over that provide by P&P format, whereas the AR test of either formats does not contribute any further predictive information in estimating the criterion variable. It could be useful in the future to test other possible predictors which might enhance our forecasting, such as a Spatial Reasoning test, personality questionnaires, interest inventories and other useful biodata.

Given the moderate correlations between the NA and MR tests (see Table 3.1), and their reasonable correlations with the predicted variable (Table 3.13), it is likely that they are measuring somewhat different intellectual aspects of the training programme and so they ought be retained in any subsequent selection procedure. Which format to use in future selection and assessment should be considered in the light of a number of factors; for example, cost of materials, number and quality of applicants, number of available vacancies, speed of decision making required, and ease of administration. The advantages and limitations mentioned in the previous chapters of using CBT in general, and CAT in particular, should also to considered.

Utility theory provides a mean of associates information concern the test validity, the situation in which tests will be used, and the significance of the good or poor decisions which may result from the test use. We can estimate the cost benefits of selecting candidates using Cronbach Utility Formula (Cronbach & Glaser, 1965) which provides an equation by which the cost benefit return of validated test can be estimated. The equation states:

$$\text{gain in dollar or pounds} = r_v \,.\, Z_x \,.\, SD_y \,.\, N - C$$

where
r_v = validity coefficient
Z_x = mean Z score on the selection test(s) of the selected candidates
SD_y = standard deviation of profit on the job performance criterion in dollar or pounds
N = number of selected applicants
C = cost of the testing programme

In our case, KPC recruits 122 high school leavers a year, for this training programme alone, from 300 applicants. The total first year salary is approximately 12,000 pounds per applicant. The cost of testing using P&P format (NA and MR tests only, no cost for obtaining HSR) is estimated at 9.40 pounds per head, giving a total cost of 2820 pounds. The multiple correlation coefficient of the tests and HSR is .528. If we assume that the company recruits applicants on average one standard deviation above the mean and that the SD_y of the criterion is 40% of salary (Schmidt et al., 1979), i.e. 4800 pounds, then we will find that the utility or the gain is about:

Gain in pounds = .528 x 1 x 4800 x 122 - 2820 = 306,376 pounds

On these assumptions, the utility or gain using P&P formats (plus HSR) is about 306,376.80 pounds per annum. If the company use the CAT format instead of the P&P one, we will find that the gain will rise to about 325,951 pounds (r_v = .561; C = 1170 pounds plus 1400 pounds for the cost of renting the 20 machines, total of 2570 pounds). The gain from using CAT format above that of P&P is about 19,574 pounds. If the computers were available in site then the profit will be 20,974 pounds.

All these calculations assumed that the company will continue using the HSR. However, if using only the P&P format, the gain will be 237,861 pounds, and for CAT formats alone the gain will be 265,049 pounds assuming the computers are rented and 266,449 pounds if they are available in the company. That is, the gain of using CAT alone over that obtained from P&P alone is 27,187 pounds assuming the computer are rented, and 28,587 pounds if they are available in the company. With using the P&P format a lone for selecting the 122 trainees, the cost is only 1:84 of the gain per year, and for the CAT format the ratio is 1:103 (assuming the computers are rented) and 1:227 if they are available in the company. In our case, the selection pays for itself 84 to 227 times over in the first year.

However, While Cronbach Utility Formula remains generally useful for estimating the cost benefits of selection, it has its limitations. The Formula does not account for variable costs, taxes, discounting, more than one year hiring, turnover, absenteeism, costs associated with incorrect rejection decisions, and cost of training. The 40% of salary estimate proposed by Schmidt et al., (1979) for SD_y is also subjective and conservative. Moreover, the Formula is assumed the stability of validity and performance variability over time, and that all applicants received the same selection treatment, the matter which is open to question (Janz, 1989; Sackett, 1989). Some modifications have been suggested by Boudreau (1983) to take account for some of the above limitations.

Behaviour is a function of the interaction between a trainee and his environment. The performance of a trainee is a product not only of what he brings with him to the training programme, but also of his experience of, and interaction with that situation. If we seek accurate and useful predictions, we should therefore analyse the trainees together with their training environment and examine the results of their interaction. On this basis, our prediction should include more than aptitude and personality factors. The Training department should start developing critical variables which allow the evaluation of each trainee in relation to his own environment. Thus, any attempt to generalise these result to other situations should be made with caution.

More importantly, while ability tests and high school ratio are known to provide a valid basis for admission decisions to a training programme, their use should be continuously monitored by the training department and institution which habitually rely on them, and routine validity studies should continue.

In conclusion, This study shows that CAT can predict a performance variable as accurately as can the P&P format. It appears that the P&P and CAT formats of the AR and MR tests are equivalent. And that one of the formats of the NA test needs to be rescaled for the test to be equivalent. The possible reasons for such a difference will be investigated in the next experiment. In particular, the next study will investigate the effects of computer anxiety, reactions to and experience with computers, and the time taken to administer each format, as well as what subjects think about the use of computerized tests for job selection and how they perceive the difficulty levels of the adaptive test. Sex differences will also be examined. The next study investigates these questions, along with other related issues.

CHAPTER 4

CHAPTER 4

Experiment 2

Testing Time and Testees' Reaction and Anxiety to Computerized Adaptive Testing for Selection Purposes.

4.1 Introduction

Although the previous experiment showed that CAT has similar predictive validity to its P&P counterpart, it failed to prove equivalence between the P&P and CAT formats of the NA test of DAT. It showed that the rank orders of scores of individuals tested in two modes closely approximated each other, but still there was a significant difference in the means of the two formats. This experiment tries to investigate possible reason for such inconsistency, along with some other important issues.

The two main features that distinguish the CAT form from the standard P&P form are computerized administration and adaptive selection of test items according to subjects' performance. In the first experiment trainees took the NA test first, then the AR and lastly the MR test. The effect of computer anxiety, if it exists, on performance may be more effective in the first stage of the testing session than in the subsequent stages. Research has shown that computerized testing increases an examinee's level of test anxiety and hence decreases performance, especially for those with relatively little computer experience (Hedl et al., 1973; Llabre et al., 1987; Lee, 1986). This may affect a test's equivalence (Hofer & Green, 1985). However, Dimock & Cormier (1991) found no evidence that format differences between P&P and CBT was influenced by either the testee's level of computer experience or anxiety.

It is important to distinguish between anxiety arising from a test per se and anxiety emerging from using a different format. The majority of the questionnaires used to measure the effect of computerized administration on an examinee's anxiety have been designed primarily to measure test anxiety or state anxiety, for instance, the State-Trait Anxiety Inventory (STAI) developed by Spielberger et al. (1983) (see e.g. Dimock & Cormier, 1991; Hedl et al., 1973) and TAS developed by Sarason (1980) (see e.g. Awdah, 1988; Llabre et al, 1987) rather than computer anxiety. These measures may be unable to show clearly how much variance is attributable to the computerized format. Computer anxiety scales could be more suitable with computerized tests, but computerized anxiety scales may not be suitable for P&P test format.

Moreover, the subject's attitude could affect test performance and hence affect both test validity and reliability (Hofer & Green, 1985; Rosen & Sears, 1987). The subject's reaction has been found to be one of the factors which affects the equivalence between P&P and CBT test formats. Meier and Lambert (1991) review the research conducted in a variety of settings investigating the examinee's reaction to computerized testing. They concluded that negative psychological reactions slow both the acceptance and the useful application of computers. Studies attempting to study reactions to computers have reached contradictory findings and have found a range of positive and negative attitudes (Weinberg & English, 1983; Rosen & Sears, 1987; Klinger et al., 1977; Moe & Johnson, 1988; Garrison & Baumgarten, 1986).

The area requires more investigation in order to clarify some important questions. For example, does having taken a computerized test previously affect the testee's score? Is there any relationship between a subject's test performance, experience with computers, test anxiety, and perceived difficulty of interacting with a computer? What do subjects think about using computerized tests for job selection and how do they perceive the difficulty levels of the adaptive test? This study investigates some of these questions, along with other related issues. It is the concern of this study that as computerized tests continue to be used in different settings by a wider range of people of various ages and levels of computer experience, the need for both more accurate measure and better understanding of subjects' attitudes to computers is increasing in order to develop more efficient and friendly machines and software (Burke et al.,1987).

Most of our accumulated information about the possible factors responsible for the differences between P&P and CAT formats came from studies comparing P&P with computerized based test formats (CBT) of a test. Although one may expect that some sources of difference could be generalized to both CBT and CAT formats, such as computer anxiety, unfamiliarity with computers, display layout, and format differences for recording responses, one wonders whether other untested factors such as examines' awareness that this test was tailored to their performance (a more difficult item coming after a correct answer and an easier item after an incorrect answer) have any effect on their performance and hence on the equivalence of the test.

The aims of this experiments were: 1) to checking the equivalence between the P&P and CAT modes of the NA test of DAT; 2) to assess the effects of computer anxiety on performance at the CAT using a questionnaire especially designed for this purpose, and to investigate whether anxiety is related to unfamiliarity with computers; 3) to examine attitudes towards the CAT testing process; 4) to investigate gender differences; 5) to assess the time taken to administer each format; and lastly, 6) to determine whether subjects notice that the test is tailored to their ability level and whether this affects their performance.

4.2 Method

4.2.1 Examinees

All 81 participants were from a sign-up sheet circulated in different psychology classes at Nottingham University Psychology Department. There were 39 male and 41 female students, ranging in age from 18 to 36 years, with a mean age of 23.19 years (SD=4.92). They were divided into 57 undergraduate and 23 postgraduate students, and the major subject of 66% of them (n=53) was psychology, whereas for the remaining 34% (n=27), major subjects varied. All subjects' first language was English.

4.2.2 Materials and Instruments

Differential Aptitude Tests (DAT)

See previous chapter.

Computer Aversion Scale (CAVS)

The computer anxiety measure used in this study is a self-report thirty-one-item questionnaire of the true-false Computer Aversion Scale (CAVS) which is based on social learning theory (Meier, 1988). A high score indicates high anxiety level. The CAVS produces four scores: (1) Efficacy Expectations for Computers, a sum of items assessing beliefs about whether or not one can perform the behaviours needed to work with a computer; (2) Outcome Expectations for Computers, a sum of items assessing beliefs about whether or not one knows what behaviours are needed to operate a computer; (3) Reinforcement Expectations for Computers, a sum of items assessing beliefs about whether or not outcomes produced by computer use meet one's own goals; and (4) Total Score, a sum of all items reflecting the cumulative effects of reinforcement, outcome, and efficacy expectations for computers. Items include "Computers have no place in my profession" and "Computers are often more enjoyable to work with than people."

Using 270 American undergraduate students to assess its reliability and validity, Meier found that the CAVS correlated significantly (r=-.53, p<.001) with the Attitudes Toward Computers Scale (ATCS) (Rosen et al.,1987) and poorly with the Social Desirability Scale (r=.06, p<.54) (Crowne & Marlowe, 1964). Alpha coefficients were found to be .89 for Total Score, .80 for Efficacy Expectations for Computers, .81 for Outcome Expectations for Computers, and .74 for Reinforcement Expectations for Computers. Therefore, this measure was chosen as a valid and reliable measure of computer anxiety.

Computer Attitude Questionnaire (CAQ)

This questionnaire, which contains 18 items, was designed to assess familiarity with computers, attitude toward computers, attitudes towards the testing process (instructions, screen, testing time, etc.) preferences in test administration, subjects' awareness of the adaptive nature of the test, and their evaluation of the test difficulty (Appendix A). The testees were asked to reveal their attitude toward aspects of computerized testing in comparison with P&P administration.

Moreover, biographical data such as age, sex, nationality, first language, main department, and educational status are obtained from all subjects (Appendix B).

4.2.3 Procedure

Subjects were randomly assigned to one of two experimental groups, which differed in the order of presentation of the forms. The first group (P&PFG, n=39) received the P&P format first, then the computerized format, the second group (CATFG, n=41) received them in the reverse order (Appendix D). The typical delay between the first and second session was eight to eleven days. Both formats were administered according to the instructions in their manuals. The experiment was conducted in individual cubicles and subjects were advised to call the examiner if help was needed. All subjects received general test instructions followed by directions for recording their answers on the answer sheet or the computer screen. Answer revisions were not permitted in the CAT format. For P&P administration, answer sheet, question booklet, scrap paper, rubber and two pencils were provided, whereas scrap paper, two pencils and an eraser were provided for the CAT version.

Completion time was recorded. Before starting the first session, the CAVS was presented, whereas the CAQ was presented immediately after the CAT session. Both questionnaires were presented in a written format.

The CAT tests were administered using a SX3U1 Notebook IBM compatible PC, with LCD monochrome screen, and a standard QWERTY keyboard. After the second session, subjects were thanked for taking part in the experiment, debriefed on the aims of the study, and given feedback about their results and given £5 for their co-operation.

4.2..4 Collection and Analysis of the Data

The dependent variables for this experiment were subjects' scores on the P&P test, the CAT test, CAVS, CAQ, and testing time. The independent variables were test administration format (P&P and CATs formats), and order of test presentation. There were also classification variables of anxiety (high and low) and sex.

The following data were collected from each subject: P&P score, CAT score, testing time, and computer anxiety score (CAVS), beside their responses on the computer attitude questionnaire (CAQ) and the biographical data sheet. All data were transferred to a computer data file and analysed using the Statistical Package for the Social Science (SPSSPC+).

4.3 Results

To assess the relationship between the various variables used in this study, all data were analysed using paired-samples t-tests, analysis of variance (ANOVA), Pearson's product-moment correlations, frequencies, cross-tabulations, and Chi-square. The tests of normality and homogeneity of variance for each format showed that the skews of the scores did not differ significantly from the normal distribution, and the variability of scores in each format was approximately the same.

4.3.1 Equivalence Between P&P and CAT Formats of DAT

The Pearson Correlation Coefficients were calculated to assess the correlation between both formats. The correlation coefficient between the two formats of NA was found to be .75 (p<.001), Significantly lower (using Fisher's r-to-z transformation) than reliabilities reported on the DAT Manual for P&P format and correlation between the two formats found in the first experiment (r=.92, .90 respectively). The correlation coefficient between the two formats was found to be .86 (p<.001) for P&PFG, and .67 (p<.001) for CATFG. The correlation coefficient was significantly lower for the CATFG.

Table 4.1 shows the means, standard deviations, and number of subjects in each group. For the first administration, those who took the P&P first (P&PFG) obtained lower means than those who took the CAT first (CATFG). T-tests revealed significant differences between the two groups (t=4.26, df=38, p<.001). The same was found in the second administration, however, there were no significant differences (t=.91, df=38, p<.366). Investigating the differences between first and second administration within each group, t-test revealed significant difference for the P&PFG (t=6.87, df=38, p<.001), where second administration (CAT) was higher. However, no significant difference was found for the CATFG (t=.27, df=40, p=.790) .

Table 4.1. Means, standard deviation, and number of cases for both groups on NA test.

	Group	Format	X	SD	N
First	P&PFG	(P&P)	27.92	6.86	39
Administration	CATFG	(CAT)	32.85	5.36	41
Second	P&PFG	(CAT)	31.87	6.75	39
Administration	CATFG	(P&P)	33.05	5.97	41

P&PFG=Paper & pencil first group; CATFG=Computerised adaptive test first group

ANOVA on the raw scores revealed that the effects of different groups were not significant (F(1,76)=1.04, p=.310). Also significant difference were found for test format, where scores in CAT were higher than those of P&P (F(1,76)=20.03, p<.001). The effect of sex differences was significant (F(1,76)=15.41, p<.001), where males obtained higher mean scores (see Table 4.2 and figure 4.1).

No significant differences were found for either the interactions between group and sex (F(1,76)=1.03, p=.314), and the interaction between sex and test format (F(1,76)=.80, p=.375). However, the interaction between group and test format was significant (F(1,76)=15.77, p<.001), the format effect being bigger for those who took the P&P first (P&PFG). The interaction between group, sex and test format was not significant (F(1,76)=3.80, p<.055).

Table 4.2 . ANOVA for effects of groups and format on NA.

Source of Variation	SS	DF	MS	F	Sig of F
Group	60.86	1	60.86	1.04	.310
Sex	898.43	1	898.43	15.41	.000 **
Format	168.86	1	168.86	20.03	.000 **
Group x Sex	60.01	1	60.01	1.03	.314
Group x Format	132.99	1	132.99	15.77	.000 **
Sex x Format	6.72	1	6.72	.80	.375
Group x Sex x Format	32.02	1	32.02	3.80	.055

1-tailed Signif: ** - .001

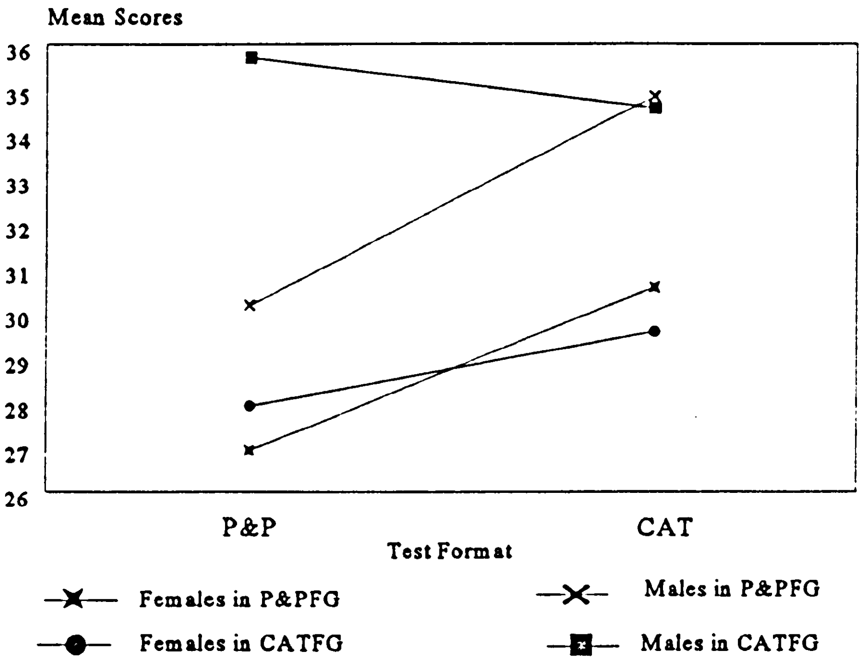

Figure 4.1 Means scores for NA test on both formats for sexes in both groups.

4.3.2 Testing Time

The response time was computed for a given trainee in both formats as the total time elapsed from the start of the test (exclusive of instructions and sample question) to the end of the test. Concerning the P&P version of DAT, the mean testing time was 23:41 min. (SD=5.26, Min.=11:22, Max.=30:00). The mean for CAT was 18:57 (SD=8.15, Min.=7:43, Max.=54:18). The CAT time was found to be 20% shorter than that of the P&P test. The difference in completion time was found to be significant between the two test formats (t=4.93, df=79, p<.001). Also, significant differences were found between different groups in completion time of CAT (t=2.03, df=71 p<.046), where CATFG took a longer time (X=20:53, SD=9.25) than P&PFG (X=16:91, SD=6.41). A significant negative correlation was found between CAT scores and time taken to complete it (r=-.266, P<.05), suggesting that those who obtained higher scores in CAT spent a shorter time completing the test than those who obtained lower scores. The correlation between the P&P test scores and completion time was not calculated because it is a fixed time test. Most of the subjects took almost all the pre-specified time for completing the P&P test (30 min.).

4.3.3 Computer Anxiety

Investigating whether there is any correlation between computer anxiety and CAT scores, no significant correlation was found between CAVS scores and CAT (r=-.16) (no significant differences in CAT scores were found between high and low anxiety groups (median split)) (t=1.11, df=78, p=.272). However, significant gender differences were found. Females had higher scores on CAVS than males (t=2.82, df=78, p<.006). However, no significant difference in CAT score was found between high anxiety females and low anxiety females (median split) (t=1.16, df=39, p=.253). Means and standard deviations on the CAVS for both sexes are shown in Table 4.3. Comparing between groups, t-test revealed no significant differences on CAVS scores between P&PFG (X=10.56, SD=5.99) and CATFG (X=8.88, SD=5.43) where t=1.32, df=78, p<.191).

Table 4.3. Means and standard deviations on the CAVS for both sexes.

	N	X	SD
Males	39	7.92	5.64
Females	41	11.39	5.36

4.3.4 **Attitudes Towards the Computer Adaptive Test**

Table 4.4 shows subjects' attitudes towards computers as the percentage choosing a certain response (testees were allowed to choose more than one choice). Overall, examinees' attitudes toward computerized testing were positive. 50% found the idea of administering tests on a computer fun, while 2.5% found it boring and 12.5% threatening.

When asked about the things they particularly liked about the computerized testing more than half of the subjects (56%) stated the potential for quick feedback, and 41% stated the clarity and simplicity of method. 42% liked it because of the lack of time pressure, 35% because it required less time to accomplish, and 22.5% because it was less fatiguing and required less effort.

On the other hand, when they were asked about the things particularly disliked about the computer, the most common response was the inability to go back and review their answer (80%). Some subjects complained of physical problems during the test session. The eyes of 17.5% subjects got tired, 8.8% found it tiring, and 2.5% got a headache. 16.3% of the subjects experienced some problems in reading the screen, while 11.3% faced difficulties in adjusting to the method.

When asked how difficult they found the questions, females found them significantly more difficult than males (t=3.45, df=78, p<.001). Subjects' Attitudes were not affected by group except in one situation. Those who took the P&P format first (P&PFG) found interacting with the computer more difficult than those in the other group (CATFG) (t=2.15, df=78, p<.035).

Table 4.4 . Percentage responding to each choice.

Q7. How do you find the idea of administering tests on a computer?

50%	1. Fun
2.5%	2. Boring
18.8%	3. Can't decide
12.5%	4. Threatening
7.5%	5. Interesting
3.8%	6. Novel

Q8. The things I particularly liked about the computerized test are:

35%	1. Required less time
41.3%	2. Clarity and simplicity of methods
42.5%	3. Lack of time pressure
56.3%	4. Potential for quick feedback
22.5%	5. Less fatiguing and less effort
2.5%	6. User friendly

Q9. The things I particularly disliked about the computerized test are:

80%	1. I could not go back and review answers
16.3%	2. Problem in reading the screen
11.3%	3. Difficulty of adjusting to the method
8.8%	4. I found it tiring
17.5%	5. My eyes got tired
3.8%	6. No examiner to ask during the test
2.5%	7. Got headache

Table 4.5 shows the Pearson's correlation coefficients between the 13 different variables. As shown, preference for taking a computerized test (PRF) correlates negatively with subject's age (Age) (r=-.27, p<.05) and CAVS scores (CAS) (r=-.27, p<.05).

Moreover, preference for taking a computerized test (PRF) correlates negatively[1] with anxiety levels cause by taking computerized tests (ANX) (r=-.42, p<.001), the difficulty in reading questions from a computer screen (READ) (r=.33, p<.05), and the difficulty of interacting with the computer (INTER) (r=.34, p<.001). Preference for the computer compared with P&P format correlates positively with the possible advantages of computerized tests over P&P tests for job selection purposes (NAD) (r=.26, p<.05).

Believing in the advantages of computerized tests for job selection (NAD) correlates negatively with the perceived easiness of correcting answer (CORR) (r=-.27, p<.05). Interestingly, CAT scores do not correlate with any single variable in the Table except with P&P scores (P&P) as previously mentioned (r=.75, p<.001). CAVS scores (CAS) correlate also negatively with difficulty in interacting with computers (INTER) (r=-.37, p<.001).

Concerning whether taking a computer-administered test affects subjects' anxiety level compared with taking P&P test (ANX), Table 4.5 shows a positive correlation with the difficulty of reading questions from a computer as opposed to reading them in a P&P form (READ) (r=.54, p<.001), what the subject thinks about the sufficiency of time given to give his/her answer (TIME) (r=.26, p<.05), and with the perceived difficulty of interacting with computer (INTER) (r=.51, p<.001).

[1] The signs in the Table have been corrected to reflect the actual direction of the relation between two variables.

The difficulty of reading questions from a computer screen compared with reading them from P&P form (READ) correlates positively with the difficulty of interacting with the computer (INTER) (r=.54, p<.001).

Table 4.5. Correlation matrix for all variables.

	AGE	ANX	INST	READ	TIME	CORR	INTER	HO	NAD	PRF	CAT	CAS
AGE												
ANX	-.1728											
INST	-.1730	.0470										
READ	-.2053	.5352**	.2343									
TIME	.1076	.2629*	.0221	.1335								
CORR	.0779	.0177	.1924	.2036	.1042							
INTER	-.1347	.5136**	.1172	.5367**	.1741	.2219						
HO	.0768	.2277	-.1157	.0951	-.1078	-.1505	.2420					
NAD	-.2186	.0683	.1547	.2053	.1219	-.2684*	.1775	.2231				
PRF	-.2739*	.4191**	.2213	.3341*	.1974	-.0011	.3440**	.2974*	.2605*			
CAT	-.2277	-.0213	-.0322	-.0096	.1166	.1655	.2131	.0707	-.0165	-.0039		
CAS	-.1020	-.2485	.1085	-.2292	-.0481	-.1238	-.3654**	-.4745**	-.1683	-.2675*	-.1621	
P&P	-.2587	-.0436	-.1569	-.0966	.0048	.1467	.1056	-.1154	-.2367	-.1895	.7493**	-.2049

1-tailed Signif: * - .01 ** - .001

AGE=Age; ANX=Anxiety; INST=Instruction Difficulty; READ=Reading the screen; TIME=Sufficiency of testing time; CORR=Making correction; INTER=Interacting with the computer; HO=Previous computer usage; NAD=Computer advantages for job selection; PRF=Format preferred; CAT=CAT score; CAS=CAVS score; P&P=Paper-and-pencil Score.

A principal factor analysis of the attitude data and testees' scores in NA and CAS with varimax rotation was conducted. In determining the number of factors, factors were considered only if they had an eigenvalue greater than 1 . Items were considered only when their loadings were greater than .50. The process produced five factors which accounted for 70% of the total variance. The factor loadings coefficients are presented in Table 4.6. Factor 1 reflects general difficulties with CAT. This factor accounts for 23.2% of the variance. The items included here seem to fit into the categories of difficulty in reading from computer screen, difficulty of interacting with computer, anxiety about the computer, and preference for P&P format over CAT format. Factor 2 reflects numerical ability and accounts for 15.8% of the variance. Factor 3 involves ease and familiarity with computer, and accounted for 12.1% of the variance. This factor includes testees' scores in anxiety questionnaire (CAS) and length of use of computer. Factor 4 reflects the belief in the advantage of answer correction. The factor accounts for 10.8 % of the variance. The items included here seem to include the possibility of making corrections and the advantages of computers for job selection. The last factor (factor 5) accounted for 8.1% of the variance and reflects the belief that testing time was sufficient.

Table 4.6. Factor loadings of the Computer Attitude Scales, CAT, P&P, and CAS.

	FACT1	FACT2	FACT3	FACT4	FACT5
Reading the screen (READ)	.80				
Interacting with the computer (INTER)	.73				
Anxiety (ANX)	.68				
Format preferred (PRF)	.64				
CAT		.89			
P&P		.92			
Previous computer usage (HO)			78		
CAVS score (CAS)			-.75		
Making correction (CORR)				77	
Computer advantages for job selection (NAD)				-.69	
Sufficiency of testing time (TIME)					.89

4.3.5 Previous Computer Experience

Although the subjects were not selected for their previous computer experience, 97.5% of them had used a computer before. Of those, 68.8% had played games on computers, 72.5% had completed one or more computer class or short course, 45% had written computer programs, 53.8% had used a computer at home, and 91.3% had used a computer at their university, school, or workplace.

Table 4.5 shows a significant negative correlation between prior computer use (HO) and scores on the Computer Anxiety Scale (CAVS) (CAS) (r=-.48, p<.001), and a significant positive correlation with preferences in using computerized tests (PRF) (r=.30, p<.05).

Moreover, 48.8% of the subjects had taken a test on a computer before, whereas 45% had not, and a mall proportion (6.3%) did not remember. Concerning their anxiety, although those who had taken computerized tests before obtained lower anxiety scores (X=8.44, SD=5.48) than those who had not (X=10.50, SD=5.64), t-test showed no significant difference between them (t=1.61, df=73, p<.113). A t-test was performed to assess whether having taken a computerized test before affected CAT scores. The analysis indicated no significant differences in CAT scores between those who had taken a computerized test before (X=32.44, SD=5.144), and those who had not (X=32.28, SD=7.275) where t=.11, df=73, p<.913.

4.3.6 Noticing the Adaptive Behaviour of CAT

One of the issues raised in the beginning of this chapter was whether subjects identify the tailored behaviour of the CAT. Table 4. 7 shows the cross-tabulation of all possible answers for question 17 (Q17, Appendix A) by those of question 18 (Q18). The Table indicates that 13 (16.3%) subjects identified the tailored behaviour of the CAT, while 8 (10%) did not notice any change in the next question's difficulty level, compared with that of the previous question. Also, 32 (40%) did not know the exact nature of CAT in either case.

Table 4.7. Identifying CAT: cross-tabulation.

(Q17) After right answer next item is ...

(Q18) After wrong answer next item is ...	More dif.	Easier	Same	don't know	Raw Total
More difficult	0	13	0	1	14 (17.5%)
Easier	4	2	0	2	8 (10.0%)
Same dif.level	1	0	8	7	16 (20.0%)
I do not know	3	6	1	32	42 (52.5%)
Column	8	21	9	42	80
Total	10.0	26.3	11.3	52.5	00.0

Moreover, 18.8% of the sample believed that the CAT questions had been arranged from easy to hard (Q16), and 13.8% from hard to easy, while 67.5% believed they had been arranged randomly (see Table 4.8). No significant gender differences were found in their answer to this question (X^2=2.73, df=2, p=.255), nor between low and high anxiety groups (X^2=2.10, df=2, p=.223).

Table 4.8 . Frequencies of possible test arrangement.

Value Label	Frequency	Percent	Cum Percent
Easy to hard	15	18.8	18.8
Hard to easy	11	13.8	32.5
Random arrangement	54	67.5	100.0
Total	80	100.0	

Those who identified CAT behaviour obtained lower mean scores (X=27.92, SD=7.5) than those who did not (X=33.24, SD=5.39). T-tests were calculated to investigate whether identifying adaptive test behaviour affects CAT scores. Interestingly, t-test values were found to be significant (t=3.04, df=78, p=.003). Investigating whether this difference is associated with sex, group, or anxiety group, Chi-Square revealed no significant differences (X^2=.010, df=1, p=.921; X^2=.258, df=1, p=.611; X^2=.088, df=1, p=.767 respectively). At the same time, no significant differences between these groups emerged on the P&P test (t=1.79, df=78, p=.077) and CAS scores (t=1.78, df=78, p<.098).

Additionally, those who identified CAT did not just get lower CAT score, but also spent a longer time completing the test than who did not (t=2.91, df=78, p<.005). Table 4.9 shows the CAT mean scores, standard deviation, and number of subjects for each group.

Table 4.9 . CAT mean, standard deviations, and number of subjects for identifying CAT groups.

	N	X	SE
Identify CAT group	13	17.65	.98
Didn't identify CAT group	67	24.52	1.74

4.4 Discussion

One of the main aims of this study was to investigate whether the scores obtained using NA-CAT of DAT can be considered equivalent to its P&P format. The correlation between the two formats of the test, the similarity of the standard deviations and the score distributions, and the significance of the difference between the two means, suggest that the two formats are not equivalent, and need to be rescaled to be so. The result confirms the findings of the first experiment in this regard. A probable explanation for the higher scores obtained on the CAT form would be that there was no time limit for completing the CAT, and that fewer items were required to complete it (20 items) as compared with the P&P format (40 items). This may reduce test fatigue and time pressure and give subjects more time to think about their answers. Another possible explanation may be related to motivational factors, in that computers provide a novel way of human assessment and cause poeple to work harder (French & Beaumont, 1991). This effect may tend to fade as new generations become gradually more familiar with using computers. It has also been presumed that CAT may have positive motivating effects on testees by presenting them with items that are sufficiently difficult to present a challenge (Betz, 1977) which may enhance their test performance. In the light of this result, organisations should not consider using both formats of NA test as an equivalent, nor exchanging their norms, cut scores or predictive validity, unless a proper rescaling for the scores is conducted. They should stick with the form that suits them most.

The correlation coefficient between the two formats was found to be .86 (p<.001) for P&PFG, and .67 (p<.001) for CATFG. The first coefficient matches Henly et al.'s (1989) findings, although the second is lower (r=0.79). In the first and second administrations, those who took P&P first (P&PFG) got lower mean scores than those who took the CAT first (CATFG). The difference was significant for the first administration but not for the second administration. In both groups, there was an increase in scores on the second administration. However, the increase was greater when the CAT followed the P&P test, suggesting that CAT benefits more from the subjects having practised on the P&P first.

The analysis of variance confirmed the finding obtained from the previous experiment concerning the NA test, and showed that the effects of different groups (P&PFG vs. CATFG) were not significant. It also revealed significant differences between the P&P and the CAT formats, demonstrating that CAT scores were higher than those of P&P. The interaction between group and format was also significant. These effects suggest that the P&P format could be more difficult than the CAT format. Confirming Sanchez (1990) finding, the gender effect was significant, males scoring higher than females. If the same entry standard is required of all candidates for a job regardless of their gender, validity evidence is needed to justify the use of the test and reduce the possibility of unfair discrimination between genders, that is, to show that those who perform poorly on the test also perform poorly on the job.

Although there was no time limit to complete the 20 items of CAT, and there was to complete the 40 items of the P&P (30 min.), the time reduction in CAT format was 20%. This leads one to believe that the reduction in time might be higher if there were no time constraints for the P&P format. This difference was found to be significant and not affected by group. The result confirms the findings of previous researchers (e.g. Oslen et al., 1989; Oslen, 1990; Moreno et al., 1984; Kiely et al., 1983). This could make selection processes faster and allow the user to use the saved time for other assessment activities. However, this finding does not suggest that the time taken to answer an individual item in CAT format is less than that of P&P test format. In fact, the response time was greater for CAT format (X=57 sec.) than for P&P test format (X=41 sec.). This result is consistent with earlier studies in showing longer testee response times for adaptive versus P&P formats (e.g. Johnson et al, 1981). The differences in testee response times were largely a function of the differing item difficulties of the two formats. Because CAT tailored the item difficulty to each individual's ability level, most examinees received more difficult items in the CAT testing condition than in P&P testing condition.

Contrary to other findings (e.g. Lee, 1986; Burke et al, 1986; Johnson & Johnson, 1981) no significant correlation was found between the amount of previous computer experience and CAT scores. One explanation for such findings could be the ease with which subjects recorded their response by using only the Y and N buttons (green label in each letter) for the entire test . This avoids giving the more experienced user an advantage over the novice. Other reasons may be related to the lengthy set of instructions and sample problems implemented before starting the actual test.

No significant difference in CAT scores were found between those who had taken computerized tests before and those who had not. This came as no surprise, as 98% of the subjects had used a computer before for various reasons, which provided them with a reasonable amount of practice dealing with this machine. However, one would like to see what the situation might be with less computer experienced subjects.

As expected, there was a significant negative correlation between previous computer experience and scores on the Computer Anxiety Scale (CAVS). This confirms the findings of Hedl et al. (1973) and supports Johnson and White's (1980) hypothesis that inexperience with computers results in high levels of anxiety. One of the limitations of such results is the use of only one self-report question (Q11. How often have you used a computer before taking this test?) rather than a series of questions to assess the amount of computer experience. This may not be enough; therefore caution is necessary in interpreting the results. A well structured measure should be used in future for such purpose.

One possible reason for such diverse findings could be attributed to the different anxiety scales used in each study. Unlike others, this study used a scale designed mainly for measuring computer anxiety, not test anxiety (eg. TAS) or state anxiety (eg. STAI). Moreover, the study agrees with Llabre et al. (1987) and Dimock & Cormier (1991) who found no systematic relationship between anxiety and performance. Also, there was no significant difference in CAT scores between high and low anxiety groups . As found by Meier (1988), significant sex differences were found in this study, where females had higher anxiety scores than males.

The result confirmed the negative relationship between computer experience and anxiety but did not reveal any relationship between computer experience and CAT scores, nor a relationship between computer anxiety and CAT scores.

Overall, examinees' attitudes toward computerized testing were positive. The reasons for this positive attitude related to potential for quick feedback, clarity and simplicity of method, lack of time pressure, shorter completion times, and less fatigue and effort. Most subjects found the idea of administering tests on a computer to be fun, interesting, and novel. Additionally, most subjects believed that they were given enough time to answer. This favourable reaction disagrees with the hypothesis expressed by Denner (1977) and Space (1981) that computerized assessments depersonalize the testing situation. However, most of the subjects complained about the inability to go back and review their answer. Other physical problems during the test session were experienced, such as eye fatigue and headache. Some of the subjects even found it difficult to read from the computer screen and adjust to the method. The small LCD monochrome screen may have been responsible for such problems. Future research should ensure reasonable size and a clearer screen to avoid such physical problems, which may effect test validity. However, generally, the CAT was well accepted by subjects. This acceptance may be enhanced with improvements in software and hardware design.

One of the important issues in human measurement is to minimize the effects of the unwanted variables that may influence examinees' performance. The final goal is to ensure that test scores reflect what they are intended to measure. For P&P tests and computer-based tests (CBT), this concern has been well documented in the literature. Although one may expect that some sources of variance discovered with P&P and CBT may be generalized to CATs, other CAT related issues deserve more attention. One of these issues concerns whether the examinee's awareness of CAT behaviour influences performance. This study found that 16.3% of the subjects identified the tailored behaviour of CAT, and 67.5% believed the items had been arranged randomly. Interestingly, those who identified CAT got significantly lower mean scores and spent longer completing the test than those who did not. This difference was not related to sex, group, or anxiety level. This important finding deserves more attention and investigation. If identifying the adaptive test affects the subject's performance, then, it seems unwise to tell subjects that they are going to take an adaptive test. The delay in completing the test could be contributed to a cognitive process taking place during the test about the difficulty level of the present item compared to the previous one, and whether that gives any indication about the correctness of the previous item. The next experiment will investigate this issue further.

CHAPTER 5

CHAPTER 5

Experiment 3

The Effects of Knowledge About Adaptive Tests on Subjects' Performance

5.1.1 Introduction

The mainl reason for using occupational tests in job selection is to generalize the manifest ability revealed from administering the test to the real job situation and to provide evidence on how well the test taker will perform on the job for which they applied. Therefore, it is important to limit any possible situational factors which may contribute to a false picture of the candidate's actual ability.

With paper-and-pencil tests, considerable effort has been made to identify factors which could influence test scores. These variables could be classified as 'examiner-related factors' (eg. smiling, nodding, presence vs. absence in testing room), or 'examinee-related factors' (eg. test anxiety level, motivation, and fatigue), or 'test-related factors' (eg. answer sheet arrangement, test length, format used for presentation), or 'situational factors' (eg. illumination, ventilation, temperature, noise level, and testing time and place).

Anastasi (1988) provides evidence on how minor aspects of the testing situation such as using desks or chairs with desk arms and the familiarity of the examiner to the test takers can influence performance. Mandler and Sarason (1952) found that instructions informing examinees that everyone was expected to finish within the time allowed had a significantly different effect on subjects with different test anxiety levels.

With the advent of computerized tests, considerable research has been devoted to identifying possible new variables. Researchers have found that previous computer experience, colour value, display layout, response devices, recording responses, and other factors (as discussed in the first chapter) affect test scores. Of course, some of the variables which relate to paper-and-pencil formats can also be generalized to computerized ones.

It is plausible that new variables will continue to arise as new components are added to the testing process. New variables which may affect the testee's performance are expected with the advent of adaptive tests. For example, as most test takers know before hand about some aspects of the test they will take (eg. test format, aspects of ability or personality measured, importance of the test), questions arise over the importance of informing the test taker about certain aspects of the adaptive test that they are going to take (Alkhadher, et al., 1994).

Subjects may, as a result of this knowledge find themselves busy detecting test behaviour, which evokes additional unwanted cognitive activities during the testing process. The perceived difficulty of the present item, for example, may be taken as a clue to the correctness of previous response (Green et al., 1984). As a result, test anxiety may increase as testees perceive easier items compared with previous ones. This may influence test score, and increase the time taken to complete the entire test, as well as each individual item.

To increase our understanding of this issue, it is important to investigate whether these possible effects are actually caused by a direct effect of knowledge about the adaptive behaviour of the CAT itself, or as a result of increased test anxiety levels generated by such knowledge. The concern, in this experiment, is with subjects' state anxiety, which may be affected by knowledge of the adaptive test's functions. State anxiety levels change as a function of the person's perception of a situation at any given point in time (Spielbereger, 1983).

Some previous attempts have been made to assess the effect of a subject's 'test wiseness' on their performance. Subjects may identify the correct options by a process of eliminating the obviously incorrect choice, by noticing an option is worded incorrectly or is too broad or too narrow to be correct, or through knowledge of the teacher's idiosyncrasies (Aiken, 1988). Moreover, the strangeness of options may also act as a cue to the correct answer (Strang, 1980).

On the other hand, it has been found that subjects who have had considerable previous experience in taking standardized tests gain advantages in test performance over those who are taking the test for the first time (Millman et al., 1965). This advantage could be attributed to intellectual skills, work habit, and problem-solving behaviour (Anastasi, 1988) acquired during previous test administrations. So it may be that the more people take ability tests, the more they are able to identify the adaptive test and the more likely their performance is to be affected, if of course, there is any such effect.

5.1.2 Method

5.1.2.1 Examinees

A sample of 120 university students between 18 and 44 years of age, with a mean age of 20.71 years (SD=3.01) were randomly assigned to one of the three experimental conditions. Subjects were recruited through advertisements directed toward the campus community at the University of Nottingham. There were 60 males and 60 female students. There were 106 undergraduate and 14 postgraduate students. All subjects spoke fluent English.

5.1.2.2 Materials

Differential Aptitude Tests (DAT)

For this study, the Numerical Ability (NA) test and the Abstract Reasoning (AR) test of the Differential Aptitude Tests (DAT) (the adaptive version) (McBride, 1986) (Form V) were used. (See experiment 1 for details).

State-Trait Anxiety Inventory (STAI Form Y-1)

The STAI Form Y-1 (Spielberger et al., 1983) was used in this experiment to measure the level of anxiety induced by the potentially stressful experimental procedure (Spielberger et al., 1983). It was defined as a transitory emotional condition characterized by subjective feelings of tension and apprehension. It has been verified in a number of studies that A-State scores increase in response to situational stress and decline under relaxed conditions (Towell, 1975). The scale consists of 20 short descriptive items and self-reported rating scales for measuring how the subject feels at the moment. The four essential qualities evaluated by the scale are feelings of apprehension, tension, nervousness and worry. The STAI has strong psychometric evidence. Its reliability ranges from .86 to .92 and extensive validity evidence is reported in its manual.

The Attitude Questionnaire (AQ)

This questionnaire contained 9 items designed for this experiment to assess subject's awareness of the tailored nature of the adaptive test and the difficulty level of the items, and the possible effects of identifying the adaptive test on his or her performance (Appendix C).

Last, biographical data such as age, sex, nationality, first language, main department, and educational status were obtained from all subjects.

5.1.2.3 **Procedure**

The subjects were randomly assigned to one of three experimental groups. The first group (Knowledge group, KG) consisted of 40 subjects The tester read to them the following statement: "*You are going to take a computer-based test called an adaptive test. Unlike the conventional paper-and-pencil test method, an adaptive test selects each question according to your performance on the previous one. If you answer an item right, the next item will be more difficult; if you answer it wrong, the next item will be easier*".

The second group (Knowledge and control for anxiety group, K&CAG) which consisted of 40 subjects, had also been informed about the basic function of the CAT as with the first group. However, in an attempt to reduce the anxiety, if any, produced by the description of the test, the examiner read the following statement: "*This is likely to work to your advantage, as it helps most people to get a better score*". The third group (Control group, CG) which consisted of 40 subjects was the control group and no information about any aspect of the CAT was given to them.

After reading the statements for KG and K&CAG, and before administering the tests for CG, the subjects were asked to complete the STAI to measure their state anxiety at that moment. Following that, the NA then the AR tests were administered for all subjects. Both tests were administered according to the instructions in their manuals. All subjects received general test instructions followed by directions for recording their answers. Answer revisions were not permitted. Scrap paper, two pencils and rubber were provided.

Completion time for the two tests together was recorded. The AQ was presented immediately after completing the two tests. The questionnaire was presented in a written format. The experiment was conducted in an individual cubicle and the subjects were advised to call the examiner if help was required. After finishing both sessions, subjects were thanked for taking part in the experiment, offered four pounds for participating in the experiment and debriefed on the aims of the study.

As it was available at that time, the SX3U1 Notebook IBM compatible PC with LCD monochrome monitor and a standard QWERTY keyboard was used.

5.1.3 Results

The dependent variables for this experiment were subject's scores on the CAT tests, STAI, AQ, and the completion testing time. The independent variables were the condition (knowledge, reassurance or control) and gender (male or female). The assumptions of normality and homogeneity for each of the dependent variables were checked and found to be satisfied. All data were transferred to a computer data file and analysed using the Statistical Package for the Social Science (SPSSPC+).

To assess the relationships between the various variables used in this study, all data were analysed using paired-samples t-tests, multivariate analysis of covariance (MANCOVA), Pearson's product-moment correlations,. frequencies, cross-tabulations, and Chi-squares.

5.1.3.1 The Affects of Knowing CAT Behaviour

Multivariate analysis of covariance (MANCOVA) was conducted in a 3 (conditions) x 2 (sexes) design using the dependent variables of NA and AR scores, and the covariate of the number of times subject had taken ability tests before (TTB). A second MANCOVA was conducted with the same independent variables and covariate, but with testing (TIME) and state anxiety (STAI) scores as the dependent variables. The covariate was essential for these analyses in order to remove the confounding influence of previous experience of taking ability tests (TTB), since a significant difference between the groups in the three experimental conditions was found on TTB (F(2,117=3.97, p<.002). Table 5.1 shows the correlations between the dependent variables and the covariate.

Table 5.1. Correlation coefficients among dependent variables and the covariate (TTB) for all groups

	NA	AR	STAI	TIME	TTB
NA		.47**	-.18	-.22*	.20
AR			.09	-.34**	.15
STAI				-.05	-.03
TIME					-.05
TTB					

*p<.01, **p<.001.
NA=Numerical Ability; AR=Abstract Reasoning; STAI=State-Trait Anxiety Inventory; TIME=Time taken to complete the test; TTB=Number of time taken test before.

Table 5.2 shows the means, standard deviations, and number of subjects for the dependent variables (NA, AR, STAI, and TIME) and the covariate (TTB). It can be seen from Figures 5.1, 5.2, 5.3, and 5.4, that for women, there is a difference between the three conditions on the mean scores on NA, AR, and STAI, but not on the TIME, whereas for men, there is a difference between the three conditions on the mean scores on NA, and TIME but not AR or STAI. However, not all these differences are significant. The MANCOVA shown in Table 5.4 examined these differences.

Table 5.2 Means and standard deviations, by treatment condition, for the NA, AR, STAI and TIME, and the covariate (TTB).

	KG		K&CAG		CG	
	F	M	F	M	F	M
	(n=20)	(n=20)	(n=20)	(n=20)	(n=20)	(n=20)
NA	28.85 (7.77)	30.95 (5.51)	31.50 (5.38)	30.45 (5.21)	33.50 (5.25)	33.65 (4.92)
AR	35.85 (7.03)	39.65 (5.70)	37.75 (6.15)	40.05 (5.63)	40.15 (4.37)	39.80 (4.42)
STAI	40.10 (13.66)	32.15 (8.18)	34.55 (8.68)	33.65 (8.40)	36.05 (7.49)	32.40 (7.33)
TIME	35.90 (10.76)	39.45 (7.67)	34.62 (6.16)	37.68 (6.10)	35.32 (5.31)	33.26 (5.37)
TTB	2.70 (1.89)	2.60 (1.39)	3.00 (1.83)	1.65 (1.27)	1.90 (1.21)	1.55 (0.83)

KG=Knowledge group; K&CAG= Knowledge and control for anxiety group; CG=Control group; NA=Numerical Ability; AR=Abstract Reasoning; STAI=State-Trait Anxiety Inventory; TIME=Time taken to complete the test; TTB=Number of time taken test before.

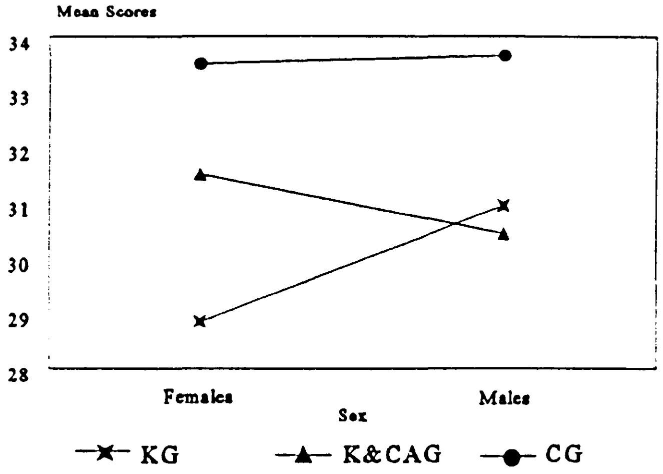

Figure 5.1 Mean scores for NA test by sex for all groups.

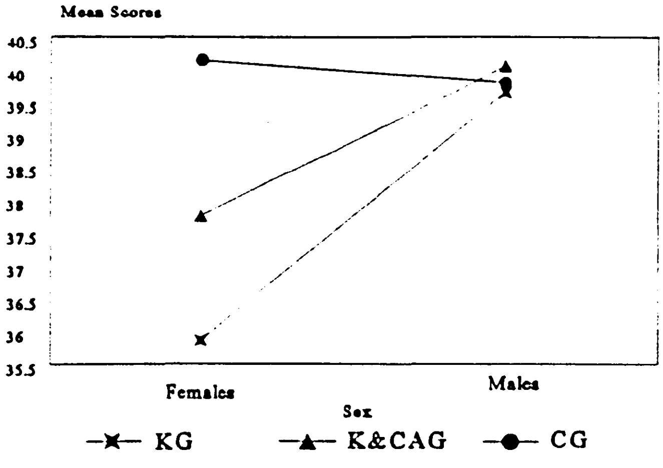

Figure 5.2 Mean scores for AR test by sex for all groups.

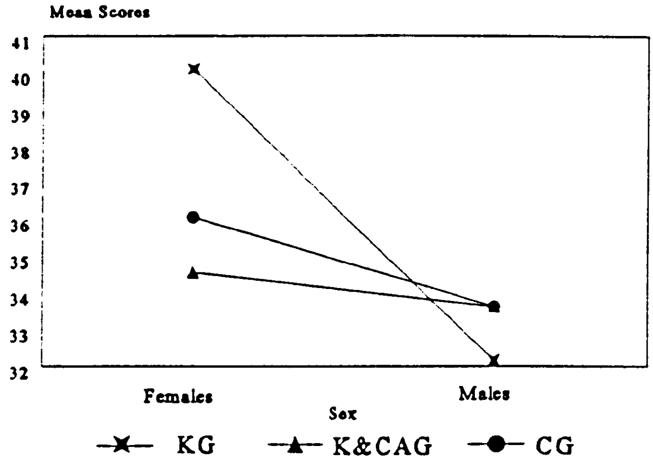

Figure 5.3 Mean scores for STAI test by sex for all groups.

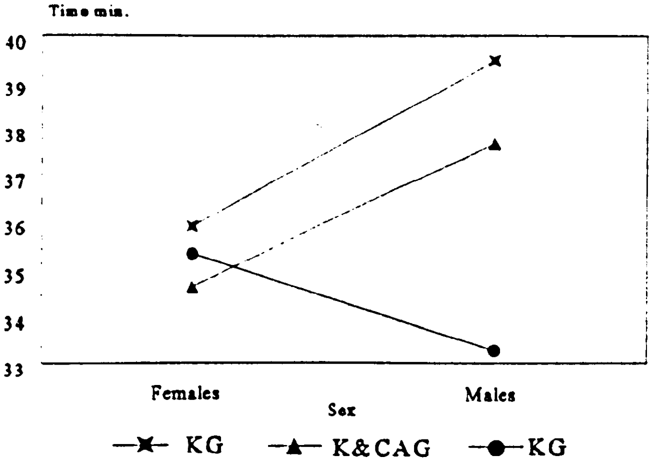

Figure 5.4 Mean scores for TIME test by sex for all groups.

The MANCOVA results are reported in terms of the Pillai-Bartlett F approximation, because that criterion has been shown to be relatively robust (Olson, 1976). The MANCOVA for the NA and AR produced significant main effects for condition (F(4,226)=4.23, p<.002), and gender (F(2,112)=3.88, p<.023). However, the group by gender interaction proved to be non-significant (F(4,226)=1.25, p<.288). The adjusted group means for each of the dependent variables are shown in Table 5.3.

Table 5.3. Adjusted means, by treatment conditions, for the NA, AR, STAI, and TIME.

	KG		K&CAG		CG	
	F	M	F	M	F	M
NA	28.34	30.55	30.67	31.08	33.86	34.39
AR	35.45	39.34	37.09	40.55	40.44	40.39
STAI	40.32	32.32	34.90	33.38	35.90	32.08
TIME	36.05	39.57	34.87	37.49	35.21	33.04

KG=Knowledge group; K&CAG= Knowledge and control for anxiety group; CG=Control group; NA=Numerical Ability; AR=Abstract Reasoning; STAI=State-Trait Anxiety Inventory; TIME=Time taken to complete the test; TTB=Number of time taken test before.

To uncover the source of the multivariate effects, a univariate analysis of covariance (ANCOVA) was used for each dependent variable. The univariates (see Table 5.4) revealed nonsignificant group x gender interactions for the NA (F(2,113)=.38, p<.686), and the AR

(F(2,113)=2.15, p<.121). However, the condition main effect was significant for the NA (F(2,113)=8.18, p<.001), and AR (F(2,113)=3.93, p<.022). The gender main effect was only significant for the AR (F(1,113)=7.85, p<.006), not for the NA (F(1,113)=1,21, p<.274). Post hoc (Scheffe˜) comparisons showed that CG had significantly higher NA mean than both KG and K&CAG while CG had significantly higher AR mean than KG. Also, males got a higher AR mean than females.

The MANCOVA for the TIME and STAI produced significant main effects for gender (F(2,112)=4.20, p<.017). However, the group and group by gender interactions are not significant (F(4,226)=2.20 p<.070; F(4,226)=1.96, p<.102 respectively). The univariates (see Table 5.4) also produced nonsignificant group x gender interactions for the TIME (F(2,113)=2.63, p<.077) and the STAI (F(2,113)=1.35, p<.263). However, the condition main effect was significant for TIME (F(2,113)=3.59, p<.031), but not STAI (F(2,113)=.84, p<.437). Concerning gender main effect, it was only significant for STAI (F(1,113)=7.30, p<.008) but not for TIME (F(1,113)=1.41, p<.237). Post hoc (Scheffe~) comparisons showed that KG and K&CAG spent significantly longer times completing both ability tests than did CG. Also, females got a significantly higher mean STAI scores than males.

Table 5.4 . Results of the univariate analysis of covariance conducted for the NA, AR, TIME, and STAI

Source of Variation	SS	DF	MS	F	Sig of F
NA					
G	428.60	2	214.30	8.18	.000 **
Sex	31.71	1	31.71	1.21	.274
G BY Sex	19.82	2	9.91	.38	.686
AR					
G	170.19	2	85.09	3.93	.022 *
Sex	169.72	1	169.72	7.85	.006 **
G x Sex	93.08	2	46.54	2.15	.121
TIME					
G	254.44	2	127.22	3.59	.031 *
Sex	50.15	1	50.15	1.41	.237
G x Sex	186.38	2	93.19	2.63	.077
STAI					
G	129.59	2	64.79	.84	.437
Sex	566.91	1	566.91	7.31	.008 **
G x Sex	209.66	2	104.83	1.35	.263

1-tailed Signif: * - .05 ** - .01; NA=Numerical Ability; AR=Abstract Reasoning; STAI=State-Trait Anxiety Inventory; TIME=Time taken to complete the test

When asking whether knowledge about CAT tailored behaviour affected them, 57.5% of those who had been informed (KG & K&CAG) thought it had affected them, whereas 41.35% did not. For those who thought it had affected them, comparing their mean scores on question Q7 (items A, B, and C) with a theoretical population mean value of 4 (the mid point on the 7 points scale), t-tests showed significant differences between the subjects' mean on Q9 (items A & C), and the theoretical value (t=3.58, df=37, p<.05, and t=6.52, df=37, p<.001 respectively) indicating that the subjects thought that the knowledge of CAT made them slower and worried them. However, no significant differences were found between the subject's mean on Q9 (item B) and the theoretical value (t=1.18, df=37, p<.10) which indicates that there is no evidence that the knowledge of CAT made them think they were doing worse or better.

In response to the open question Q9 (item D), 6 subjects indicated that the knowledge increased their tendency to give up, 4 subjects felt very pressed to do well, and 14 subjects tried to assess how well they were doing during the test.

5.1.3.2 Noticing CAT Behaviour.

It seem useful here to distinguish between being informed about CAT behaviour and noticing such behaviour. Those who have been informed about CAT behaviour before the test session started may not necessarily notice it. Table 5.5 shows those who noticed CAT behaviour in each group (those who notice that every time they think they answer a question right the next question becomes more difficult, and when they answer it wrong the next item becomes easier) and those who did not. Remember that CG had not been told about CAT behaviour.

Table 5.5. Noticing CAT behaviour: cross-tabulation by group.

		Noticing CAT		
		Yes	No	Row Total
Group	KG	19	21	40
	K&CAG	20	20	40
	CG	11	29	40
Column Total		50 40.8 %	70 59.2%	120 100%

KG=Knowledge group; K&CAG= Knowledge and control for anxiety group; CG=Control group

With regard to those who had already been informed about CAT tailored behaviour (KG and K&CAG), no significant differences were found between those who noticed CAT tailored behaviour and those who did not, on any of the dependent variables (Table 5.6).

Table 5.6. Means, standard deviations, and t-test for those who identified CAT tailored behaviour and those who did not in G1 and G2 on the dependent variables.

	Noticed CAT		Did not Notice CAT			
	X	SD	X	SD	t	P
NA	29.92	6.65	30.90	4.55	.78	.439
AR	38.53	4.68	38.14	6.68	.30	.769
STAI	35.32	10.92	34.93	7.17	.19	.850
TIME	37.47	4.83	36.40	8.84	.66	.510

NA=Numerical Ability; AR=Abstract Reasoning; STAI=State-Trait Anxiety Inventory; TIME=Time taken to complete the test;

Verifying the results obtained from the first experiment, 11 subjects out of the 40 (27.5%) in the CG who had been given no information about CAT, did identify the CAT 'tailored' function. Those who identified CAT behaviour got significantly lower NA scores than those who did not (t=3.32, df=38, p=.002), and spent a significantly longer time completing the tests (t=3.10, df=38, p=.004). No significant differences were found for the other variables (Table 5.7).

Table 5.7. Means, standard deviations, and t-test for those in G3 who identified CAT tailored behaviour and those who did not on the dependent variables.

	Notice CAT		Did not Notice CAT			
	X	SD	X	SD	t	P
NA	30.10	6.02	34.90	3.11	3.32	.002 **
AR	39.00	2.72	40.34	2.16	1.64	.110
STAI	34.36	8.22	34.17	9.41	.06	.953
TIME	36.11	1.37	33.60	2.54	3.10	.004 **

1-tailed Signif: * - .05 ** - .01; NA=Numerical Ability; AR=Abstract Reasoning; STAI=State-Trait Anxiety Inventory; TIME=Time taken to complete the test; TTB=Number of time taken test before.

5.1.3.3 Prior Test Experience and Ability to Identify CAT

As seen in Table 5.1, taking ability tests before (TTB) does not correlate with any of the dependent variables or with perceiving the difficulty level of the questions (question Q1) (r=.06). Moreover, those who identified CAT tailored behaviour did not differ significantly from those who did not in the number of times they had taken ability tests before (Table 5.8).

Table 5.8 t-test for those who identified CAT tailored behaviour and those who did not on the amount of previous experience in each group.

	Identified CAT			Did not Identify CAT				
	N	X	SD	N	X	SD	t	P
KG	18	2.22	1.35	22	3.00	1.80	1.52	.138
K&CAG	20	2.50	1.57	20	2.15	1.84	.65	.522
CG	11	1.73	1.10	29	1.72	1.03	.01	.993
Total	49	2.22	1.40	71	2.24	1.62	.05	.958

KG=Knowledge group; K&CAG= Knowledge and control for anxiety group; CG=Control group

5.1.4 Discussion and Conclusion

Measuring the testees' actual ability and limiting the situational factors that could contaminate their score has been a major concern for both test developers and users. One purpose of this experiment was to investigate the possible effects of knowing about adaptive test behaviour on subjects' test performance, and whether such possible effects are a direct result of an increase in test anxiety levеL

This study demonstrated that a knowledge of CAT behaviour negatively affected subjects' performance on both NA and AR, and caused subjects to spend a longer time completing the tests. Knowledge did not increase the state anxiety level (STAI) for those who had been informed, since there was no significant difference among the three conditions in the amount of state anxiety indicated. Thus, these findings suggest that it is not anxiety, but rather knowledge of CAT tailored behaviour, that impacted on subjects' performance. In all the analyses, no interactions were found between conditions and gender. Conditions differed significantly on the NA, AR, and the time taken to complete the tests (TIME), where the control group (CG) got higher NA and AR scores and spent a shorter time completing the tests. The genders only differed significantly on the AR and the STAI, where males got higher AR mean scores, and females higher STAI mean scores. More than half of the subjects informed about CAT tailored behaviour believed it made them perform more slowly and worried them.

No other studies have been found on this issue. It is important for future researchers to replicate this study and to extend it with other populations such as high and low ability subjects. It would be interesting also to investigate other ability tests. This study used only Numerical and Abstract Reasoning tests.

It is unclear, however, how such effects has such knowledge. Green et al. (1984) believe that "to the extent that the applicant knows or concludes that item difficulty depends on previous responses, the perceived difficulty of the present item may be taken as a clue to the correctness of earlier responses. It is a form of feedback, however subtle". Moreover, unwanted mental processes may be activated during the test as testees try to assess their performance on the previous item. That could take considerably more time. If we assume that a process of feedback is working as the adaptive test proceeds, one wonders whether feedback in itself has such ability to alter a testee's performance. To clarify the picture, it seems important to assess the possible effects of such feedback on a testee's performance. This issue will be tackled in the next experiment.

Comparing the control group score on NA with that of the experiment 2 CATFG (those who took NA-CAT version first), t-tests reveal no significant differences between the two groups (t=.649, df=79, p>.10). However, although there was a significant difference between the sexes on the NA in the first experiment, no such differences were found in this experiment. One possible explanation is the difference in populations used in both experiments. The first experiment used mainly psychology students (66%), whereas this experiment used students from various departments at Nottingham University.

It was postulated at the beginning of this chapter, that as a result of the possible effect of 'test wiseness', the more people take ability tests, the more they become able to identify adaptive tests. The results of this study do not support this hypothesis. However, the results confirm the second experiment's finding concerning the difference in performance on the NA and time taken to complete the test, between those who noticed CAT tailored behaviour and those who did not, when there was no prior knowledge about the basic properties of CAT. This give us another explanation for the source of unequivalence found in the first and second experiment between the P&P and CAT formats of the NA test.

The implications of this study are of considerable importance in the use of the adaptive tests. It seems unwise (on the basis of these findings alone) to inform testees before the test session about the actual function of the adaptive test, as this may affect their performance and delay their responses. This implication seems more important when the decision based on the testee's results takes into account the amount of time he or she spends to complete the test, and when the decision is of vital interest to him or her. This notion is not limited to the occupational setting, but also applies to any situation where adaptive tests are used. In a recession, when more employees lose their jobs and more competitive candidates struggle for a limited places, the need for more refined and fair measurement is even greater.

Experiment 4

The Effects of Immediate Knowledge of Results on the Testee's Score, Anxiety, and Answering Time

5.2.1 Introduction

The previous experiment posed the question of whether knowledge of the basic function of adaptive tests influences testees' performance on ability tests. The study demonstrated that the knowledge of CAT behaviour negatively affects subjects' performance on both NA and AR, and causes subjects to spend a longer time completing the tests. The findings suggest that it is not anxiety, but rather the knowledge of the CAT tailored behaviour, that has an impact on subjects' performance. The study suggested that a form of feedback may act during adaptive testing which may have a negative effect on testees' performance. However, does feedback about the correctness of response during the testing process itself have the capability to negatively effect testee's performance? This is what the next experiment attempted to find out. Item by item feedback was provided to the testees to make success and failure more obvious. The feedback was designed to reflect the actual correctness of testees' answers. The assumption was that success in answering an item correctly would be viewed by a testee as a form of positive reinforcement, while failure to do so would be viewed as negative reinforcement. The kind of reinforcement received may differentially motivate or de-motivate the testee.

The effect of Immediate Knowledge of Results (IKR) on examinee's test performance is unclear, and the studies in this area yield conflicting results. Some researchers have found that IKR improves test performance. For example, Rocklin & Thompson (1985) found that IKR improved performance on verbal ability tests for those undergraduate students given an easy test, but not for those given a difficult one. But they also found a reduction in test anxiety. Betz (1977) found that for a high ability group, mean test scores under IKR conditions were significantly higher than those under non-IKR conditions on both the conventional and adaptive tests. She found the same results with a low ability group, but in this case the difference was statistically significant only for the conventional testing strategy. Some researchers attributed the increase in test performance to the increase in testee motivation and reduction in test anxiety associated with IKR (e.g. Weiss & Betz, 1973). It may also be that IKR facilitates ongoing test accuracy and serves a cueing function by offering information that will help with the next answer.

On the other hand, Strang & Rust (1973) found that providing IKR reduced test performance and increased both student nervousness and testing time. Wise et al. (1986) found that IKR led to both increased anxiety levels and decreased test performance. Harold et al. (1973) also found that IKR resulted in a loss of accuracy, and also produced an increase in task completion time when coupled with the definition of the task as a test. They also found an increase in nervousness. In their study using a 32 item arithmetic test with elementary school children, Wise and Wise (1987) compared the use of computer administration with immediate item feedback, computer administration without immediate item feedback, and paper administration. They did not find any significant differences among the treatment groups in terms of mean test score. However, they found that item feedback increase significantly the state anxiety of high Math achievers. Other research has found no differences in performance between those who had IKR and those who had not (e.g. Wise, et al., 1989).

Thus, the purposes of this experiment were to examine the effects of IKR on: (a) the testee's score and anxiety; and, (b) the average time (in seconds) taken to answer a question in cases of right and wrong answers. The study also attempted (c) to investigate whether negative or positive feedback on the answer to the previous item causes subjects to spend a longer or a shorter time completing the next item.

5.2.2 Methods

5.2.2.1 Subjects

The subjects were 80 students recruited from a sign-up sheet at the University of Nottingham. All the subjects were undergraduates. There were 40 males and 40 females in the sample, ranging in age from 18 to 26 years, with a mean age of 20.06 years (SD=2.03). For all subjects, the first language was English, and major degree subject varied widely. The subjects were randomly assigned to the two treatment conditions used in this study (40 subjects in each condition group).

5.2.2.2 Materials

Selby MillSmith Adaptive Ability Test (CAT-NA)[1]

This adaptive ability test was produced by Seilby MillSmith Limited in 1988 to be used specifically in the UK and Europe. It consists of three modules available in software form, designed to test Language Ability, Numeric Ability, and Administrative Ability. Each test lasts no longer than 20 minutes. They have been developed specifically for use in selection, training and career development for operatives, clerical and supervisory staff, and management.

The Numerical Ability test module (NA), which was used in this experiment, was designed to measure the aspects of numeracy generally required in job situations. It assess a testee's ability to estimate and solve problems with speed and relative accuracy up to

[1] As modifying the DAT to provide the necessary item by item feedback was not possible, this new adaptive test was used for this experiment. I wish to acknowledge with thanks Selby MillSmith Ltd. for their help and cooperation, in providing and where necessary modifying their software.

about 'O' Level or GCSE Level C standards.

Each questions has five answer choices. The test starts with a short locator test to determine the initial difficulty level. Then the speed and accuracy of the candidate's work is continuously monitored. The computer selects the set of items which is closest to the candidate's ability level. If he/she answers a set of items correctly in a short period, the computer will shift him/her to a higher level of difficulty. If he/she answers many items incorrectly, or spends a long time completing the set, the computer will present items of lower difficulty. The test consists of three difficulty levels with three blocks of items in each, in order to be able to re-allocate an individual to different levels during the test. The test session is terminated when the individual performance is consistent, and this may occur at any stage during a test. The mean difficulty level for each level and block is reported in the test manual. There is no overlap in item difficulties between adjacent blocks.

The testee's score on the test depends on the number of correct answers, their difficulty level, and the length of time (in seconds) taken to answer them. Once the testee has completed the test, the results are stored in a file, ready to be presented to the user. The results file contains the raw score, the correctness of the item, time taken, the answer selected, and the correct choice.

The NA test has a test re-test reliability coefficient of .67. The test manual reported the correlation coefficient between CAT-NA and the Saville Holdsworth Ltd Personnel Test Battery (PTB) to be .55 with Numerical Computation (NP2), and .03 with the Verbal Meaning (VP5). This indicates both convergent and discriminant validity. Also the correlation coefficient with GCE 'O' levels is reported to be .26 with 'O' level points scores, .31 with Maths Grade, and .09 with English language grade. This information was reported in the test manual.

State-Trait Anxiety Inventory (STAI Form Y-1)

See experiment 3.

5.2.2.3 **Procedure**

The subjects were told before the testing session that they would be given a numerical ability test. The subjects were randomly assigned to one of two experimental groups: the 'feedback group' (FBG) or the 'no feedback group' (NFBG). Both groups consisted of 40 subjects, with an equal number of males and females. The computerized CAT-NA was modified to inform those in the FBG about the accuracy of their answers after each item ('That answer was right/wrong' as appropriate). This message was displayed for 4 seconds, whereas for the NFBG a 'Please wait' message was presented on the screen for the same period. The 4 seconds did not count toward the time limit for the test. Between blocks 1 and 2 and between blocks 2 and 3 of the questions, the message

'The next set of questions will be more difficult/easier' (as appropriate) was displayed for FBG only. The STAI was used both before and after administration of the CAT-NA to determine if differential change in situation anxiety existed between treatment group. The tests were administered according to the instructions in their respective manuals. Answer revisions were not permitted. All subjects received general test instructions followed by directions for recording their answers. Scrap paper, two pencils and an eraser were provided. After all data had been collected, subjects were thanked for taking part in the experiment, debriefed on the aims of the study, given feedback about their results, and offered four pounds for participating in the one hour experiment.

The experiment was conducted in an individual cubicle and the subjects were advised to call the examiner if help was required. The SX3U1 Notebook IBM compatible PC with LCD monochrome monitor and a standard QWERTY keyboard was used.

All data were transferred to a computer data file and analysed using the Statistical Package for the Social Science (SPSSPC+).

5.2.3 Results

5.2.3.1 The Effects of IKR on Testee's Scores and Anxiety

Low and nonsignificant correlations were found between the CAT-NA and both the pre-STAI and post-STAI scores (r=-.006 and r=-.040 respectively). However, as was expected, a significant correlation was found between the pre-STAI and the post-STAI (r=.665, p<.001).

Table 5.9 (see also Figures 5.5 & 5.6) shows the means and standard deviation for the dependent variables (CAT-NA and Post-STAI) and the covariate (Pre-STAI) broken down by group condition and gender. The mean scores for the CAT-NA for both groups were less than that reported in the test manual (N=445, X=90.59, SD=16.82, which was based on a sample drawn from 11 different universities and polytechnics across the UK). For the control group, the mean pre- and-post-STAI scores were similar to those reported in the test manual (males: N=296, X=36.47, SD=10.02; females: N=481, X=38.76, SD=11.95).

Table 5.9. Means and standard deviation for CAT-NA, post-STAI, and pre-STAI broken down by group and gender.

	NFBG			FBG		
Variable	F (n=20)	M (n=20)	Total (n=40)	F (n=20)	M (n=20)	Total (n=40)
CAT-NA	77.30 (16.59)	86.05 (19.13)	81.68 (18.22)	60.50 (19.34)	67.70 (16.09)	64.10 (14.61)
Pre-STAI	35.80 (9.97)	35.45 (6.02)	35.63 (8.13)	38.35 (6.73)	36.65 (5.79)	37.50 (6.26)
Post-STAI	35.35 (9.70)	36.65 (6.84)	36.00 (8.31)	41.00 (4.90)	41.24 (5.44)	41.13 (5.11)

Note, NFBG=no feedback group; FBG=feedback group; below each mean the corresponding standard deviation is shown in parentheses.

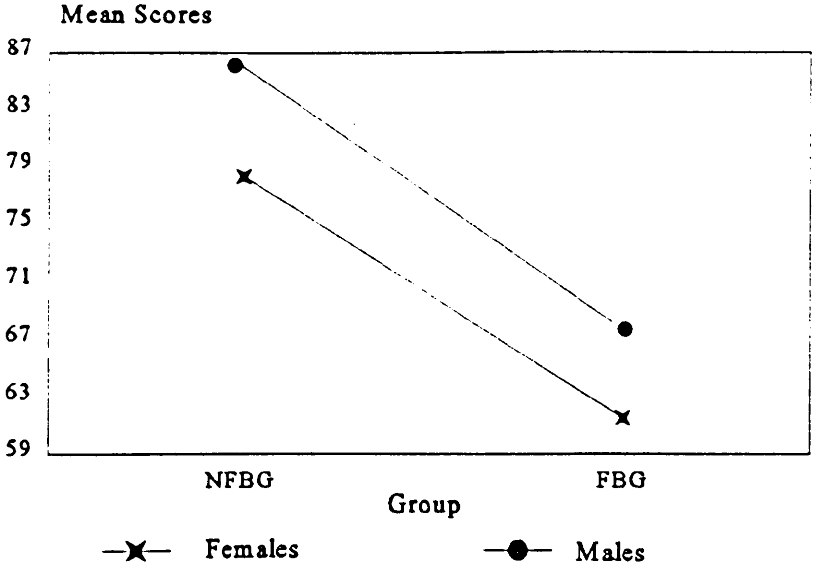

Figure 5.5. CAT-NA scores for both sexes in both condition groups.

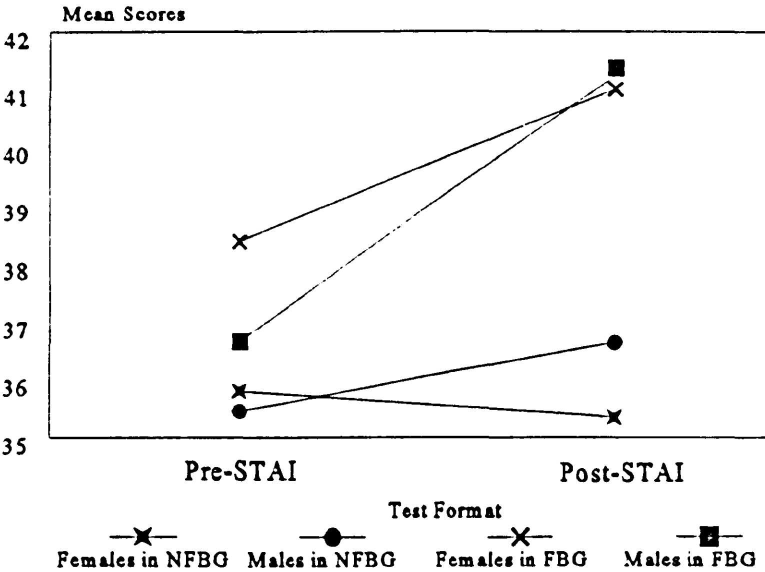

Figure 5.6. STAI mean scores before and after CAT-NA for both sexes in both groups.

In order to meet the assumptions of the analysis of variance, a square root transformation was carried out before analysis of the data. ANCOVA (Table 5.10) on the CAT-NA and post-STAI was significant for group (F(1,76)=23.86, p<.001; F(1,76)=11.25, p<.001 respectively). However, the effect of sex difference was significant only with CAT-NA (F(1,76)=5.01, p<.05). The NFBG and males got higher CAT-NA mean scores than their FBG and females counterparts, and FBG got higher post-STAI mean scores than NFBG. No significant differences were found for the interactions between group and sex on either dependent variables, or between sexes on post-STAI.

Table 5.10. ANCOVA on the CAT-NA and post-STAI.

Source of Variation	SS	DF	MS	F	Sig of F
CAT-NA					
G	6315.16	1	6315.16	23.86	.000 **
Sex	1326.40	1	1326.40	5.01	.028 *
G x Sex	8.47	1	8.47	.03	.858
Post-STAI					
G	302.37	1	302.37	11.25	.001 **
Sex	40.85	1	40.85	1.52	.221
G x Sex	.17	1	.17	.01	.937

1-tailed Signif: * - .05 ** - .01; CAT-NA=Computerised adaptive test-Numerical Ability test; G=Group; Post-STAI=State-Trait Anxiety Inventory.

5.2.3.2 The Effect of IKR on Time Taken to Answer the Questions

Investigating the effect of IKR on time taken to answer the questions, Table 5.11 shows the means and standard deviations of the times taken to complete individual items. The results are shown according to groups, gender, and correctness of the answer (to see whether, on average, wrong/right response take shorter/longer time to complete). Table 5.12 and Figure 5.7 reveal significant differences between treatment groups (F(1,2149)=7.66, p=.006), and item correctness (F(1,2149)=97.88, p<.001), but not between genders (F(1,2149)=.70, p=.403). The feedback group and incorrect answers yielded longer response times. Of the possible two-way interactions, group by gender was significant (F(1,2149)=6.71, p=.010), as was gender by item correctness (F(1,2149)=4.34, p=.037), but group by item correctness was nonsignificant (F(1,2149)=.43, p=.510). Interestingly, male subjects took the longest time to respond when in the FBG, but the shortest time when in the NFBG. Also, females took the longest time to answer incorrectly (X=58.74, SD=50.72), but took the shortest time with correct answers (X=33.82, SD=33.55) (note that males took X=52.81, SD=48.74 with wrong answer, and X=36.60, SD=33.77 with right answer). 3 way interaction between the three independent variables was also found to be significant (F(1,2149)=7.46, p<.001).

Table 5.11. Means and standard deviations of the time taken (seconds) to complete an individual item, for wrong and right answer.

	NFBG			FBG		
The item is	F (n=579)	M (n=603)	Total (n=1182)	F (n=506)	M (n=469)	Total (n=975)
Wrong	62.17 (56.20)	43.48 (47.21)	52.49 (52.45)	55.80 (45.58)	64.01 (48.46)	59.37 (46.90)
Right	33.02 (33.31)	34.00 (35.01)	33.51 (34.17)	34.81 (33.86)	39.98 (31.82)	37.36 (32.95)
Total	37.75 (39.41)	35.59 (37.45)	36.65 (38.42)	39.37 (37.68)	44.28 (36.50)	41.73 (37.18)

Note: NFBG=no feedback group; FBG=feedback group; below each mean the corresponding standard deviation is shown in parentheses; n=number of items attempted.

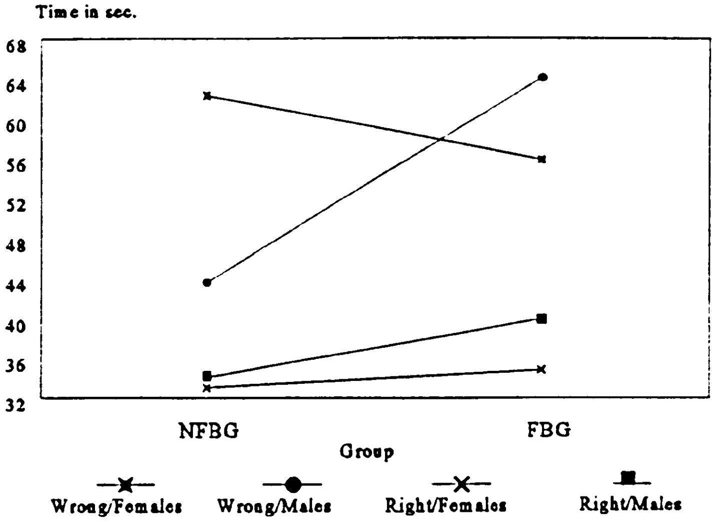

Figure 5.7. Mean by group, gender, and correctness of item, for time taken to answer the item.

Table 5.12. ANOVA for effects of group, gender, and correctness of item on time taken (in seconds) to answer the item.

Source of Variation	SS	DF	MS	F	Sig of F
G	10447.10	1	10447.10	7.66	.006 **
SEX	954.91	1	954.91	.70	.403
IC	133483.38	1	133483.38	97.88	.000 **
G x SEX	9151.24	1	9151.24	6.71	.010 **
G x IC	590.93	1	590.93	.43	.510
SEX x IC	5914.36	1	5914.36	4.34	.037 *
G x SEX x IC	10171.72	1	10171.72	7.46	.006 **

1-tailed Signif: * - .05 ** - .01; G=Group; SEX=Sex; IC=Item Correctness.

Tables 5.13, 5.14 and Figure 5.8 attempt to investigate whether negative or positive feedback on the answer of the previous item (i.e. informing him/her about the correctness of the previous item-right/wrong) causes the subject to spend a longer or a shorter time completing the next item. The prime concern here is with the dependent variable (time), the correctness of the previous item, and interactions with group and gender. As shown in Table 5.14, a significant difference was found for item correctness (F(1,2069)=4.34, p<.05), where subjects spent a shorter time on the next item when getting negative feedback on the previous item (X=36.42, SD=27.90) than with positive feedback (X=40.59, SD=40.27). However, neither the interactions between item correctness and group nor item correctness and gender are significant (F(1,2069)=1.12, p=.291; F(1,2069)=.60, p=.439 respectively).

Table 5.13. Means and standard deviations by group, gender, and correctness of previous item, for the time taken (in seconds) to answer the next item.

	NFBG			FBG		
previous item was...	F (n=579)	M (n=603)	Total (n=1182)	F (n=506)	M (n=469)	Total (n=975)
Wrong	33.02 (27.51)	30.48 (25.06)	31.68 (26.22)	40.85 (30.60)	41.10 (26.39)	40.96 (28.75)
Right	39.57 (41.64)	37.60 (39.78)	38.57 (40.70)	39.97 (40.15)	46.51 (38.77)	43.18 (39.59)
Total	38.56 (39.84)	36.42 (37.81)	37.47 (38.82)	40.17 (38.22)	45.50 (36.81)	42.73 (37.62)

Note: NFBG=no feedback group; FBG=feedback group; below each mean the corresponding standard deviation is shown in parentheses; n=number of items attempted.

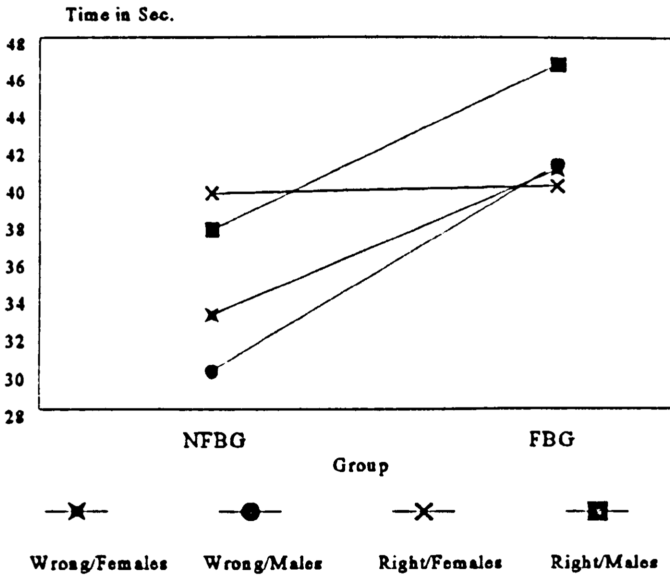

Figure 5.8. Means by group, gender, and correctness of previous item, for time taken to answer the next item.

Table 5.14. ANOVA for effects of correctness of item, group, and gender on time taken (in seconds) to answer the next item.

Source of Variation	SS	DF	MS	F	Sig of F
IC	6338.78	1	6338.78	4.34	.037 *
G	15477.21	1	15477.21	10.59	.001 **
SEX	725.42	1	725.42	.50	.481
IC x G	1629.86	1	1629.86	1.12	.291
IC x SEX	877.28	1	877.28	.60	.439
G x SEX	7137.90	1	7137.90	4.88	.027 *
IC x G x SEX	618.94	1	618.94	.42	.515

1-tailed Signif: * - .05 ** - .01; G=Group; SEX=Sex; IC=Item Correctness.

5.2.4 Discussion and Conclusion

The previous experiment demonstrated that knowledge of the CAT behaviour negatively affects subjects' performance on that ability test and causes them to spend a longer time completing the task. It suggested that a form of feedback acts during the adaptive test which has a negative effect on testees' performance. The findings of this experiment support that hypothesis, showing that the control group (NFBG) got significantly higher mean scores than those who had been provided with item-by-item feedback (FBG). Also males got significantly higher mean scores than females. However, no group by gender interaction was found. This finding tallies with that of Wise et al (1986), who found that item feedback led to decreased test performance. It also agrees with Strang and Rust (1973), who found an increase in testing time with feedback. Moreover, this study shows that the FBG got higher post-STAI mean scores than the NFBG, suggesting that IKR increases anxiety as suggested by other studies (Strang & Rust, 1973; Wise et al. , 1986 & 1989; Harold et al., 1973).

Unlike other studies, this study attempted to assess the effect of feedback on the time taken to answer individual items. It found that providing feedback significantly increased response time. Also, answering incorrectly tooke more time than answering correctly. Moreover, the interaction between group and gender was significant. Although both sexes were disadvantaged in the feedback condition, male subjects were disadvantaged more than female subjects (8.69 as opposed to 1.62 seconds). On the other hand, although females took significantly more time with wrong responses than males, they spent significantly shorter time on correct response. This interaction was found to be significant. As a main effect, the experiment revealed that different feedback about the correctness of the previous item (right or wrong) affects the time taken to answer the next item. Contrary to what was expected, incorrect answers caused subjects to spend a shorter time answering the subsequent item than correct answers. These findings need more investigation to understand the underlying reasons for such differences.

Although the use of computers for testing made the process of providing item-by-item feedback more practical, the negative effects of such a process on testee performance can be seen to outweigh any possible usefulness. Wise and Wise (1987) go so far as to say that " the use of such feedback in computer-administered tests is not recommended until its effects are better understood". Of course, this study does not suggest eliminating the CAT because of such limitations. However, preventing the testee from knowing the basic function of CAT and eliminating feedback about response accuracy can prevent problems, such as those demonstrated here. It is important when interpreting testee's or group's results to pay attention to the possible impact of the conditions under which the measurements were made.

Further studies are needed to investigate other related issues such as the effects of feedback on different populations of testees such as high vs. low ability testees and different ethnic groups. It could be that more difficult items are more frequently failed by low ability testees and that failure raises anxiety which decrease their performance in the test. More studies are needed also to assess the effects of IKR on different types of tests (e.g. verbal and mechanical tests), in order that the dynamics of item-by-item feedback can be clarified further. It is important for CAT developers to understand the factors which affect the test's validity.

CHAPTER 6

CHAPTER 6

Experiment 5

The Differential Effects of Item Difficulty Arrangement on Performance Using Computerized Adaptive Tests.

6.1 Introduction

Normally, when a testee is ready to be tested using an adaptive test, the computer begins by presenting an item of medium difficulty, aimed at the average testee. If the answer to that item is right, the next item becomes more difficult. If the answer is wrong the computer presents an easier item. The computer avoids presenting very difficult or very easy items which do not contribute meaningful information about the testee. However, initial items of medium difficulty to one testee may seen as harder or easier for other testee. Testees whose ability level is below the middle have items coming at them from the top down (Wainer et al., 1990). As a result, low ability testees will typically experience early failure, and high ability testees may experience early success, until the test adapts itself to the testee's ability level (Wise et al., 1989).

When a testee attempts easy items first, he/she somehow becomes better able to answer subsequent items, and when he/she faces difficult items, his/her frustration depresses the ability to answer subsequent items correctly (Gerow, 1980). This is especially true for those testees of average or below average ability. With an easy to hard arrangement (E-H) of item presentation, it is more likely that testees will experience a higher level of initial success than when items are arranged from hard to easy (H-E) or randomly (R).

On the basis of the learned helplessness theory, Petiprin and Johnson (1990) stated that when encountering difficult items, testees experience symptoms of learned helplessness, for example, frustration and diminished self-efficacy, which could lead to decreases in performance on the next items. However, if the testees answer easier items first, the resulting self-confidence could improve subsequent performance on similar items.

With conventional paper-and-pencil tests, the effect of item arrangement has long been a concern of test developers and users, not to mention the dilemma which faces teachers when developing their own academic achievement tests. Research investigating this issue has yielded conflicting results. Some researchers have found that an E-H ordering is most beneficial to testee's performance, especially when testing time is limited (e.g. Sax and Cromack, 1966; Hambleton and Traub, 1974; Hambleton, 1986). With speeded tests, where usually there is not enough time to complete the whole test within the time limit, placing easy items early in the test naturally enhances testees' performance

(Leary & Dorans, 1985). This result has led the authors of many textbooks to suggest that the optimal ordering of test items is E-H (e.g. Aiken, 1988; Gronlund, 1981; and Hopkins & Stanley, 1981). Others have indicated no performance differences when different forms of the same test are developed (E-H, H-E, or R) (e.g. Plake, 1981; Plake et al., 1981; Huck and Bowers, 1972; Sweeney et al., 1970; Barcikowski & Olsen, 1975). Leary and Dorans (1985) reviewed a number of studies involving aptitude tests and found significant main effects for item difficulty arrangement in two out of five studies (one using a speeded test and one using a power test). Some studies have found evidence of student preferences for tests arranged from E-H over R or H-E order (Allison & Thomas, 1986; Tuck, 1978). Perhaps testees, from their cumulative experience with P&P tests, expect (Power) tests to be arranged from E-H. The expectation of item order in itself has been found to affect performance (Hambleton and Taub, 1974). This raises an important issue of what will happened, in the case of a CAT, when a testee is given a test arranged in a different order from what he/she expected.

Other researchers have investigated the possible interaction between item order and anxiety. Bradshaw and Gaudry (1968) found that anxiety increases for those who experience initial failure as a result of facing difficult items at the beginning of the test, whereas anxiety decreases for those who experience initial success. Wainer and Kiely (1987) feel that test anxiety and frustration are increased with inappropriate starting points. However, Plake, et al. (1982) and Smouse and Munz (1986) found a nonsignificant relationship between anxiety and item order.

One should, however, be cautious in generalizing these results to adaptive testing, since adaptive tests differ from paper-and-pencil tests in various ways, such as on testing time and number of items. Moreover, one major obstacle in these studies is the lack of effective control over the order in which testees approached the test items. The testee who faces a difficult item may jump to another easier one until he/she gains the knowledge, the confidence, or the time to come back and answer it. Allison and Thomas (1986) asked elementary, secondary school, and university students for the test-taking strategy usually used in taking an achievement test, and found that the students usually selected E-H as the preferred strategy, followed by the items as presented (e.g. randomly) as second choice, and the H-E strategy as the least liked.

With adaptive tests, and with some other computerized tests, the testee has no choice except to answer the items in the order they are arranged and presented. That is, the next item will not be presented unless the present one is been answered. This reveals more clearly any effects of item sequence on the testee's performance.

The purpose of this study, then, is to examine (a) the differential effects of item difficulty arrangement on performance using computerized adaptive tests; and (b) whether there is any difference between sexes and high vs. low ability testees. Also, (c) the interactive effects of anxiety and item difficulty arrangement will be investigated.

6.2 Method

6.2.1 Examinees

The subjects consisted of 80 university students enrolled from a sign-up sheet in different undergraduate courses at the University of Nottingham. There were 40 male and 40 female students. They ranged in age from 18 to 36 years, with a mean age of 20.53 years (SD=2.87). For all subjects, the first language was English, with a variety of major degree subjects. The subjects were randomly assigned to the two treatment conditions used in this study (40 subjects in each group).

6.2.2 Materials

Selby MillSmith Adaptive Ability Test (CAT-NA)

See experiment 4[1].

Interpreting Data (ID)

The ID is part of the Advanced Work Aptitude Profile & Practice Set (Advanced WAPPS) developed by Saville & Holdsworth Ltd. The test is designed to measure the

[1] I wish to acknowledge with thanks Selby MillSmith Ltd. for their help and cooperation, in providing and where necessary modifying their software.

ability to make correct decisions or inferences from numerical or statistical data. It is related to several numerical aptitudes and is intended to measure the testee's ability to cope with figures in a practical and realistic context. The test consist of 22 items and the time limit is 10 minutes. The test manual reported an internal consistency reliability (Cronbach's Coefficient Alpha) of 0.73. The test is used here to classify the subjects into high and low ability groups before carrying out the experiment.

State-Trait Anxiety Inventory (STAI Form Y-1)

See experiment 3.

6.2.3 **Procedure**

The subjects were randomly assigned to one of two experimental groups, easy items first (EIF), or hard items first (HIF). Both groups consisted of 40 subjects, with an equal number of males and females in each. Seven to ten days before the second session started, the Interpreting Data (ID) was administered to discriminate between high and low ability subjects. This test was administered in the paper-and-pencil format. The STAI was used both before and after administration of the CAT-NA to determine if differential change in situation anxiety existed between treatment group. All tests were administered according to the instructions in their manuals. The use of calculators was not allowed with the ID. All subjects received general test instructions, followed by directions for recording their answers. For CAT-NA, answer revisions were not permitted. Scrap paper, two pencils and an eraser were provided. The experiment was conducted in an individual cubicle, and the subjects were advised to call the examiner if help was required. After finishing both sessions, subjects were thanked for taking part in the experiment, debriefed on the aims of the study and given feedback about their results plus four pounds for participating in a one hour experiment.

The SX3U1 Notebook IBM compatible PC with LCD monochrome screen and a standard QWERTY keyboard was used. All data were transferred to a computer data file and analysed using the Statistical Package for the Social Science (SPSSPC+).

6.3 Results

The correlations between the variables used are shown in Table 6.1. A significant correlation was found between the CAT and ID (r=.44, p<.001). This correlation is less than that reported in the Selby MillSmith Adaptive Ability Test manual (r=.55). The other significant correlation was found between pre- and-post-STAI (r=.54, p<.001). As found in experiment 4, the correlations between CAT-NA and both pre-and-post STAI were non-significant.

Table 6.1. Correlations between the variables used.

Correlations:	CAT-NA	Pre-STAI	Post-STAI	ID
CAT-NA		.2437	-.1113	.4397**
Pre-STAI			.4401**	.1137
Post-STAI				.0387
ID				

1-tailed Signif: * - .01 ** - .001
CAT-NA=Selby MillSmith Adaptive Ability Test; STAI=State-Trait Anxiety Inventory; ID=Interpreting Data;

The dependent variables for this experiment were subjects' scores on the CAT-NA tests and post-STAI. The independent variables were the difficulty level of the pre-items (EIF and HIF), gender (female and male) and the ability level of the subjects (HA or LA). The tests of normality and homogeneity of variance for each dependent variables showed that the skews of the scores did not differ significantly from the normal distribution, and the variability of scores in each format was approximately the same. The data were analysed using three-way factorial analyses of covariance (ANCOVA), where the pretest state anxiety (pre-STAI) was the covariate.

The mean scores of the first seven items presented to both groups were calculated to check their difficulty level. The mean score for the easy items was 6.40 out of possible 7.00 (mean time taken= 126.37 sec.), suggesting that practically all subjects answered all of the items successfully. However, the mean for the hard items was 2.40 (mean time taken= 511.84 sec.), indicating that most subjects failed to answer the items successfully. Both initial item forms showed the intended difficulty level needed for this study. These scores were not part of the CAT-NA score.

Table 6.2 (Figure 6.1) shows the means and standard deviations for the dependent variables (CAT-NA and Post-STAI), ID, and the covariate (Pre-STAI), broken down by conditions, ability levels, and gender. High and low ability subjects in the HIF group scored lower in the CAT-NA than their counterparts in the EIF group. The effect of the difficulty level of initial items showed a parallel pattern for low and high ability subjects. Moreover, both the EIF and HIF groups got higher post-STAI mean scores than pre-STAI mean scores (Figure 6.2). The mean scores on the CAT-NA for both groups were less than that reported on the test manual (N=445, X=90.59, SD=16.82, based on a sample drawn from 11 different Universities and Polytechnics across the UK).

Male and female subjects (in both LA and HA groups) in the HIF group scored lower in the CAT-NA than their counterparts in the EIF group. The effect of the difficulty level of initial items revealed a somewhat parallel pattern for male and female subjects. Moreover, both groups got higher post-STAI mean scores than pre-STAI mean scores. Again, the mean scores on the CAT-NA for both groups were less than that reported on the test manual (N=445, X=90.59, SD=16.82, based on a sample drawn from 11 different Universities and Polytechnics across the UK).

Table 6.2. Means and standard deviations by group condition and ability level, for the CAT-NA, post-STAI, ID and the covariate.

	NA		post-STAI		pre-STAI		ID		
	X	SD	X	SD	X	SD	X	SD	N
EIF	84.85	15.06	39.95	7.79	37.15	8.18	15.42	2.50	40
LA	79.90	12.87	38.05	7.76	34.85	7.69	13.55	1.87	20
F	75.00	10.36	40.10	6.38	34.60	8.99	13.40	1.83	10
M	84.30	15.89	36.00	8.78	35.10	6.62	13.70	2.00	10
HA	89.80	15.76	41.85	7.54	39.45	8.19	17.30	1.38	20
F	89.70	17.81	42.50	6.60	40.20	6.86	17.50	1.26	10
M	89.90	14.38	41.20	8.68	38.70	9.67	17.10	1.52	10
HIF	78.10	14.60	42.02	9.50	35.70	5.35	15.75	2.88	40
LA	70.00	14.82	42.65	10.23	35.10	6.30	13.30	1.65	20
F	68.43	12.62	44.28	11.43	35.28	7.31	13.57	1.74	10
M	80.33	17.37	38.83	5.74	34.66	3.44	12.66	1.36	10
HA	84.20	11.81	41.40	8.94	36.30	4.29	18.20	1.32	20
F	86.67	14.61	39.16	11.35	36.00	6.38	18.00	1.67	10
M	83.14	10.86	42.35	7.99	36.42	3.34	18.28	1.20	10

Note: EIF=Easy items first; HIF=Hard item first; LA=low ability; HA=High ability.

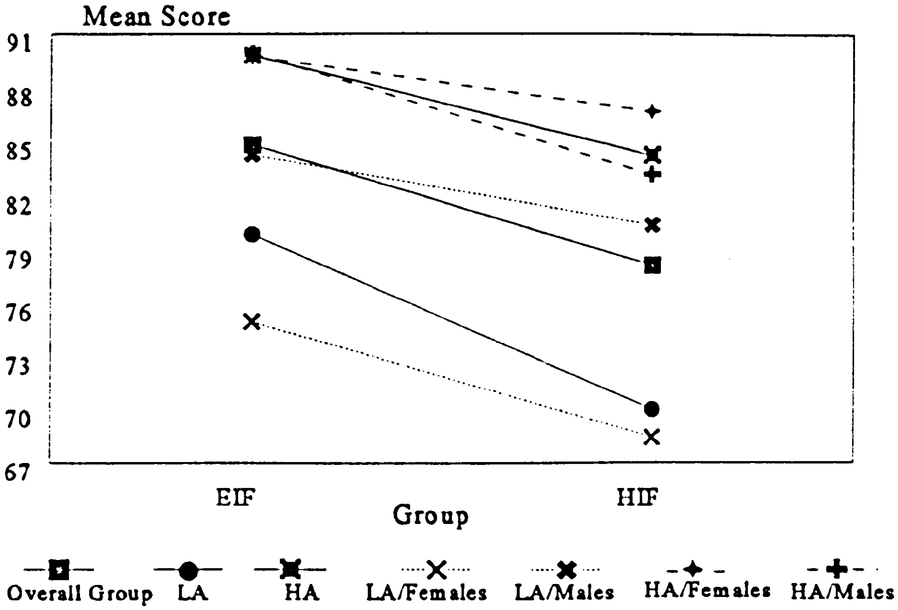

Figure 6.1 CAT-NA means by group, ability level, and gender.

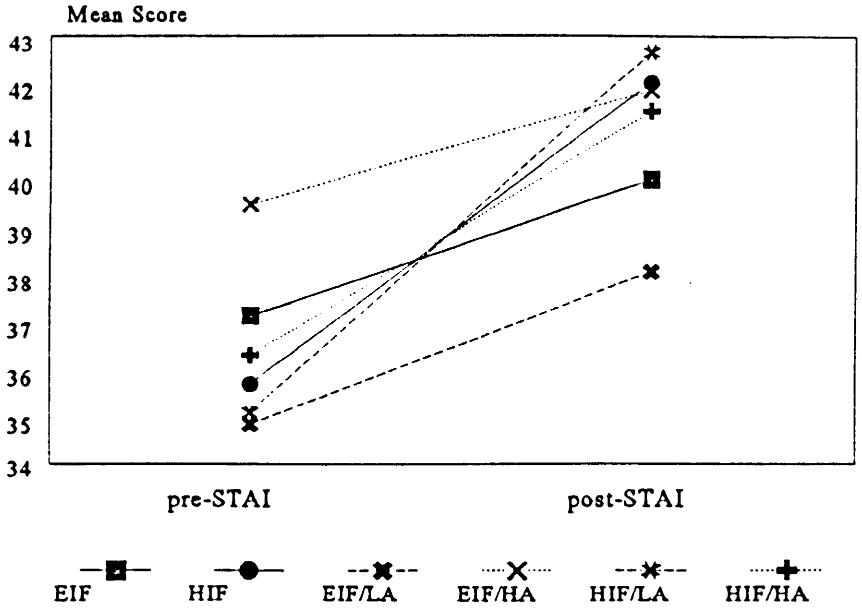

Figure 6.2 Pre-STAI and post-STAI means by group, and ability level.

The only significant differences revealed by ANCOVA (see Table 6.3) are for the main effects of group (F(1,73)=4.140, p=.046) and ability level (F(1,73)=8.158, p=.006) for the CAT-NA. The EIF group and subjects with a high ability level got significantly higher CAT-NA mean scores than their counterparts.

Table 6.3. Results of ANCOVA conducted for the CAT-NA and post-STAI with pre-STAI as a covariate for treatment group, ability level, and gender.

Source of Variation	SS	DF	MS	F	Sig of F
CAT-NA					
G	771.489	1	771.489	4.140	.046 *
ABG	1520.169	1	1520.169	8.158	.006 **
SEX	413.142	1	413.142	2.217	.141
Gx ABG	14.325	1	14.325	.077	.782
Gx SEX	.921	1	.921	.005	.944
ABGx SEX	608.823	1	608.823	3.267	.075
Gx ABGx SEX	71.095	1	71.095	.382	.539
Post-STAI					
G	167.323	1	167.323	2.724	.103
ABG	.005	1	.005	.000	.993
SEX	80.455	1	80.455	1.310	.256
Gx ABG	32.441	1	32.441	.528	.470
Gx SEX	8.181	1	8.181	.133	.716
ABGx SEX	155.397	1	155.397	2.530	.116
Gx ABGx SEX	19.331	1	19.331	.315	.577

1-tailed Signif: * - .05 ** - .01

G=Group; ABG=Ability Group; SEX=Sex

6.4 Discussion and Conclusion

The purpose of this experiment was to investigate the differential effects of initial item difficulty on performance using computerized adaptive tests. The results support the assumption that the difficulty level of the initial items has a significant effect on overall test score. Generally, those in the EIF group performed better than those in the HIF, regardless of their ability level or gender. Hard initial items decreased the scores of both high and low ability testees and vice versa. Although there was no significant interaction effect between group and ability level, Figure 6.1 revealed that those with a low ability level were affected relatively more negatively (deceased 9.90 points) by hard initial items than were high ability subjects (decreased 5.6 points). If we consider that the initial items of medium difficulty level are somewhat difficult for low ability and easy for high ability testees, then we will expect that those who perceived initial items as difficult items (low ability testees in this case) will perform worse than those who perceived initial items as easy items (high ability testees in this case). It seems that the initial average difficulty items used with CAT may enhance the performance of high ability subjects and depress the performance of those with low ability.

A similar pattern was found with pre- and-post-STAIs. Regardless of subgroup, subjects always ended the test with higher anxiety. Again, there were no significant interactions. Figures 6.2 revealed interesting results. Although those in the HIF started the test with relatively low anxiety, they ended the test with remarkably high anxiety.

A possible reason for the difference in results between this study and some of the other studies mentioned earlier could be attributed to the effective control over the order in which testees approached the items in this computerized adaptive test. Here, subjects were been forced to answer the items in the order in which they were been arranged, a factor which is difficult to control with traditional paper-and-pencil tests.

The notion that initially easy items will improve performance on subsequent items and that difficult items will lower performance on following items, as suggested by learned helplessness theory and others, is confirmed by this study. Figure 6.1 suggested that those with a low ability level were affected relatively more negatively by hard initial items than were high ability subjects. The effect was present for all subjects in condition groups, regardless of their gender and ability level.

One possible solution to the general effects of the difficulty level of the initial items is to provide 'a locator', which has items of different levels of difficulty to assess the testee's initial ability before the actual test begins. It is important that the testee knows that there is no score for these initial items. The suggested 'locator' can be considered as part of the examples, if appropriate. Another possible solution is to start from easy rather than average difficulty items, so all examinees experience early success (Wainer et al., 1990). Of course, this may increase the length of the testing session and limit speed of administration, thereby undermining one of the advantages of the adaptive test.

CHAPTER 7

CHAPTER 7

Discussion and Conclusion

7.1 Introduction

The goal of CAT is to avoid presenting items to testees that are either too easy or too difficult and thus less informative for the tester. This is done through CAT-tailoring the test to the testee by presenting items which have appropriate levels of difficulty. The objective of the work described in the thesis was to assess the psychometric as well as the psychological benefits and limitations of adaptive testing using both real employees and college students. Specifically, the thesis investigated the issue of equivalence between P&P and CAT formats; the predictive validity for both formats; the time taken to complete both formats; the effect of feedback on time taken to answer items; computer anxiety, testee's reactions and attitudes towards computer adaptive testing; the effects of previous computer experience and prior test experience on performance and ability to identify CAT; the effects of feedback and knowing CAT behaviour on testee's scores and anxiety; and finally, the differential effects of item difficulty arrangement on performance in adaptive tests. To achieve these goals, one field study and four laboratory experiments were undertaken.

This final chapter is classified into four sections. The first section summarizes the key findings. The second section discusses these findings and their implications. Section three gives some guidelines and advice for future research. Finally, the fourth section presents some concluding comments.

7.3 Summary of Findings

After a general preface and two introductory and literature review chapters on computerized-based testing (CBT) (Chapter 1) and the computerized adaptive test (CAT) (Chapter 2), the objectives, methods, results, and discussions of one field study and four laboratory experiments were described in Chapters 3 to 6. The main findings presented in this thesis can be summarised as follows:

1) The current research strongly suggested the equivalence between the P&P and CAT formats for the AR and MR tests of the Differential Aptitude Tests (DAT), but failed to do so for the NA test. One of the formats of the NA test needs to be rescaled to be equivalent. With all tests, the mean score for the second administration was higher than for the first, suggesting that practice effects were taking place (Experiment 1, Chapter 3 & Experiment 2, Chapter 4).

2) The CAT version of DAT can predict a performance variable at least as accurately as can the P&P format (Experiment 1, Chapter 3).

3) Overall, testees' attitudes toward several aspects of computerized testing were positive. The results confirmed the negative relationship between computer experience and computer anxiety but did not reveal any relationship between computer experience and CAT scores, nor a relationship between computer anxiety and CAT scores. Previous experience of a computerized test also did not significantly affect CAT scores (Experiment 2, Chapter 4).

4) Knowledge of CAT behaviour negatively affected subjects' performance, and caused them to spend longer completing the tests, but did not increase the level of their state anxiety (Experiment 3, Chapter 5). It seems that a form of feedback acts during the adaptive test which has a negative effect on testees' performance and response time. This assumption was confirmed (Experiment 4, Chapter 5). Different knowledge about the correctness of the previous item (right or wrong) affected the time taken to answer the next item (Experiment 4, Chapter 5). That is, subjects spent a shorter time on the next item when getting negative feedback (wrong) on the previous item than with positive feedback (right). Also, where there was no prior knowledge about the basic properties of CAT, self-identification of adaptive behaviour by the testee affected his/her scores and delays the completion of the test (Experiment 2, Chapter 4 & Experiment 3, Chapter 5).

5) Using CAT version of DAT resulted in a 20% reduction in time to complete the test. The reduction in testing time is due to the presentation of 20 items in the CAT mode, compared with 40 items in the P&P mode. However, the response time for answering

each individual item was higher for the CAT format than for the P&P format (Experiment 2, Chapter 4).

6) The initial average difficulty item(s) presented by a CAT seems likely to enhance the performance of high ability subjects and to depress the performance of those with low ability. The difficulty level of the initial items had a significant effect on overall testees' scores, regardless of their ability level . Also, subjects always ended the test with higher anxiety (Experiment 5, Chapter 6).

7.3 Implications and Discussion of Results

7.3.1 Occupational Psychologists and the Use of Computerized Tests in Organizations

Occupational psychologists are interested in the areas of counselling and personal development, employee relations and motivation, the design of work and its environment, organizational development, personnel selection and assessment, human-machine interaction, performance appraisal and career development, and training. They work in small as well as large organizations, where usually computers are part of the work environment. Schoenfeldt and Mendoza (1991) showed that many of the activities carried out by occupational psychologists lend themselves to possible computerization or automation. Included among these activities are selection, placement and training of employees, the design and management of performance evaluation systems, the development of a system to handle career progression, and planning of organizational interventions. They presented examples showing computers linking areas of human resource and organizational behaviour that have been historically seen as distinct things. For example, selection systems incorporate input from other segments of the organization, and job analysis information is associated with planning data as an indication of current job activities/ requirements and future activities/skills that will be needed, so the selection can be more dynamic, and new employees can be chosen to have the capabilities needed for future organizational changes.

Although the focus of this thesis is on the application of computers in testing for the purpose of selection and assessment, it is important to notice that computers are used with other activities inside the organization for the same purpose. Computers have been used in selection scenarios for conducting interviews, keeping applicant records, presenting check lists and rating scales, and for self- and psychological-monitoring. The logic used in CAT for human assessment has been exported to other areas of interest to occupational psychologists. Coovert (1986), for example, showed how a computerized adaptive process can be adapted to generate task statements for a job analysis, so that the computer presents only those tasks which are relevant for the job. However, to date, as

Denton (1987) noticed, most occupational psychologists use computers only for complex statistical analysis.

Despite the possible limitations mentioned throughout this thesis, experimental and live-testing studies indicate that CAT works. However, developments in CAT have been fairly slow to percolate through to selection and assessment uses. Most of the software available to date is limited largely to educational and military uses, and most has been developed in the United States. The only adaptive tests that the author knows to have been developed in Britain are the Selby-MillSmith adaptive test, and the Micropat mentioned earlier. However, these are not IRT-based tests. The author had great difficulty during the research in finding sufficient number of CAT programs to use with the experiments carried out. Many reasons have been suggested for this slow adaptation of CAT to occupational uses. First, the development of IRT-based CATs is a costly process, and unless there is a large number of candidates to be tested, the price for better measurement is high. Second, the concept of the adaptive test is still complicated and vague for most test users and even for psychologists. Perhaps the mathematical complexity of the theory is the main reason. Third, as Smith and Robertson (1989) noted, publications on CAT tend to occur in journals which are not routinely scanned by personnel psychologists (such as the *Journal of Educational Measurement, The Journal of Computer Based Instruction, Behaviour Research Methods & Instrumentation, and the International Journal of Man Machine Studies*), and hence, the potential benefits that computers can offer to human assessment are often unrecognised. Fourth, the numbers of companies which are currently using P&P tests (and computerized tests) for selection and assessments, although expanding, are still relatively small; and those which use basic computerized tests seem happy with what they have got. Fifth, some types of tests available now do not seem to profit much from computer technology. These tests include work samples, planning tasks, tests with flowcharts, 'what-would-you-do' tests, open-ended tests, and tests consisting of less than 20 items. Broad general ability tests, and tests which require a complex administration and scoring procedure, seem most likely to benefit from computerisation.

All these reasons for the slow adoption of CAT for occupational use (and others mentioned in chapters 1 and 2 about the limitations of CBT and CAT) have discouraged some researchers like Schmidt (1993), who wrote sadly that " Computerized testing-and especially computer-adaptive testing (CAT)-has a twenty-year history of being the wave of the future that will sweep all before it, and then failing to live up to its promise. ...CAT is an important element in what we have to offer organizations. but it has turned out to have more limitations than we initially realised".

The intervention of the computer in organization life has solved many crucial problems, but created others as well (see e.g. Frese (1987)). As seen from the literature reviews conducted earlier in Chapters 1 and 2, computers have contributed by solving important problems and improving productivity for organizations (e.g. saving time, cost and effort, and provide more precise measurement of testee's ability), whilst, on the other hand, creating important issues (e.g. computerphobia and serious legal issues). Gardner et al. (1986) investigated the effects of computers on the delivery of services, where computers were used, among other things, for assessment and evaluation. They reached the conclusion that " for every promise there are a dozen corresponding pitfalls, each one waiting to engulf individuals and systems and to create as many new problems as the innovations solve" (p.155). Also, the major conclusion Frese (1987) reached was that issues of human-computer interaction cannot and should not be separated from organizational issues. Another problem facing organizations when they come to decide the proper computer hardware to purchase for testing, is the rapid improvement of computer equipment, which makes the choice risky. It is difficult in this situation to keep pace with new CBT standards, that are replacing those of yesterday's.

Finally, occupational psychologists need to be aware of the legal and ethical issues associated with the use of computerized tests. Computerized (adaptive) tests may be vulnerable to many of the same legal attacks as P&P tests. As with conventional tests, occupational psychologists should ensure that tests comply with good testing practice, including such ethical issues as fairness and privacy. CBT should be carefully examined to ensure that it does not produce new legal questions for the practice of human assessment. For example, unfamiliarity with computers could be correlated with ethnicity, gender, age, and socioeconomic status, so any effects due to unfamiliarity might appear statistically as poorer performance by some groups (Bersoff & Hofer, 1990). Violations may arise from different intentional or accidental practice, such as careless entry of data, presenting a response device that is not appropriate for testees, unreasonable interpretations of the results reported by computers, or, broadly, unfair discrimination between sexes or ethnic groups. However, the great standardisation of CBT administration gives the computerized test an advantage over P&P tests.

7.3.2 Equivalence Between P&P and CAT

One rarely finds a P&P test developed later as another version of a computerized test. It is usually the reverse. If we need to make use of the psychometric characteristics of the P&P format in a computerized form, an important question arises concerning the equivalence of scores of tests administered by the different methods. As argued in Chapters 1 and 2, it cannot be assumed that scores from one format of a test are equivalent to those of another one. It may seem that to transfer a test item from P&P mode to computerized mode does not make much difference. However, even when the two versions are made as similar as possible (e.g. matching content, colour, and layout), differences may still be present. Therefore, test equivalence must be established empirically, and until this has been done, the computerized version cannot be considered a substitute for its P&P counterpart. The current research strongly suggested equivalence between the two modes for the AR and MR tests of DAT, but failed to do so for the NA test. Because the correlations between the scores of the two forms are relatively high, suggesting that both formats measure the same construct, the CAT version of the NA test can be equivalent to its P&P version if the scores of either format are rescaled so as to be comparable with the scores of the other test (APA, 1986). It would probably be easier to rescale the CAT scores to maintain the P&P norms and cutting scores.

For the NA test, the mean shift was slight, amounting to about two raw score points, in favour of the CAT version. For the AR and MR tests the two formats retain somewhat more similar mean scores. Previous researchers found that scores were sometimes lower for the computerized test than for the P&P test (e.g. Kovac, 1989; Lee et al., 1986), and sometimes higher (e.g. Greaud & Green, 1986). Kovac (1989) suggested that each ability test should be examined separately before any general conclusion on the whole battery of such tests can be achieved.

A probable explanation for the non-equivalence found with the NA test would be that the CAT format did not allow the possibility of looking back, reviewing items, altering responses or delaying the attempt to answer questions. In fact, allowing testees to do so has raised objection from some researchers such as Green et al. (1984), Lee et al. (1986), and Sarvela (1991). They argued that if a testee modifies an answer afterwards, a recalculation based on all the items administered thus far should be carried out. The items selected in the foregoing part of the session will no longer be the most appropriate, given the recalculated estimate of the position on the latent trait. CAT relies upon a response to each item as it is presented; consequently, item skipping or preview is difficult to reconcile with it. As Noonan and Sarvela (1991) predicted, " a test-wise student could notice that the items are getting easier and decide to go back and change earlier answers. Also, students could try to look back at items continually in order to get clues that help them answer other items". Careful work carried out by Lee et al. (1986) and Sachar and Fletcher (1978) showed that the difference between mean scores of the two formats they obtained was attributable to not allowing review of earlier items. Green (1991) suggested that if a test is given in P&P simulation of the computer, one item per sheet, with no looking at earlier sheets, the difference will disappear. He argued that if the computer is not to allow review, the score may need adjustment before using P&P norms, because the P&P version allows review. It would be interesting to see what happened if we prevented testees from backtracking on the P&P format of the NA test, in order to simulate computerized testing. On the other hand, two empirical studies carried out by Lunz et al. (1992) and Wise et al. (1989) did not support the above assumption. They found that testees' changes are very small, to the extent that final ability estimates do not significantly differ whether changing previous answers is permitted or not. Disallowing review annoyed testees, although developers and users may see it as a safeguard to prevent the omission of any item that the testee should answer. Ironically, examinees are always advised in school, or when taking selection tests, to check their answers and to ensure they did not fill out the wrong answer. Now, suddenly, we are asking them to forget what might appear to be their 'right' to revise and modify their answers. Finally, while the role of CAT is to provide additional and more accurate information about the testee, the possibility exists that other information may be lost instead.

Computerized testing may look, from the testees' point of view, more interesting, novel, and challenging. The novelty of the medium used for assessment may increase the testee's motivation to answer the test items. It has also been presumed that CAT may have positive motivating effects on testees by presenting them with items that are sufficiently difficult to present a challenge (Betz, 1977), which may enhance their test performance. Many years ago, Vroom (1964) suggested that performance is determined by ability and motivation. Thus, whatever the testee's actual ability, motivational characteristics contaminate measures of ability by changing performance on these measures in ways which are situation-specific (Hyland, 1987). If we accept that the goal of ability testing is to predict applicant performance in real job situations, then it is also important to understand how motivational factors can affect performance, and hence the equivalence between the two formats. This effect may tend to fade as new generations become gradually more used to using computers.

The size and the quality of the computer screen can also have an effect on what are usually graphical, coloured, or long test items. As the author noticed from the testees' response on the attitude questionnaire (Experiment 2), they experienced difficulty in reading from the computer, and some physical problems during the test session such as eye fatigue and headache. Moreover, although instructions are more standardised on the computerized method, it is a difficult task to keep instructions of both formats equal. Answering items using P&P differs substantially from interacting with a computer screen and an input device. During the introductory stage of the P&P test, for example, the tester normally tries to establish a friendly climate with his/her testee, and pay individual attention to his/her inquiries. However, confronting computers in an early stage of the testing session may be seen by test takers as an unnatural starting step.

Practice effects were detected during the study. There was an increase in scores on the second administration. The increase was relatively greater when the CAT followed the P&P test, suggesting that CAT benefits more when subjects practise with P&P first. This is a further reason why organizations should not consider using both formats of the NA test as an equivalent, nor exchanging their norms, cut scores or predictive validity unless a proper rescaling for the scores is conducted.

7.3.3 Predictive Validity and Utility of CAT

Another aim of the thesis was to assess the predictive validity of the NA, AR, and MR of the DAT for the oil refinery training programme (Chapter 3). The criterion selected for the success in the training programme was the students' scores in the training courses. In test validation, a discrimination is often made between job performance and training criteria. Tests which predict success in training courses may not predict job success and vice versa. Therefore, a follow-up study should be done to collect data about trainees' performance in the actual job in order to assess the validity of the test in that job.

Ghiselli (1973) attempted to generalise the results from validation studies of 20 categories of personnel selection tests for 21 occupational categories, carried out in North America over the past years. He found that the overall average correlation of ability tests with training performance was .45, and .35 with job performance criteria. Smith (1986) and Hunter and Schmidt (1989), using meta analytic techniques have supported these findings. Perhaps the correlations achieved in this study with the overall performance, particularly for NA tests (Table 3.13), are not so far from these figures. The CAT format gave a marginally higher correlation with trainees' performance than its P&P counterpart, but no significant differences between correlations were found (using Fisher's r-to-z transformation). These findings confirm those by Bloxom (1989), who found that CAT and P&P formats produced similar validities, regardless of the assumed model (e.g. Rasch, two or three models). Garrison and Baumgarten (1986) also compared the efficiency and the predictive validity of adaptive one-parameter and conventional formats of a test of knowledge of mathematics. They found that the CAT was more efficient (i.e. the average information function of the adaptive items exceeded the average information function of the P&P test, over the entire range of theta). They compared the tests' correlations with a measure of academic performance at the end of the school quarter in which the tests were administered, and found very similar validity coefficients for the two formats. But, what do these correlations mean? Are these validity coefficients likely to improve selection? And at what cost? The major question here is whether the benefits of using CAT in place of the P&P test are sufficient to justify the costs.

Realising that both formats of a test are commercially available, and that the benefits to employers of using ability tests are clear, and assuming the equivalence between the P&P and the adaptive formats is established , the next question became, which format should be used for testing? Murphy and Davidshfer (1994) noted that the question is not whether to use computerized tests but, rather, when and how. In fact, many factors determine which will be the best format. Such factors are: costs; predictive validities; testees', selection officers', and employers' acceptance of the testing method; period of use (long/short term investment); and the kind and number of testees that need to be tested. The advantages and limitations mentioned in the previous chapters of using CBT in general, and CAT in particular, should also be considered. Some of these factors have already been discussed. Cook (1988) believes that the cost of selection is fairly easy to calculate, but it is more difficult to calculate the return on selection. When he came to answer the question of how to select selection tests, he named five criteria: validity; cost; practicality (whether it is easy to administer the test in the situation specified); generality (the number of types of employees the test can be used for); and legality. At that time, he was speaking about the P&P format test. However, these criteria can also be generalised to evaluate computerized (adaptive) tests against P&P test formats.

As we saw earlier, little difference in validity between the two formats is to be expected. A practical test is one that is not difficult to introduce. Practicality is determined by, for example, the time required for testing, the level of preparation required from testers, and the suitability of the testing venue. The P&P test is easy to administer, move, and handle, and does not required electricity, or large rooms. The P&P format tests are familiar to many applicants. Generality is mainly determined by the kinds of tests used, regardless of the mode used for administration. However, in some specific situations, like administering tests to people with disabilities, computers can be equipped to handle the situation more efficiently than can P&P tests. Concerns about legality were discussed earlier and in Chapter 1.

Many authors have been rather optimistic about the cost effectiveness of computerized testing in the long term. For example, Wiess and Vale (1987) stated, "The time saved by administering tests by CAT can be used for additional testing of other abilities to increase the predictive validity of a test battery...". Green et al. (1982) also stated that "The introduction of the computer-administered adaptive version of the ASVAB has great potential for improved personnel assessment in the Armed Forces, quite apart from the immediate savings realized in the recruit testing process". Robertson and Smith (1989) claimed that in many cases where adaptive testing is used, costs could be reduced by up to 50 per cent.

However, no literature on specific cost analyses has been found. In most cases, the reduction of items associated with using CAT is used as an efficiency measure, but the cost is not always considered. For example, trying to develop a realistic expectation about the psychometric benefits of CAT from studies employing real data, Bloxom (1989) found that adaptive tests are more efficient than the P&P format, i.e. they require fewer items than P&P formats to obtain the same level of measurement precision. The reduction in testing time produced by CAT, from a practical point of view, is overestimated. In the current research, as mentioned earlier, a reduction of 20% in completion time was been found. In a practical situation, when we have, say, 20 testees allocated to 20 machines, we will find that some testees will complete the untimed adaptive test in a time longer than that allocated for the P&P version (in the present study, one testee took 54 minutes, although the time specified for the P&P version is 30 minute, he could not be stopped!). The tester has to wait till those testees who are not finished have completed the test(s). Although most CAT testees finish earlier than people taking the P&P test, the time gained from those who finished earlier is wasted by waiting for those who take longer than usual. The tester cannot leave the testing room when some people are still struggling with the test. The situation may be even worse, when this unexpected delay destroys the pre-specified testing schedule.

Costs of implementing CAT normally include hardware, software, operating, maintenance of tests and hardware, and training. How many people need to be tested at one time should be considered before deciding how may machines need to be purchased and how many supervisors need to be trained. The availability of any of these on site (such as computers used for other administrative and secretarial activities) obviously helps to reduce the cost. It should not be expected that the cost of purchasing an adaptive test will be similar to that of conventional P&P or CBT formats of the same test. Developing a CAT version of a test is time and money consuming, and this is reflected in the test price.

In his study within Dutch Railways, Schoonman (1989) estimated the profits arising from implementation of CAT to come from four sources: lower costs for P&P tests; higher efficiency in administration; easier maintenance of the tests in use; and higher predictive validity. He advised that from economic, practical, and scientific points of view, it seems unwise to spend a long time on the development of a full three-parameter-based testing system which is more expensive and only slightly better that two-or one-parameter models. The Rasch model, he argued, is a proper choice in starting IRT-based testing. He proposed a 'scenario' for implementation of computerized test, which of course will not happen overnight. Computerized (adaptive) tests will not be suitable for every organization. Organizations which select a small number of candidates each year may find P&P methods the most cost effective.

7.3.4 Attitudes Towards Computerized Adaptive Test

The attitude of candidates is an important issue when considering the use of new technology. Negative psychological reactions may slow both the acceptance and the useful application of computers, and may affect the data obtained during the testing process (Romanczyk, 1986; Munger & Loyd, 1989). Personnel selection is an activity that influences testees' as well as society's image of an organization. Whether employed or not, it is always good practice for the candidate to feel that the testing process is fair and does not prevent him/her from exhibiting his/her actual ability. Selection should be viewed as a process which has impacts on the candidate as well as the organization. The psychological impact of rejection or acceptance on candidates may vary according to the context in which the decision is taken, the method used, and the personal characteristics of candidates (Robertson & Smith, 1989). If candidates view computerized testing as a method which does not reflect their true ability, for one reason or another, the impact on them will be painful. If we accept that the practical goal of ability testing is to predict performance in real situations, then it is also important to know how attitude toward the computer might affect performance (Glass & Knight, 1988).

In this study, the overall testees' attitude toward computerized testing was positive. The reasons for this positive attitude related to potential for quick feedback, clarity and simplicity of method, lack of time pressure, shorter completion times, and less fatigue and effort. Most subjects found the idea of administering tests on a computer fun, interesting, and novel. Additionally, most subjects believed that they were given enough time to answer. On the other hand, most of the subjects complained about the inability to go back and review their answers. Other physical problems during the test session were experienced, such as eye fatigue and headache. Some of the subjects even found it difficult to read from the computer screen and adjust to the method. The small LCD monochrome screen may have been responsible for such problems. Future research should ensure reasonable size and a clearer screen to avoid such physical problems which may affect test validity. But generally, the CAT was well accepted by subjects. This acceptance may be enhanced with improvements in software and hardware design. When the subjects were asked whether taking a computer-administered test affected their anxiety level compared with taking a P&P test, their response correlated positively with the difficulty of reading questions from a computer as opposed to reading them in a P&P form; what they thought about the sufficiency of time given to answer; and the perceived difficulty of interacting with a computer.

In Experiment 2, the author was confronted with subjects of whom 97.5% had used a computer before. Their experience with computers was shown to be significantly and negatively correlated with computer anxiety, and positively correlated with a preference for taking computerized tests. However, this experience did not show any effect on the subjects' performance on the CAT. The author expects that practice sessions before actual testing help eliminate any possible advantage that more experienced subjects have over those with less experience. In fact, the APA guidelines (1986) encourage users to train test-takers on proper use of the computer equipment before the actual test begins.

That high level of arousal caused by high level of anxiety disrupts performance is well known. However, there was no relationship between anxiety and performance in any of the experiments done here. One possible explanation is testees' knowledge that no important decision resulted on their performance. In reality, when a decision has to be taken about hiring or firing someone, the situation may be different. Indeed, in situations where studies are designed to assess the equivalence between the P&P and the computerized formats using real candidates, candidates are usually aware of the fact that their performance on the computer mode, usually, will not be used for decision making, and it would be unethical for the tester to say otherwise.

Quible and Hammer (1984) argue that employees "fear being replaced by the new sophisticated, more efficient equipment; and they experience anxiety when confronted by having to learn to use new sophisticated, technical equipment". Therefore, any efforts, as Bloom (1985) suggested, to treat computer anxiety should be directed toward attitude change.

As the results of the current thesis have revealed, preference for taking a computerized test correlates negatively with anxiety caused by taking computerized tests, and with the difficulty of interacting with the computer. Preference for the computer format compared with the P&P format correlates positively with a belief in the possible advantages of computerized tests over P&P tests for job selection purposes. Compared with P&P formats, CAT would be a more enjoyable experience for low ability testees, because they would mainly be confronted with items they can answer, instead of a lot of difficult items.

As we try to perceive how testees and users react to computerized testing, we should be aware of the possibility of individual differences in acceptability of this method of assessment. In their comprehensive study of 'computerphobia', Brosnan and Davidson (1994) noticed that the majority of studies reported that between one quarter and one third of their sample, of whom females were the majority, registered as computerphobic. They indicated that computer literacy may become a 'critical filter' for women who will therefore not have access to the many careers which need skills in dealing with computers. Conversely, Anderson (1987) found that the sexes did not differ in their level of anxiety. Older students, as he noticed, revealed less anxiety than younger ones, something which the author's own results confirmed. Kratochwill et al. (1991) stated that "rather than seeking an answer to the question of whether computers should be used in assessment, perhaps we should attempt to identify types of individuals who may be especially uncomfortable with computerized assessment and attempt to design environments that make use of computers more acceptable to these groups".

The questionnaire in the current study, however, was limited in scope. Due to time constraints, subjects were not asked about their attitudes towards P&P testing. Research on the acceptability of computerized testing needs to be more methodologically sound before any rigid conclusions can be drawn. Romanczky (1986) pointed out that the groups used for study may yield significant differences in reports of acceptability. Witt and Elliott (1985) noticed that in existing studies, measures tend to be quite informal and lack the psychometric characteristics (systematic, reliable, and valid) necessary to draw valid conclusions. Information is needed about acceptability of computer assessment from the individual who draws conclusions, makes inferences, and develops assessment programmes (Kratochwill et al., 1991).

7.3.5 Noticing and Knowing the Adaptive Behaviour of CAT

The current research has addressed a practical issue concerning whether testees' awareness of CAT behaviour (without prior knowledge about the basic properties of CAT) influences their performance, and found that around one quarter of the subjects identified the tailored behaviour of CAT, and consequently obtained significantly lower mean scores and spent longer time completing the test than those who did not. This finding may suggested that a cognitive process takes place during the test about the difficulty level of the present item compared to the previous one, which affects negatively the testee's performance.

As most adaptive tests start testing by presenting item(s) of medium difficulty level, it is presumed that those with high and low ability will notice CAT more than those in the middle of the ability range because the decrease (for high ability testees) and increase (for low ability testees) in item difficulty will be more obvious for them compared with those with average ability. If items become easier and easier (mostly with low ability testees) it could mean that the testee is making more wrong answers. That could increase anxiety and decrease performance, and vice versa. It is also presumed that testees will be more able to notice the fluctuation in the difficulty level of the items at the beginning of the test than at the end of the test. As far as the author knows, these hypotheses have not yet been tested by any researcher.

Experiment 3 has demonstrated that knowledge of CAT behaviour negatively affects a subject's performance on the ability test used, and causes him/her to spend a longer time completing the task. More than half of the subjects informed about CAT tailored behaviour believed it made them perform more slowly and worried them. The experiment found that a form of feedback acts during the adaptive test which has a negative effect on testees' performance and anxiety. As a result, if all items are not totally independent, there is a danger of contaminating future items. If a testee is consistently answering items incorrectly, the negative feedback (inferred from lower difficulty level of subsequent items) can be destructive to motivation on future items and vice versa. (Noonan & Sarvela, 1991). Test developers often worry about the motivational consequences of subtle cues about the correctness of the testee's responses (Wechsler, 1974).

Experiment 4 attempted to assess the effect of feedback on the time taken to answer individual items. It was found that providing feedback significantly increase response time. Also, answering incorrectly took more time than answering correctly. On the other hand, although females took more time with wrong responses than males, they spent a shorter time on correct response. The study revealed that different feedback about the correctness of the previous item (right or wrong) affects the time taken to answer the next item. Contrary to what was expected, incorrect answers caused subjects to spend a shorter time answering the subsequent item than correct answers. These findings need more investigation to understand the underlying reasons for such differences.

It seem unwise, on the basis of these findings, to inform the testees before the test session about the actual function of the adaptive test as this may affect their performance and delay their responses. This implication seems more important when the decision based on the testee's results takes into account the amount of time he or she spends to complete the test, and when the decision is of a vital interest to him or her. It does not seem desirable to inform testee about either the correctness of his/her answer or their estimated position on the trait being measured. Both types of feedback could be demotivating. With some types of adaptive test which have no fixed number of items, it seems also difficult, and undesirable, to inform the testees about the number of items or the time remaining. However, not to do so could be demotivating because the testee has no cues where or when the test will end. In other words, the testee has no control over the situation.

7.3.6 The Differential Effects of Item Difficulty Arrangement on Performance

The current research investigated the differential effects of initial item difficulty on performance using computerized adaptive tests and found that the difficulty level of the initial items has a significant effect on overall test score, regardless of whether the testee is in a high or low ability group. Hard inital items decrease the scores of both high and low ability testees, and vice versa. If we consider that the initial items of medium difficulty level are likely to seem difficult for low ability and easy for high ability testees, then we will expect that those who perceived inial items as difficult items (low ability testees in this case) will perform worse than those who perceived initial items as easy items (high ability testees in this case). It seems that average difficulty level items enhance the performance of high ability subjects and depress the performance of those with low ability. It was also found that subjects always end the CAT test with higher anxiety. In an interesting study by Wise et al. (1986), item feedback did not affect anxiety or performance when items were presented in an E-H order, but anxiety increased and performance decreased with random presentation of items.

This may be considered as one of the limitations of CAT and developers should acknowledge this problem. As a possible solution, developers may programme the test to start with a 'locator', which has items of different levels of difficulty to assess testee's initial ability before the actual test begins. The proposed locator has no scores, but is used to estimate roughly the testee's ability and help estimate the best inial items for him/her to start with. The suggested 'locator' can be considered as part of the examples, if appropriate. Another possible solution is to start from easy rather than average difficulty items, so all examinees experience early success. However, this may increase the length of the testing session and limit the speed of administration which may reduce the efficiency of CAT.

As most P&P tests start with easier items than their CAT counterpart, the result provides us with one possible reason for the unequivalence found in some cases between these two formats. Users of CAT should take into account the effects of inial items on both high and low testees' scores, and hence interpret the scores with caution.

One of the important issue emerging from this study seems to be the evidence it provides against the practice of making the order of presentation of items in a test different for different testees, to reduce the chance of cheating without caring about the order of the difficulty levels of the items.

7.4 Future Research

Suggestions for future research have been mentioned in different places throughout this thesis. This section provides suggestions and advice for future research. Some of these suggestions are drawn from the empirical work carried out here, others are from calls encountered in the literature reviews. The suggestions are as follows:

1. The current research used only numerical, mechanical, and abstract reasoning tests for a specific population of trainees and university students. Future research may include other kinds of testees and real job candidates, and examine additional types of tests, to further the knowledge of CAT effects.

2. Future research should explore the effects of using CAT with particular groups (e.g. by sex, race, age, education, and those unfamiliar with computers) as a primary variable of interest rather than as a secondary area. Research in this area is scanty.

3. Unfortunately, the choice of the three tests used with the field study was not carried out by job analysis study as normally should be the case. Job analysis studies are extremely useful in indicating how relevant the tests are, and what cut scores distinguish good from bad performance. The Kuwait Petroleum Corporation claimed that a job analysis study for the refinery job had previously been done, but refused to make it available to the researcher because, as they said, it was confidential. Therefore, the selection of these three tests was done by quick scanning of the training curriculum to help identify the tests which appear relevant to the job. The author wishes that he could have included the Space Relations test to the battery, but the time factor was critical.

4. For assessing the testees' previous experience with computers, the questionnaire used only one self-report question (Q11. "How often have you used a computer before taking this test?") rather than a series of questions to assess the amount of computer experience.

This may not be enough. A well structured measure should be used in future for such a purpose.

5. Further studies are needed to investigate other related issues such as the effects of feedback on different populations of testees, e.g. high vs. low ability testees and different ethnic groups. It could be that more difficult items are more frequently failed by low ability testees and that failure raises anxiety which decrease their performance in the test. More studies are also needed to assess the effects of IKR on different types of tests (e.g. verbal and mechanical tests), in order that the dynamics of item-by-item feedback can be further clarified.

6. Not a single study has been found on the effects of knowledge about the basic function of CAT on testees' performance. The experiment conducted in this thesis seems to be the first attempt. It is important for future researchers to replicate this study and to extend it with other populations such as high and low ability subjects. It would be interesting also to investigate other ability tests. This study used only Numerical and Abstract Reasoning tests.

7. Due to time constraints, subjects in the present research were not asked about their attitudes towards P&P testing in order to make a comparison between testees' attitudes to the two modes of testing. Future research on this issue needs to take more account of this limitation.

8. It is presumed that those with high and low ability will notice CAT more than those in the middle of the ability range, because the increase (for high ability testees) and decrease (for low ability testees) in item difficulty will be more obvious for them than for those with average ability. It is also presumed that testees will be more able to notice the fluctuation in difficulty levels of items at the beginning of the test than at the end of the test, until the test approaches the testee's level of ability. These hypotheses have not been tested and certainly warrant further research.

9. P&P tests tend to use items which are most discriminating for those testees in the middle of the trait continuum, whereas CAT is more accurate than P&P tests for assessing those testees who are much higher or much lower than other testees on the dimension being assessed. This hypothesis has not been tested empirically (Bloxom, 1989).

10. CATs help reduce test anxiety by presenting items which challenge but do not discourage testees. This helps maintain a constant level of motivation in answering the test items. No study addressing this issue has been found. Studies are needed to investigate this hypothesis.

11. To date, the only factor used for adapting the difficulty of the CAT is testees' answer(s) to previous item(s). Theoretical as well as practical bases are needed to assess the possibility of adapting the difficulty of the test according to testee's level of anxiety as well as his/her answer to previous item(s). That is, easy items are presented when the

testee's anxiety level is assessed as high, to control the possible effects of high anxiety on the testee's performance. Assessing testee's anxiety during the testing process can be achieved, for example, by using a special device fixed on testee's hand which is connected to the computer used. Again, this untested idea needs more careful consideration.

12. Many have argued that CAT should do more than provide a score or adapt itself to candidates' abilities. CAT should be able to diagnose specific difficulties, for example, determine what candidate is doing wrong.

13. Research should surely address the as yet under-exploited potential computers have for presenting dynamic and high resolution graphics. Although there is some evidence which suggests an economic gain and greater accuracy using computerized tests, it does not appear that there is very much psychological gain to be obtained from simply producing a computerized version of an existing traditional test, since the psychological functions that make up these domains are not changed by the format of the test. The issue which could be more important is whether a computer is more efficient in assessing human ability by, for example, measuring different dimensions of an ability which would not be possible to measure by conventional means of assessment. Hence, this will provide us with the potential for expanding our conception and understanding of a specific ability by adopting a new testing format. For example, a clear and simple dynamic motion can be depicted to help understand the particular property of the motion involved. This limits the possibility of confounding understanding of the item with other examinee's skills such

as reading ability. That should improve the validity of the test (Wise & Plake, 1989). Although computers have a great capability for manipulating and presenting such dynamic displays and materials, attempts to use such capabilities are still limited. CAT, particularly when combined with novel test items, such as those involving movement, colour, speech, sound, and interactive graphics, could result in dramatic improvements in the diversity, efficiency and accuracy of psychological measurement.

14. Developers and users of CAT should be aware of Green et al.'s (1991) technical guidelines for developing and using CAT. Such issues are validity, effects of initial items, estimation of item parameters, dimensionality, measurement error, item pool characteristics, human factors, and CAT equivalence with its P&P counterparts.

15. Finally, one of the limitations of this research is the lack of real consequences for testees completing an experiment. No grade or important decision was affected by scores on the experiment. As a result, the testees may have completed the experiment without the necessary motivation, and even adequate anxiety, to do their best. Future research should try overcome this problem.

7.5 Concluding Comments

The aim of this thesis has been to develop realistic expectations about the psychological and psychometric implications of using CAT. CAT has numerous advantages and a potential for improving the efficiency and accuracy of testing, and has many potential areas of future contribution within personnel selection and assessment. This potential can be realised if proper consideration is given when designing, developing, and implementing these testing systems, and if professional standards are maintained by developers and users. Before the adoption of CAT, a number of significant psychometric, psychological, practical, and ethical considerations, mentioned throughout this thesis, must be addressed and taken into account. Finally, in the words of Schoonman (1989), " *Computerized testing is only dawning in practice. Although we expect computerized testing to gain in importance in the coming years, hurdles have to be taken. The psychometrical, technical, and economical barriers are overlooked too often. Computerized testing is not simply buying hardware and transferring paper based tests to it*".

REFERENCES

References

Agapitou, G.(1993).Evaluation of the equivalence of the automated mode of administration and the time limit effect on performance in selection tests. Unpublished dissertation. University of Hull.

Acker,W.(1980).In support of microcomputer based automated testing: A description of the Maudsley Automated Psychological Screening Test (MAPS). *British Journal on Alcohol and Alcoholism, 15,* 144-147.

Acker,W.(1982). A computerized approach to psychological screening: the Bexeley-Maudsley Automated Psychological Screening and the Bexeley-Maudsley Category Sorting Test. *International Journal of Man-Machine Studies, 17,*361-370.

Aiken,L.R.(1988). Psychological testing and assessment. (pp.435-447).Boston:Allyn and Bacon.

Alkhadher,O., Anderson,N. & Clarke D. (in press). Computer-Based Testing: A Review of Recent Developments in Research and Practice. *The European Work and Organizational Psychologist.*

Alkhadher,O., Clarke D. & Anderson,N.(1994). The Effects of Knowledge about Adaptive Tests on subjects' performance. Paper presented to The British Psychological Society-Occupational Psychology Conference, 3rd-5th January 1994, Birmingham.

Allison,D.E. & Thomas,D.C(1986). Item difficulty sequence in achievement examinations: Examinees' preferences and test taking strategies. *Psychological Report, 1986, 59*867-870.

Allred, L.J.(1986). Sources of non-equivalence between computerised and conventional psychological tests. Unpublished doctoral dissertation, John Hopkin University, Baltimore.

American Psychological Association.(1986).Guidelines for computer-based tests and interpretations. Washington DC:Author.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education.(1985). Standards for educational and psychological tests. Washington,D.C.:American Psychological Association.

American Psychological Association Committee on Professional Standards (COPS) and Committee on Psychological Tests and Assessment (PTA) (1986). Guidelines for computer based tests and interpretations. Washington, DC: Author.

Anastasi, A.(1988). Psychological testing. New York: Macmillan Publishing Co.

Anderson, N.R. and Shackleton, V.J. (1986).Recruitment and selection: A review of developments in the 1980s. Personal Review, 15(4),19-26.

Anderson, R.E.(1987). Females surpass males in computer problem-solving: Findings from the Minnesota Computer Literacy Assessment. *Journal of Educational Computing Research, 3(1)*, 39-51.

Arvey,R.D., Strickland,W., Drauden,G.,& Martin,C.(1990). Motivational components of test taking. *Personnel Psychology, 43*,695-716.

Awdah,A.S.(1988).The effect of multiple-choice item response changes on test scores and the relation of such changes to anxiety and item difficulty. *Dirasat, 15(1)*,68-80.

Baker,F.B.(1989). Computer technology in test construction and process. In R.Linn. *Educational Measurement*, Macmillan Publisher Co:NY,USA.

Barcikowski,R.S. & Olsen,H.(1975). Test item arrangement and adaptation level. *Journal of Psychology, 90*,87-93.

Bartram,D.(1994a). Computer-based Assessment. In C.L.Cooper & I.T.Robertson (Eds.),*International Review of Industrial and Organisational Psychology, 9*,31-69.

Bartram,D.(1994b). What is so important about reliability? The need to consider the standard error of measurement. *Selection & Development Review, 10(1)*, 1-3.

Bartram,D.(1994c). The role of computer-based test interpretation (CBTI) in occpational assessment. BPS Workshop in advances in selection and assessment, Nottingham University:UK.

Bartram,D. (1985, December).The automation of psychological testing procedures: Towards some guideline for management and operation. Paper presented to the conference on The Management and Operation of Computer-based Testing-procedures. London.

Bartram,D. (1989a). Computer-based assessment. In P.Herriot (Ed.),Handbook of assessment in organizations.(pp.369-390). London:Wiley.

Bartram,D. (1989b). Emerging trends in computer-assisted assessment. Paper presented for the Hohenheim Conference on Individual and Organizational Aspects of Selection. Stuttgart.

Bartram,D. & Bayliss,R.(1984). Automated testing: Past, present ,and future. Journal of Occupational Psychology, 57,221-23.

Bartram,D. & Dale,H.C.A. (1983). Micropat version 3.0: A description of the fully automated personnel selection testing system being developed for the Army Air Corps. Ergonomics Research Group, University of Hull, Report ERG/Y6536/83/7.

Bartram,D., Lindley,P.A., & Foster,J.(1992). *Review of psychological tests for assessment in vocational training.* British Psychological Society Books.

Beaumont,J.G.(1985a).The effect of microcomputer presentation and response medium on digit span performance. International Journal of Man-Machine Studies,22,11-18.

Beaumont,J.G.(1985b). Speed of response using keyboard and screen-based microcomputer response media. International Journal of Man-Machine Studies,23,61-70.

Beaumont,J.G. & French,C.C.(1987).A clinical field study of eight automated psychometric procedures: The Leicester/DHSS project. International Journal of Man-Machine Studies,26, 661-682.

Bejar, I.I.(1975). *An investigation of the dichotomous, graded, and continuous response level latent trait models.* Unpublished dissertation, University of Minnesota.

Bejar,I.I., Weiss,D., & Gialluca,K.(1977). An information comparison of conventional and adaptive tests in the measurement of classroom achievement (Research report 77-7). Minneapolis: University of Minnesota, Department of Psychology (No.N00014-76-C-0627).

Bergstrom,B.A. & Lunz,M.E. (April, 1991). confidence in pass/fail decision for computer adaptive and paper and pencil examinations. Paper presented at the annual meeting of the American Educational Research Association. Chicago, IL, USA.

Bersoff,D.N. & Hofer,D.J.(1991). Legal issues in computerized psychological testing. In Gutkin,T.B. & Wise,S.L. *The computer and the decision-making process.* Hillsdale, NJ:L.E.A.Erlbaum.

Betz,N.E.(1977).Effects of immediate knowledge of results and adaptive testing on ability test performance. Applied Psychological Measurement,1,259-266.

Betz,N.F., & Weiss,D.J.(1975). Empirical and simulation studies of flexilevel ability testing (Research Report 75-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program (NTIS No.A013185).

Bevan, N. (1981). Is there an optimum speed for presenting text on a VDU ? International Journal of Man-Machine Studies,14,59-76.

Biskin,B.H. & Kolotkin,R.L.(1977).Effects of computerized administration on Scores on the. Applied Psychological Measurement, 1(4),543-549.

Bloom,A.J.(1985). An anxiety management approach to computer phobia. *Training and Development Journal, Jan,* 90-94.

Bloom,B.L.(1992). Computer-assisted psychological intervention: A review and commentary. *clinical Psychology Review,12,*169-197.

Bloxom,B.(1989). Adaptive testing: a review of recent results. *Zeitschrift Fur Differentielle und Diagnostische Psychologie, 10(1),* 1-17.

Boudreau,J.W. (1983). Economic considerations in estimating the utility of human resource productivity improvement programs. *Personnel Psychology, 36,* 551-557.

Bradshaw,G.D. & Gaudry,E.(1968). The effect of a single experience of success or failure on test anxiety. *Australian Journal of Psychology,20,*219-223.

Braun,H.I., & Holland,P.W.(1982).Observed score test equating: A mathematical analysis of some ETS equating procedures. In P.W.Holland & D.B.Rubin (Eds.),*Test equating* (pp.9-49).New York: Academic Press.

Bresolin,M.J.,Jr (1984). A comparative study of computer administration of the Minnesota Multiplasic Personality Inventory in an inpatient psychiatric setting. Unpublished doctoral dissertation. Loyola University of Chicago.

Brierley,H. (1971). A fully automated intellectual test. British Journal of Social and Clinical Psychology, 10 ,286-288.

Brinton,G. & Rouleau,R.A.(1969).Automating the hidden and embedded figures test. Perception and Motor Skills, 29, 401-402.

Brosnan,M.J. & Davidson,M.J.(1994). Computerphobia: Is it a particularly female phenomenon. The Psychologist,7(2), 73-78.

Burke,M.J., & Normand,J.(1987).Computerized psychological testing: Overview and Critique. Professional Psychology: Research and Practice,18(1),42-51.

Bunderson,C.V., Inouye,D.K., & Olsen,J.B.(1989). The four generations of computerised educational measurement. In R.L.Linn (Ed.), *Educational Measurement* (3rd edition).New York: Macmillan,pp.367-408.

Burke,M.J.(1984).*Eastman Kodak computerised clerical test validation report.* Psych Systems Technical Report. Baltimore, MD: Psych Systems.

Burke,M.J.(1993).Computerised psychological testing. In Schmitt,N., Borman,W.C.(Eds.), *Personnel Selection* (pp.203-239). San Francisco: Jossey-Bass.

Burke,M.J., & Normand,J.(1987).Computerized psychological testing: Overview and Critique. *Professional Psychology: Research and Practice,18(1)*,42-51.

Burke, M.J.,Normand, J.,Raju, N.S. (1987). Examinee attitudes toward computer-administered ability testing. Computers in Human Behavior, 3,95-107.

Burkead, J.E. & Sampson, J.P.(1985).Computer-assisted assessment in support of the rehabilitation process. *Rehabilitation in Counselling Bulletin,28,*262-274.

Butcher, J.N. (1987). *Computerized Psychological Assessment, A practitioner's Guide.* New York: Basic books, INC.,Publishers.

Butcher,J.N.(1987).The use of computers in psychological assessment: An overview of practices and issues. In J.N. Bucher (Ed.), *Computerized psychological assessment, A practitioner's Guide.*(pp.26-49).New York:Basic Books,INC, Publisher.

Butcher,J.N., Keller,L.S. & Bacon,S.F. (1985). Current developments and future directions in computerized personality assessment. Journal of Consulting and Clinical Psychology,53(6),803-815.

Byers,A.P.(1981).Psychological evaluation by means of an on-line computer. Behaviour Research Methods and Instrumentation, 13(3), 585-587.

Byrnes,E. & Johnson,J.H.(1981).Change technology and the implementation of automation in mental health care setting. Behaviour Research Methods And Instrumentation, 13,573-580.

Calvert,E.J. & Waterfall,R.C.(1980).A system of automated psychological testing. Workshop on Intellectual Function Testing, Royal Hospital and Home for Incurable,London.

Calvert,E.J. & Waterfall,R.C. (1982). A comparison of conventional and automated administration of Raven's Standard Progressive Matrices. International Journal of Man-Machine Studies, 17,305-310.

Cook, M. (1988). *Personnel selection and productivity.* Chichester, Wiley.

Cornwell,J.M., White,E., & Rupinski,M.(1993). The effect of individual differences on preferences and performance involving Human-Computer interaction. Paper presented to the Computer in Psychology,93 conference.York,UK.

Coovert,M.D.(1986). Altering artificial intelligence heuristics to provide data for specific application. In S.Goel (chair), *Advances in tailoring job analysis methods for specific applications.* Symposium conducted at the annual conference of the Society for Industrial and Organizational Psychology, Chicago.

Crowne,D., & Marlowe,D.(1964). *The approval motive.* New York: Wiley.

Cronbach,L.J. & Gleser,G.(1965). *Psychological Testing and Personnel Decision.* Urbana,IL: University of Illinois Press.

Cudeck,R. (1985).A structural comparison of conventional and adaptive versions of the ASVAB. Multivariate behavioral research, 20(3),305-322.

Dann,P.L., Irvine,S.H., & Collis, J.M. (19). *Advances in computer-based human assessment.* London: Kluwer Academic Publisher.

Davis,C., & Cowles,M. (1989). Automated psychological testing: Methods of administration, need for approval, and measures of anxiety. *Educational and Psychological Measurement,49,* 311-321.

Denner,S. (1977). Automated psychological testing: A review. British Journal of Social And Clinical Psychology, 16, 175-179

Denny,J.P.(1966). Effects of anxiety and intelligence on concept formation. Journal of Experimental Psychology, 72, 596.

Denton, L.(1987). *Computers: Partners in practice.* APA Monitors, p 7.

Dimock,P. & Cormier,P.(1991).The effects of format differences and computer experience on performance and anxiety on a computer-administered test. *Measurement and evaluation in counselling and development,24,*119-126.

Divig,D.R.(1988). *Two consequences of improving a test battery* (CRM 88-171). Alexandria, VA: Center for Naval Analyses.

Drasgow,F. & Hulin,C.L.(1990). Item Response Theory. In M.D.Dunnette & L.M.Hough (eds.), Handbook of industrial and organizational psychology (2nd ed.).

Dunn,T.G., Lushene,R.E. & O'Neil,H.F. (1972).Complete automation of the MMPI and a study of its response in latencies. Journal of Consulting and Clinical Psychology.39, 381-387.

Dutro,K.(1983).The development of a MUMPS-based rehabilitation psychology computer applications. Paper presented at the Annual Convention of the American Psychological Association.

Elithorn,A.,Mornington,S. & Stavrou,A.(1982). Automated Psychological testing: Some Principles and practice. International Journal of Man-Machine Studies,17,247-263.

Elithorn,A.,Jones,D. & Kerr,M.O.(1963).A binary perceptual maze. American Journal of Psychology,76,506-508.

Elithorn,A. & Telford,A. (1969). Computer analysis of intellectual skills. International Journal of Man-Machine Studies, 1, 189,209.

Elwood,D.L.(1969). Automation of psychological testing. American Psychological,24,287-289.

Elwood,D.L.(1972a).Test retest reliability and cost analyses of automated and face to face intelligence testing. International Journal of Man-Machine Studies,4,1-22.

Elwood,D.L.(1972b).Automated WAIS testing correlated with face-to-face WAIS testing: A validity study. International Journal of Man-Machine Studies, 4(2),129-137.

Elwood,D.L. & Clark,C.(1978).Computer technology. Computer administration of the Peabody Picture Vocabulary Test to young children. Behavioral Research Methods and Instrumentation, 10,43-46.

Elwood,D.L. & Griffin,R.H.(1972). Individual intelligence testing without the examiner: Reliability of an automated method. Journal of Consulting and Clinical Psychology,38,9-14.

Erdman,H.P.Klein,M.H. & Greist,J.H.(1985).Direct patient computer interviewing. Journal of Consulting and Clinical Psychology, 53,760-773.

Evans,A. (1994). *The data protection act: A guide for personnel managers.*

Eyde,L.D. (1987). Computerized psychological testing: An introduction. *Applied Psychology: An International Review,36(3/4)*, 223-235.

Eyde,L.D., & Kowal,D.M. (1987). Computerized test interpretation services: Ethical and professional concerns regarding U.S. producers and users. *Applied Psychology: An International Review,36(3/4)*, 401-417.

Federico,P.A.(1991).Measuring recognition performance using computer-based and paper-based methods. *Behaviour Research Methods, Instruments & Computers,23,*341-347.

Fekken,G.C., & Holden,R.R.(1989). Psychometric evaluation of the Microcomputerised Personality Research *Form. Educational and Psychological Measurement,49,*875-882.

Fowler,R.D. (1985). Landmarks in Computer-assisted psychological assessment. Journal of Counselling and Clinical Psychology, 53(6),748-759.

Fowler,R.D. & Butcher,JN.(1987). International applications of computer-based testing and interpretation. *Applied Psychology: An International Review,36(3/4)*, 419-429.

Frese,M.(1987). Human-computer interaction in the office. *International Review of Industrial and Organizational Psychology*, 117-156.

Gaines,B.R.(1981).The technology of interaction-dialogue programming rules. International Journal of Man-Machine Studies, 14,133-150.

Gambrell,J.B. & Sandfield,R.E. (1977) Computers in the school: Too much too soon? High School Journal,69,327-331.

Gardner, J.M., Souza,A., Scabbia,A., Breuer,A. (1986). Microcare-promises and pitfalls in implementing microcomputer programs in human services agencies. *Computers in Human Behaviour, 2,* 147-156.

Garrison,W. and Baumgarten,B.(1986). An application of computer adaptive testing with communication handicapped examinees. Educational and psychological measurement,46,23-35.

Gedye,J.L. & Miller,E.(1969).The automation of psychological assessment. International Journal of Man-Machine Studies,1,237-262.

Gerow,J.R.(1980).Performance on achievement tests as a function of the order of item difficulty. *Teaching of Psychology,7(2),*93-94.

Ghiselli,E.E. (1973). Validity of aptitude tests in personnel selection. *Personnel Psychology, 26(4)*, 461-477.

Gitzinger,I.(1990)Acceptance of test presentation on a personal computer by clinical subjects. Psychotherapie psychosomatik medizinisch psychologie, 40(3-4),143-145.

Glass,C.R. & Knight,L.A.(1988). Cognitive factors in computer anxiety. *Cognitive Therapy and Research, 12(4)*, 351-366.

Greaud,V.A. & Green,B.F.(1986). Equivalence of conventional and computer presentation of speed tests. Applied Psychological Measurement,10(1),23-34.

Green,B.F. (1991). Guidelines for computer testing. In T.B Gutkin & S.L Wise (1991). *The computer and the decision-making process.* Hillsdale, NJ: L.E.A. Erlbaum.

Green,B.F., Bock,D.R., Humphreys,L.G., Linn,R.L., Reckase,M.D. (1984).*Technical guidelines for assessing computerized adaptive tests. Journal of educational measurement,21(4)*,347-360.

Green,C.L.(1982). The diagnostic accuracy and utility of MMPI and MCMI computer interpretive reports. *Journal of Personality Assessment, 46,* 359-365.

Greist,J.H., Klein,M.H. & Van Cura,L.J.(1973).A computer interview for psychiatric Patient target symptoms. Archives of General Psychiatry,29,247-253.

Gronlund,N.E.(1981). *Measurement and evaluation in teaching(4th ed.).*New York: Macmillan.

Gutkin,T.B. & Wise,S.L.(1991). *The computer and the decision-making process.* Hillsdale, NJ:L.E.A. Erlbaum.

Hakel,M.D.(1986).Personnel selection and placement. Annual Review of Psychology,37,351-80.

Hambleton,R.K.(1986). Effects of item order and anxiety on test performance and stress. Paper presented at the annual meeting of Division D, the American Educational Research Association, Chicago.

Hambleton R.K. & Traub,R.E.(1974). The effects of item order on test performance and stress. *Journal of Experimental Education, 43,*40-46.

Hambleton,R.K., & Swaminathan,H.(1985). *Item response* theory. Boston: Kluwer-Nijhoff Publishing.

Hambleton,R.K., Zaal,L. & Pieters,J.P(1991).Computerised adaptive testing: Theory, applications, and standards. In R.K.Hambleton, & J.N.Zaal (Eds), *Advances in Educational and Psychological Testing: Theory and Applications.* Boston: Academic Publishers.

Hansen,K.E.,Johnston,J.H. & Williams,T.A.(1977). Development of an on-line management information system for community, mental health centres. Behaviour Research Methods and instrumentation, 19,139-143.

Harold,R.S. & James,O.R.(1973). The effects of immediate knowledge of results and task definition on multiple-choice answering. The journal of Experimental Education,42(1),77-80.

Harrell,T.H.,Honaker,L.M., Hetu,M. & Oberwager,J.(1987). Computerized versus traditional administration of the Multidimensional Aptitude Battery-Verbal Scale: An examination of reliability and validity. *Computers in Human Behaviour, 3*,129-137.

Harrell,T.H. & Lombardo,T.A. (1984).Validation of an automated 16PF administration procedure. Journal of Personality Assessment, 48,638-642.

Harvey,A.L.(1987).Differences in response behaviour for high and low scores as a function of item presentation on a computer assisted test. Unpublished doctoral dissertation, University of Nebraska, Lincoln.

Hattori, T (1990).Measuring of verbal ability with computerized adaptive testing. *Japanese Journal of Educational Psychology*, 1990, 38(4), 445-454.

Hedl, J.J.Jr, O'Neil,H.F.Jr & Hansen,D.N.(1973). Affective reaction toward computer-based intelligence testing. Journal of Counselling and Clinical Psychology,40,217-222.

Hedlund,J.L., Evanson,R.C., Sletten,I.W., & Cho,W.D.(1980).The computer and the clinical prediction. In J.B.Sideowski, J.H.Johnson, & T.A.Williams(Eds.).Technology in Mental Health Care Delivery Systems(pp.201-235).Norwood,NJ:Ablex.

Hedlund,J.L. & Vieweg,B.W. (1988).Automation in Psychological Testing. *Psychiatric Annuals(4)*,217-227.

Henly,S.J., Klebe,K.J., McBride,J.R., & Cudeck,R.(1989).Adaptive and Conventional Versions of the DAT: The first Complete Test Battery Comparison. *Applied Psychological Measurement,*13 (4),363-371.

Herr,E. & Best,P. (1984). Computer technology and counselling: The role of the professional. Journal of Counselling and Development,631,192-195.

Hofer,P.J.(1985). Developing standards for computerized psychological testing. Computers in Human Behaviour, 1,301-315.

Hofer,P.J. & Green,B.F.(1985).The challenge of competence and creativity in computerized psychological testing. Journal of Counselling and Clinical Psychology,53(6),826-838.

Honaker,L.M., Harrell,T.H., Buffaloe,J.D.(1988). Equivalency of Microtest computer MMPI administration for standard and special scales. *Computer in Human*

Behaviour,4(4),323-337.

Hopkins,K.D. & Stanley,J.C.(1981). *Educational and psychological measurement and evaluation(6th ed.)*.Englewood Cliffs, NJ: Prentice-Hall.

Hsu,T.C., Shermis,M.D.(1989). The development and evaluation of a microcomputerized adaptive placement testing system for college mathematics. Journal of Educational computing Research, 5(4), 473-485.

Huba,G.J.(1988).Comparability of traditional and computer Western Personnel Test (WPT) versions. *Educational and Psychological Measurement,48*,957-959.

Huck,S.W. & Bowers,N.D.(1972). Item difficulty level and sequencing effects in multiple choice achievement tests. *Journal of Educational Measurement,9*, 105-111.

Hulin,C.L., Drasgow, F., & Parsons,C.K. (1983). Item response theory: Application to psychological measurement. Homewood, IL:Dow Jones-Irwin.

Hunt,E., Pellegrino,J.W., Frick,R.W., Farr,S.A. & Alderton,D.(1988). The ability to reason about movement in the visual field. Intelligence,12,77-100.

Hyland,M.E. (1987). Performance, motivation and anxiety: The construct of "efforts" from a control theory perspective. In S.H.Irvine & S.E.Newstead (Eds). Intelligence and cognition: Contemporary form of reference, Martinus Nijholl Publisher, Dordrecht, Netherlands.

Janz,T. (1989). Case study on utility: utility to the rescue, a case of staffing program decision support. In M.Smith. & I.Robertson (eds). *Advances in selection and assessment*. London:Wiley.

Jay,T.(1985).Defining and measuring computerphobia. In R.Eberts and C.Eberts(Eds.). Trends in Ergonomics/Human Factors, Volume II (pp.321326). Amsterdam: North-Holland Publisher.

Johnson,D.F. & Mihal,W.L.(1973).The performance of blacks and whites in computerized versus manual testing environments. American Psychologist,28,694-699.

Johnson,J.H. & Johnson,K.N. (1981). Psychological considerations related to the development of computerized testing stations. Behaviour Research Methods and Instrumentation, 13(4),421-424.

Johnson,M.F., Weiss,D.J., & Prestwood,J.S.(1981). *Effects of immediate feedback and pacing of item presentation on ability test performance and psychological reactions to testing*. Research Report 81-2. Minnepolis: University of Minnesota, Department of

Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Johnson,D.F.,& White,C.B. (1980). Effects of training on computerized test performance in the elderly. Journal of Applied Psychology,65,357-358.

Johnson,J.H. & Williams,T.A. (1980).Using on-line computer technology to improve service response and decision-making effectiveness in a mental health admitting system. In J.B.Sidowski, J.H.Johnson, & T.A.Williams,(Eds.), Technology in mental health care delivery systems. Norwood.NJ: Ablex.

Katz,L. & Dalby,S.(1981).Computer and manual administration of Eysenck Personality Inventory. Journal of Clinical Psychology, 37,586-588.

Kent,T. & Albanese,M. (1987).A comparison of the relative efficiency and validity of tailored tests and conventional quizzes. Evaluation and the health professions,10(1),67-79.

Kiely,G.L.,Zara,A.R. & Weiss,D.J.(1983,January).Alternate forms reliability and concurrent validity of adaptive and conventional tests with military recruits. Draft report submitted to Navy Personnel Research and Development Center, San Diego,CA.

Kingler,D.E.,Miller,D.,Johnson,J.H. & Williams,T.A. (1977). Process evaluation of an on-line computer-assisted unit for intake assessment. Behaviour Research Methods & Instrumentation, 9, 110-116.

Kintz,B.L., Delprato,D.J., Mettee,D.R., Persons,C.E. & Schappe,R.H. (1965). The experimenter effect. Psychological Bulletin, 63,223.

Klepac,R.K.(1984).Micro-computers in behaviour therapy: A sample of applications. The Behaviour Therapist,7,79-83.

Klingler,D.E.,Johnson,J.H. & Williams,T.A.(1976).Strategies in the evolution of an on-line computer-assisted unit for intake assessment of mental health patients. Behaviour Research Methods & Instrumentation, 8,95-100.

Klonoff,H. & Clark,C.V.(1975).Measuring staff attitudes toward computerization. Hospital and Community Psychiatry, 26,823-825.

Knights,R.M.,Richardson,D.H. & McNarry,L.R.(1973).Automated vs clinical administration of the Peabody Picture Vocabulary Test and Colored Progressive Matrices. American Journal of Mental Deficiency,78(2),223-225.

Koch,W.R., Dodd,B.G., & Fitzpatrick,S.J. (1990).Computerized adaptive measurements of attitudes. Measurement and evaluation in counselling and development,23(1), 20-30.

Kovac,R.(1989).The effects of computerized selection tests on job applicant performance. Unpublished Ph.D, DePaul University.

Kratochwill,T.R.(1985).Selection of target behaviours in behavioral consultation. Behavioral Assessment,7,49-61.

Kratochwill,T.R. & Doll,E.J.(1991) In Gutkin,T.B. & Wise,S.L. *The computer and the decision-making process.* Hillsdale,NJ:L.E.A. Erlbaum.

Krug,S.E.(1988).*Psychware Sourcebook: Third edition.* Kansas City,MO,Test Corporation of America.

Krus,D.J., & Ceurvost,R.W.(1978). Computer assisted construction of variable norms. Educational and Psychological Measurement, 38,815-818.

Kubinger,K.D., Formann,A.K. & Farkas,M.G.(1991).Psychometric shortcomings of Raven's Standard Progressive Matrices, in particular for computerized testing. European Review of applied psychology, 41(4),295-300.

Larkin,K.C. & Weiss,D.J.(1974). *An empirical investigation of computer-administered pyramidal ability testing.* Research Report 74-3. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Leary,L.F. & Dorans,N.J.(1985). Implications for altering the context in which test items appear: a historical perspective on an immediate concern. *Review of Education Research, 55*,387-413.

Lee,J.A.(1986).The effects of past computer experience on computerized aptitude test performance. Educational and Psychological Measurement,46,727-733.

Lee,J.A., Moreno,K.E., & Sympson,J.B.(1986). The effects of mode of test administration on test performance. *Educational and Psychological Measurement*,46,467-474.

Levy,A.L. & Barowsky,E.I.(1986).Comparison of computer-administered Harris-Goodenough Draw-a-Man test with standard paper-and-pencil administration. *Perceptual and Motor Skills,63*,395-398.

Llabre,M.M., Clements,N.E., Fitzhugh,K.B., Lancelotta,G., Mazzagatti,R.D. & Quinones,N. (1987). The effect of computer-Administered testing on test anxiety and performance. Journal of Educational Computing Research,3(4),429-432.

Lord,F.M.(1980). Applications of item response theory to practical testing problems. Hillsdale.NJ:Erlbaum.

Lucas,R.W.(1977).A study of patients' attitudes to computer interrogation. International Journal of Man-Machine Studies,9,69-86.

Lunz,M.E., Bergstrom,B.A., & Wright,B.D.(1992). The effect of review on student ability and test efficiency for computerised adaptive tests. Applied Psychological Measurement,16(1),33-40.

Lukin,M.E.,Dowd,E.T.,Plake,B.S., and Kraft,R.G.(1985). Comparing computerized versus traditional psychological assessment, Computers in human behaviour, 1,49-58.

Mandler,G., & Sarason,S.B.(1952). A study of anxiety and learning. *Journal of abnormal and social psychology,47*,166-173.

Martin,C.L. & Nagao,D.H.(1989).Some effects of computerised interviewing on applicant response. *Journal of Applied Psychology,74*,72-80.

Matarazzo,J.D.(1983).Editorial on computerized psychological testing. Science, 221 (4608), (22 July),323.

Matarazzo,J.D.(1990). Psychological Assessment Versus Psychological Testing. *American Psychologist,45(9)*,999-1017.

Mathisen,K.S.,Evans,F.J.,Meyers,K. & Kogan,L.(1985).Human factors influencing patient-computer interaction. Computers in Human Behaviour, 1,163-170.

Maurelli,V.A., & Weiss,D.J.(1991).Factors influencing the psychological characteristics of an adaptive testing strategy for test batteries (Research Rep. No.81-4).Minneapolis: University of Minnesota, Department of psychology, computerized adaptive testing laboratory.

Mazzeo,J. & Harvey,A.L.(1988).The equivalence of scores from automated and conventional educational and psychological tests:A review of the literature (College Board Report No.88-8,ETS RR No.88-21).Princeton,NJ: Educational Testing Service.

McBride,J.R.(1980). Adaptive verbal ability testing in a military setting. In D.J. Weiss (Ed.), *Proceedings of the 1979 computerized adaptive testing conference*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

McBride,J.R.(1986)*A computerized adaptive edition of Differential Adaptive Tests.* Paper presented at the meeting of the American Psychological Association, Washington DC.

McBride,J.R. & Martin,J.T.(1983).Reliability and validity of adaptive ability tests in a military setting. In D.J.Weiss(Ed.), New Horizons in testing:Latent trait test theory and computerized adaptive testing (pp.223-236).New York:Academic Press.

McCormic,E.J. & Ilgen,D.R.(1987). *Industerial and organizational psychology.* Englewood Cliffs, New Jersey:Allen and Unwin.

Mchenry,J.M., Hough,L.M., Toquam,J.L., Hanson,M.A., & Ashworth,S.(1990). Project a validity results: The relationship between predictor and criterion domains. *Personnel Psychology,43*,335-354.

Mead A. & Drasgou F. (1993). Equivalence of computerises and paper and pencil ability tests: A meta-analysis. Psychological Bulletin, 144 (in press).

Meier,S.T. & Geiger,S.M.(1986).Implications of computer-assisted testing and assessment for professional practice and training. Measurement and Evaluation in Counselling and Development, 19(1),29-34.

Meier,S.T.(1988).Predicting individual differences in performance on computer-administered tests and tasks: Development of the Computer Aversion Scale.*Computer in human behaviour*,4,175-187.

Meier,S.T. & Lambert,M.E.(1991).Psychometric properties and correlates of three computer aversion scales. Behaviour research methods, instrument & computers,23(1),9-15.

Millman,J.,Bishop,C.H & Ebel,R.(1965). An analysis of test-wiseness. *Educational and psychological measurement,25*,707-726.

Moe,K. and Johnson,M.(1988).Participant's reactions to computerized testing. Journal of educational computering research,4(1),79-86.

Moreland,K.L.(1987).Computerized psychological assessment: What's available. In J.N. Bucher (Ed.),*Computerized psychological assessment, A practitioner Guide.*(pp.26-49).New York:Basic Books, INC, Publisher.

Moreno,K.E., Wetzel,C.D., McBride,J.R., & Weiss,D. (1984).Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB). Applied psychological measurement,8(2),155-163.

Morris,L.W.,& Fulmer,R.S.(1976).Test anxiety (worry and emotionality) changes during academic testing as a function of feedback and test performance. Journal of Educational Psychology,68,817-824.

Moser,K.; Selig,J.; Johannes,G.; Rebstock,M.(1990). Process variables in a computer-assisted performance test. Diagnostica,36(4),321-328.

Munger, C.F. & Loyd,B.H.(1989). Gender and attitudes toward computers and calculators: The relationship to Math performance. *Journal of Educational Computing Research, 5(2)*, 167-177.

Murphy,K.R. & Davidshofer,D.O.(1994). Psychological Testing: Principles and applications. Englewood Cliffs,NJ:Prentice-Hall, Inc.,

Noonan, J.V. & Sarvela, P.D. (1991). Implementation decisions in designing computer-based instructional testing programs. In T.B Gutkin & S.L Wise (1991). *The computer and the decision-making process.* Hillsdale, NJ:L.E.A. Erlbaum.

Newstead,S. & Dennis,I.(1994). Examiners examined: The reliability of exam marking in psychology. *The Psychologist, 7(5)*, 216-219.

Nurius,P.S.(1990b).A review of automated assessment. *Computer in Human Services,6(4)*,265-281.

O'Brien,T.,Dugdale,V.(1978).Questionnaire administration by computer. *Journal of the Market research Society,20*,228-237.

Offerman,L.R. & Gowing,M.K.(1993). In N.Schmitt and W.Borman (Eds), Personnel Selection in Organizations. San Francisco: Joss-Bass Publication.

Oslon,C.L.(1976). On choosing a test statistic in multivariate analysis of variate. *Psychological Bulletin,83*,579-586.

Olsen,J. (1990).Applying computerized adaptive testing in schools. M*easurement and evaluation in counselling and development;23(1)*,31-38.

Olsen,J.B.,Maynes,D.D.,Slawson,D., & Ho,K. (1986,April). Comparison and equating of paper-administered, computer-administered, and computerized adaptive tests of achievement. Paper presented at the meeting of the American Educational Research Association,San Francisco.

Olsen,J.B., Maynes,D.D., Slawson,D. & Ho,K.(1989).Comparisons of paper-administered, computer-administered, and computerized adaptive achievement tests. *Journal of Educational Computing Research*, 5(3),311-326.

O'Neil,H.F., & Baker,E.L. (1991). Issues in intelligent computer-assisted instruction: Evaluation and measurement. In Gutkin,T.B. & Wise,S.L. *The computer and the* *decision-making process.* Hillsdale,NJ:L.E.A.Erlbaum.

Overton,W.G. & Scott,K.G.(1972).Automated and manual intelligence testing: Data on parallel forms of the Peabody Picture Vocabulary Test. American Journal of Mental Deficiency,76,639-643.

Owen,R.A.(1975).A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American statistical association,*70,351-356.

Park,K.S. & Lee,S.W.(1992). A computer-aided aptitude test for predicting flight performance of trainees. *Human Factors, 34(2)*, 189-204.

Pearson,J.S.,Swenson,H.P.,Rome,H.P.Mataya,P. & Brannick,T.L. (1965).Development of a computer system for scoring and interpretation of MMPI in a medical clinic. Annals of the New York Academy of Sciences,126,682-692.

Perriolate,R.(1987).L'informatisation des tests au service de psychologie de la S.N.C.F.Applied Psychology an International Review,36(3-4),299-310.

Petiprint, G.L. and Johnson,M,E. (1990). Effects of gender, attributional style, and item difficulty on academic performance. *The Journal of Psychology, 125(1)*,45-50.

Plake,B.S.(1981). Item arrangement and knowledge of arrangement on test scores. *The Journal of Experimental Education, 49,*56-58.

Plake,B.S., Ansorge,C.J., Parker,C.S. & Lowry,S.R.(1982). Effects of item arrangement, Knowledge of arrangement test anxiety and sex on test performance. *Journal of Educational Measurement, 19,*49-57.

Plake,B.S., Thompson,P.A., & Lowry,S.(1981). Effect of item arrangement, Knowledge of arrangement and test anxiety on two scoring methods. *The Journal of Experimental Education, 49,*214-219.

Pocius,K.E.(1991). Personality factors in human-computer interaction:A review of the literature. Computers in Human Behaviour,7,103-135.

Quible,Z. & Hammer,J.N.(1984). Office automation's impact on personnel. *Personnel Administration, 6,* 25-32.

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

Reckase,M.D. (1986). The use of tailored testing with instructional programs. Final report. (Research report ONR-86-1).Iowa: American College, Testing Program (N00014-

82-K-0716).

Ridgway,J., MacCulloch,M.J. & Mills,H.E.(1982).Some experiences in administering a psychometric test with a light pen and microcomputer. International Journal of Man-Machine Studies, 17,265-278.

Robertson,I.T. & Smith,M. (1989). Personnel selection methods. In M Smith & I. Robertson. *Advances in selection and assessment.* London:Wiley.

Rock,D.L. & Nolen,P.A.(1982).Comparison of the standard and computerized version of the Raven Colored Progressive Matrices Test. Perceptual and Motor Skills,54,40-42.

Rocklin,T. & Thompson,J.M.(1985).Interactive effects of test anxiety, test difficulty, and feedback. Journal of Educational Psychology,77,368-372.

Rocklin,T. & O'Donnell,A. (1987).Self-adapted testing: A performance improving variant of computerized adaptive testing. Journal of educational psychology,79(3).315-319.

Rolls,S. & Harris,J.(1993).The effects of computer based test interpretation (CBTI) in staff selection-implications for usage and training. Paper presented at the BPS Occupational Psychology Conference, Brighton.

Romanczyk,R.C.(1986). *Clinical utilization of microcomputer technology.* New York: Pergamon.

Roper,B.L., Ben,P., Yossef,S., & Butcher,J.N.(1991). Comparability of computerized adaptive and conventional testing with the MMPI-2. *Journal of Personality Assessment, 57(2)*, 278-290.

Rosen,L., Sears,D. & Weil,M. (1987). Computerphobia. *Behaviour Research Methods, Instruments, and Computers,19*,167-179.

Sacher,J. & Fletcher,J.D.(1978). Administering paper-and-pencil tests by computer ,or the medium is not always the message. In D.J.Weiss (Ed.),Proceedings of the 1977 Computerized Adaptive Testing Conference (pp.403-420). Minneapolis: Department of Psychology,University of Minnesota.

Sackett,P.R.(1989). Comment on selection utility analysis. In M.Smith & I.Robertson. *Advances in selection and assessment.* London:Wiley.

Sampson,Jr.J.P. (1983). Computer-assisted testing and assessment: Current status and implications for the future. *Measurement and Evaluation in Guidance,15(3)*,293-299.

Sampson,Jr.,J.P. & Pyle,R.K.(1983). Ethical issues involved with the use of computer-assisted counseling, testing, and guidance systems. The Personnel and Guidance Journal,61(5),283-287.

Sampson,Jr.,J.P. & Weiss, D.J..(1981). *Predictive validity of conventional and adaptive tests in an air force training environment.* Manpower and Personnel Division, Minneapolis: Department of Psychology, University of Minnesota, Minnesota.

Sanchez,M.(1991). *Evaluation of the automated form of the MGIB tests.* Unpublished Ms.c,Hull University.

Sands,W., & Gade,P.A.(1991). An application of computerized adaptive testing in U.S. army recruiting. *Journal of Computer-based Instruction, 10(3,4)*, 87-89.

Sarason,I.(ed.) Test anxiety: Theory, Research and Application, New Jersey: Lawrence Erlbaum Associates, Publishers,p.9,1980.

Sax,G. & Cromack,T.R.(1966). The effects of various forms of item arrangement on test performance. *Journal of Educational Measurement,1966,3,*309-311.

Sattler,J.N. & Theye,F.(1967).Procedural, situational, and interpersonal variables in individualized intelligence testing. *Psychological Bulletin,68,*347-360.

Schmidt,F.L.(1993). Personnel psychology at the cutting edge. In N.Schmitt and W.Borman (Eds), Personnel Selection in Organizations. San Francisco:Jossy-Bass Publication.

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

Schmidt,F.L., Urry,V.W., Gugel,J.F. (1978). Computer assisted tailored testing: examinee reactions and evaluations. *Educational and Psychological Measurement, 38,* 265-273.

Schmitt,N.,Gilliland,S.W., Landis,R.S., & Devine,D.(1993). Computer-based testing applied to selection of secretarial applicants. *Personnel Psychology,* 49,149-165.

Schoenfeld,L.F. (1989). Guidelines for computer-based psychological tests and interpretations. Computer in Human Behaviour, 5(1),13-21.

Schoenfeldt,L.F. & Mendoza,J.L.(1991). The use of the computer in the practice of industrial/Organisational psychology. In Gutkin,T.B. & Wise,S.L. *The computer and the decision-making process.* Hillsdale,NJ: L.E.A.Erlbaum.

Schuldberg,D.(1990).Varieties of inconsistency across test occasions: Effects of computerised test administration and repeated testing. *Journal of Personality Assessment,55*,168-182.

Schoonman,W.(1989). *An applied study on computerised adaptive testing.* Offsetdrukkerij Kanters B.V., Alblasserdam: Netherlands.

Silver,E.M., Bennett,C. (1987). Modification of the Minnesota Clerical Test to predict performance on video display terminals. *Journal of Applied Psychology,72*,153-155.

Skinner,H.A. & Allen,B.A.(1983).Does the computer make a difference? Computerized versus face-to-face versus self-report assessment of alcohol, drug, and tobacco use. Journal of Consulting and Clinical Psychology,51,267-275.

Skinner,H. & Pakula,A.(1986).Challenge of computers in psychological assessment. Professional Psychology: Research and Practice,17(1),44-50.

Slack,W.V. & Slack,C.W.(1977).Talking to a computer about emotional problems: A comparative study. Psychotherapy: Theory, Research, and Practice,14,156-164.

Smith,M. & Robertson,I.(1989). *Advances in selection and assessment.* London:Wiley.

Smith,R.(1963). Examination by computer. Behavioral Science, 8,76-79.

Smouse,A.D. & Munz,D.C.(1986). The effects of anxiety and item difficulty sequence on achievement testing scores. *Journal of Psychology,68,* 181-184.

Space,L.G.(1981).The computer as psychometrician. Behaviour Research Methods & Instrumentation, 13(3),595-606.

Spielberger,C.D., Gorsuch,R.L., Lushene,R.E., Vagg,P.R., & Jacobs,G.A. (1983). *STAI manual.* Palo Alto,CA: Consulting Psychologist Press.

Spintti,J.P., & Hambleton,E.K.(1977).A computer simulation study of tailored testing strategies for objective-based instructional programs. Educational and Psychological Measurement, 37,139-158.

Stoloff,M & Couch,J.(1992).*Computer use in psychology.* American Psychological Association,Inc, Washington,DC,USA.

Stout,R.L.(1981). New approaches to the design of computerized interviewing and testing systems. Behaviour Research Methods & Instrumentation,13,436-442.

Strang,H.R.(1980).Effect of technically worded options on multiple-choice test performance. *Journal of educational research, 73*,262-265.

Strange,H.R. & Rust,J.O.(1973).The effects of immediate knowledge of results and task definition on multiple-choice answering. Journal of Experimental Education,42(1),77-80.

Sweeney,D.L., Smouse,A.D., Rupiper,O., & Munz,D.C.(1970). A test of the inverted U hypothesis relating achievement anxiety and academic test performance. *The Journal of Psychology, 74,*267-273.

Swenson,W.M.(1960). A preliminary investing of the possibilities of application of computer devices to the scoring and interpretation of structured personality tests and their use in a medical center. International Business Machines Corporation Medical Symposium Proceedings,2,401-415.

Tylor,T.R.(1983).Computerized testing. South Africa Journal of Psychology,13(1),23-31.

Telfer,R.(1985).Micro-computer based psychological testing and record-keeping. Defense Force Journal, 54, September/October.57-61.

Temple,D.E. & Geisinger,K.F.(1990). Response latency to computer-administered inventory items as an indicator of emotional arousal. *Journal of Personality Assessment*,54(1&2),289-297.

The data protection registrar,(1989).*Data protection Act 1984.*

The Psychological Corporation (1988). DAT Computerized Adaptive Edition: User's Manual,IBM PC Version.

The Psychological Corporation (1986). Differential Aptitude Tests Form V: British Manual.

Thompson,J.G. & Weiss,D.J.(1980). *Criterion-related validity of adaptive testing strategies*. Research Report 80-3. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Thompson,J.A. & Wilson,A.L. (1982). Automated Psychological Testing. International Journal of Man-Machine Studies,17,279-289.

Trapp,A. & Hammond,N(1991).*The CTI directory of psychology software*. CTI Centre for Psychology, University of York, York,UK.

Traver,D.(1986).Vocational evaluation upgrade system. Training and technology for the disabled, Discovery III conference paper. University of Wisconsin-Stout,Menomonie.

Stout Vocational Rehabilitation Institute.

Tuck,J.P.(1978). Examinee's control of item difficulty sequence. *Psychological Report,42,* 1109-1110.

Vale,C.D.(1981).Design and implementation of microcomputer-based adaptive testing system. Behaviour Research Methods and Instrumentation, 13,399-406.

Vale,C.D. & Weiss,D.J. (1975).A study of computer administered stradaptive ability testing (Research Report 75-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program,(AD AO18758).

Vroom, V.H.(1964). Work and Motivation. New York: Wily.

Wagman,M.(1983).A factor analytic study of the psychological implications of the computer for the individual and society. Behaviour Research Methods & Instrumentation,15,413-419.

Wainer, H. (1990). Computerized adaptive testing: A primer, Hillsdale, NJ: Erlbaum.

Wainer, H., Dorans,N.J., Green,B.F., Mislevy,R.J., Steinberg,L., and Thissen,D. (1990). Future challenges. In Wainer H. (Eds), Computerized adaptive testing: A primer, Hillsdale, NJ:Erlbaum.

Wainer,H., & Kiely,G.L.(1987). Item clusters and computerised adaptive testing: A case for testlets. *Journal of Educational Measurement 24,* 185-201.

Walker,N.W. & Myrick,C.C.(1985).Ethical considerations in the use of computers in psychological testing and assessment. Journal of School Psychology,23(1),51-57.

Ward,W.C. (1984). Using microcomputers to administer tests. Educational Measurement: Issues and Practices,3(2),16-20.

Waters,B.K. (1977).An empirical investigation of the stratified adaptive computerized testing model. Applied psychological measurement,1(1), 141-152.

Wechsler,D. (1974). *Manual for the Wechsler Intelligence Scale for Childern-Revised.* San Antonio, TX:Psychological Corporation.

Weinman,J.A.(1982).Detailed computer analysis of performance on a single psychological test. International Journal of Man-Machine Studies,17,321-330.

Weiss,D.J.(1973).In R.C.Sapinkopf(1978),A computer adaptive testing approach to the measurement of personality variability. Dissertation abstract International,38,10B,4993.

Weiss,D.J.(1985).Adaptive testing by computer. Journal of Consulting and Clinical Psychology,53(6),774-789.

Weiss,D.J. & Betz,N.E.(1973). Ability measurement: Conventional or adaptive? Research report 73-1.Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, MN, 1973. (AD 757788).

Weiss,D.J., & McBride,J.R.(1984).Bias and information of Bayesian adaptive testing. *Applied psychological Measurement, 8,*273-285.

Weiss,D.J. & Vale,D.C. (1987). Adaptive testing. Applied Psychology: An International Review,36(3/4),249-262.

Weiss,D.J. & Yoes,M.(1991). Item response theory. In R.K.Hambleton, & J.N.Zaal (Eds), Advances in Educational and Psychological Testig: Theory and Applications. Boston: Academic Publishers.

Weiss,D.J. & Yoes,M.(1991). Item response theory. In R.K.Hambleton, & J.N.Zaal (Eds), Advances in Educational and Psychological Testing: Theory and Applications. Boston:Academic Publishers.

Whitely,S.E.(1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45,*479-494.

Wilson,S.L., Thompson,J.A.& Wylie,G. (1982). Automated psychological testing for the severely physical handicapped. International Journal of Man-Machine Studies,17,291-296.

Wise,S.L. & Plake,B.S.(1989).Research on the effects of administering tests via computers. Educational Measurement: Issues and Practice,8(3),5-10.

Wise,S.L. & Plake,B.S.(1990).Computerized-based testing in higher education. *Measurement and evaluation in counselling and development,*23,3-10.

Wise,S.L., Plake,B.S., Eastman,L.E., Boettcher,L.L., & Ukin,M.E., (1986).The effects of item feedback and examinee control on test performance and anxiety in a computer-administered test. Computers in Human Behaviour, 2,21-29.

Wise,S.L., Plake,B.S., Johnson,P., & Roos,L.(1992). A comparison of self-adapted and computerised adaptive tests. *Journal of Educational Measurement, 29(4),*329-339.

Wise,S.L. & Wise,L.A.(1987). Comparison of computer-administered and paper-administered achievement tests with elementary school children. *Computers in Human Behaviour, 3,* 15-20.

Witt,J.C. & Elliott,S.N. (1985). Acceptability of classroom management strategies. In T.R. Kratochwill (Ed). *Advances in school psychology* (Vol.4,pp.251-288). Hillsdale,NJ.:Lawrence Erlbaum Associates.

Zachary,R.A., & Pope,K.S.(1984).Legal and ethical issues in the clinical use of computerized testing. In M.D.Schwartz(Ed.),Using computers in clinical practice (pp.151-164).New York:Haworth Press.

APPENDICES

Appendix A

Computer Attitude Questionnaire

You have now completed a computerised test. It would help us if you could tell us what is your opinion of the whole procedure and your thoughts about the test. For some questions you are asked to circle one number from 1 to 7, for others please circle the statements that match your judgment.

1. Taking a computer-administered test made me feel more anxious than taking a paper-and-pencil test.
Strongly Agree 1 2 3 4 5 6 7 Strongly Disagree

2. The test instructions were difficult to understand.
Strongly Agree 1 2 3 4 5 6 7 Strongly Disagree

3. Reading questions from a computer screen is more difficult than reading them from a paper-and-pencil form.
Strongly Agree 1 2 3 4 5 6 7 Strongly Disagree

4. I had not enough time to give my answers.
Strongly Agree 1 2 3 4 5 6 7 Strongly Disagree

5. It was not easy to me to make any corrections in my answers.
Strongly Agree 1 2 3 4 5 6 7 Strongly Disagree

6. Interacting with the computer is a difficult task.
Strongly Agree 1 2 3 4 5 6 7 Strongly Disagree

7. How do you find the idea of administering tests on a computer?
(Please circle as many as apply)
a. Fun.
b. Boring.
c. Can't decide.
d. Threatening.
e. ....
f. ....

8. The things I particularly liked about the computerised test are ....

(please circle as many as apply)

a. Required less time.

b. Clarity and simplicity of methods.

c. Lack of time pressure.

d. Potential for quick feedback.

e. Less fatiguing and less effort.

f. ....

g. ....

9. The things I particularly disliked about the computerised test are ....

(please circle as many as apply)

a. I could not go back and review answers.

b. Problem in reading the screen.

c. Difficulty of adjusting to the method.

d. I found it tiring.

e. My eyes got tired.

f. ....

g. ....

10. Have you taken a test on a computer before?

a. Yes.

b. No.

c. I can't remember.

11. How often have you used a computer before takeing this test?

Never 1 2 3 4 5 6 7 Frequently

12. Describe how you have used computers. Check as many as apply. If you have never used a computer mark the first one only.

a. I have never used a computer.

b. I have played games on the computer.

c. I have completed one or more computer class or short course.

d. I have written computer programs.

e. I have used a computer at home.

f. I have used a computer at my university, school, or work.

g. ....

h. ....

13. For job selection purposes, computerised tests has no advantages over paper-and-pencil test.

Strongly Agree 1 2 3 4 5 6 7 Strongly Disagree

14. I prefer taking paper-and-pencil test to computer-administered test.
Strongly Agree 1 2 3 4 5 6 7 Strongly Disagree

Now could you please tell us what you think about the difficulty level and the sequence of the questions.

15. How difficult did you find the questions?
Very Easy 1 2 3 4 5 6 7 Very Difficult

16. I think the questions have been arranged from .....

a. Easy to hard.
b. Hard to easy.
c. Random arrangement.

17. I have noticed that every time I think I answered a question right the next question becomes ....

a. More difficult.
b. Easier.
c. Same difficulty level.
d. I don't know.

18. I have noticed that every time I think I answered a question wrong the next question becomes ....

a. More difficult.
b. Easier.
c. Same difficulty level.
d. I don't know.

Thank you for your time and help.

Appendix B

Please fill out the following section about yourself.

Last Name: ....

First Name: ....

Age: ....

Sex: ....

First Language: ....

Main Department: ....

Are you ...... Undergraduate ...... Postgraduate

Year of Study ....

Address: ....
....
....

Telephone: ....

Don't write any thing here.
Group: ....
Subject No. : ....
Disk No.: ....

KCAT: ....
CAT: ....
CATT1: ....
CAT2: ....
CATT2: ....
ANX.: ....
AQ: ....

Appendix C

Computer Attitude Questionnaire

You have now completed a computerised test. It would help us if you could tell us what is your opinion and thought about the test. For some questions you are asked to circle one number from 1 to 7, for others please circle the statements that match your judgment.

1. How difficult did you find the questions?
Very Easy 1 2 3 4 5 6 7 Very Difficult

2. I think the questions have been arranged from .....
a. Easy to hard.
b. Hard to easy.
c. Random arrangement.

3. I have noticed that every time I think I answered a question right the next question becomes .....
a. More difficult.
b. Easier.
c. Same difficulty level.
d. I haven't noticed anything in this regard.

4. I have noticed that every time I think I answered a question wrong the next question becomes .....
a. More difficult.
b. Easier.
c. Same difficulty level.
d. I haven't noticed anything in this regard.

5. How often have you taken an ability test before (e.g. Verbal, Numerical, Mechanical Reasoning tests). Not an achievement exam like you take usually at school?
Never 1 2 3 4 5 6 7 Frequently

6. Do you think knowing it is an adaptive tests affects you?
a. Yes
b. No.

If yes..

7. How do you think it affected you?

a. Made me much quicker 1 2 3 4 5 6 7 Made me much slower

b. Made me do much worse 1 2 3 4 5 6 7 Enable me to do much better

c. Worried me a lot, Strongly Agree 1 2 3 4 5 6 7 Strongly Disagree

Thank you for your time and help.

Appendix D

The purpose of producing CBT and CAT is to provide a better kind of test than the P&P format by, for example, speeding up the process of testing, reducing the number of items, cuting the cost of testing, producing new types of items ...etc. It is not the intention of these new formats to make testees think or sit differently, or to be more difficult to read and answer. Usually, CBTs and CATs are designed to be equivalent to an existing P&P test. However, inevitably, given the difference in nature between a computer and other written media, full equivalence is not always achieved. Testees also differ with regard to their interactional style and preferences toward the testing media, though systems are designed for the average user and not to fit with the preferences of each individual. Some of the factors responsible for these diferences have been dealt with throughout the thesis. The following is a list of factors which reflects the difference in nature between the two ways of testing:

Physical conditions:	Seating	
	Table height	
	Lighting	
	Ventilation	
	Temperature	
	Noise level	
Sensory:	Contrast	Size and color of screen
	Screen lighting	Multimedia (e.g. sound, moving objects
	Glare	Character attributes
	Reflection	
Motor:	Hand, head, and eye movements.	
Cognitive:	Number of items visible	

Visibility of graphs
Computer ability
Response device (pencil / keyboard / mouse ...etc.)
Distance between stimulus field and response device
Attributions to the computer including covert measurement.

Test related factors: Testing time limit
Time allowed for each item
Delay between items
Speed of response
Knowledge of where about the testee is in the test
Speed of the machine
Hardware and software used
Possibility of correcting previous answers
Paging back
Instructions
Providing item-by-item feedback
Difficulty level of initial items including the psychological benefits of a warm up stage
Type of test (power / speeded).

Testee related factors: Experience with computers
Attitude toward computers
Computer anxiety
Motivation
Preferences regarding HCI features
Testee's expectations about the nature of the test.

With the two experiments conducted to assess the equivalence between the P&P and CAT formats (Experiment 1 & 2), subjects were randomly assigned to one of two experimental groups, which differed in the order of presentation of the forms. The first group (P&PFG) received the P&P format first (Figure A), then the computerized format (Figure B), the second group (CATFG) received them in the reverse order. The typical delay between the first and second session was eight to eleven days. Three of the ability tests of DAT (Numerical Ability, Abstract Reasoning, and Mechanical Reasoning) in both format were used in the experiments. The system has the ability to score and record the results, and to print an individual profile report. In addition, the user can interrupt a testing session at any time and resume testing later. On the basis of the subject's response and item parameter values, the ability estimate is updated using the Bayesian technique. Given this updated ability estimate, the most informative of the remaining items is selected, after which a new ability estimate is calculated. The process continues until twenty items have been administered. Equipercentile equating is used to convert the ability estimate into raw scores, and these raw scores are used in this study for analysis purposes.

Both formats were administered according to the instructions in their manuals. The experiments was conducted in both sessions in individual cubicles where the physical conditions are kept as constant as possible. However, because of the heat and noise generated from the computer used, some change in the testing room temperature and level of noise were noticed.

As the author noticed from the testees' response on the attitude questionnaire (Experiment 2), a few experienced difficulty in reading from the computer, and some physical problems during the test session such as eye fatigue and headache. Computerized testing looks, from the testees' point of view, more interesting, novel, and challenging. The novelty of the medium used for assessment may increase the testee's motivation to answer the test items. Moreover, although instructions are more standardised on the computerized method, it is a difficult task to keep instructions of both formats equal. Answering items using P&P differs substantially from interacting with a computer screen and an input device. During the introductory stage of the P&P and CAT tests, the tester tries to establish a friendly climate with his/her testee, and pay individual attention to his/her inquiries.

Subjects were advised to call the examiner if help was needed. All subjects received general test instructions followed by directions for recording their answers on the answer sheet or the computer screen. Answer revisions were not permitted in the CAT format. For P&P administration, answer sheet, question booklet, scrap paper, rubber and two pencils were provided, whereas scrap paper, two pencils and an eraser were provided for the CAT version. To answer an item in the P&P format, testees needed to find and 'black' the right box on the answer sheet using a pencil, as opposed to pressing the appropriate keys on the keyboard. The differences in responding style may impose different hand, head, or eye movements upon the testee, or differeces in general pasture.

The CAT tests were administered using a Notebook IBM compatible PC with LCD monochrome screen, or standard PC, and a standard QWERTY keyboard. After the second session, subjects were thanked for taking part in the experiment, debriefed on the aims of the study, and given feedback about their results and given £5 for their co-operation.

Because of copyright on the tests used, the author is unable to show samples of the materials used through out this thesis.

Answer

1. Multiply
521
41   A 23761
----   B 65412
C 21361
D 76230
E none of these

2. ? = 30 % of 54
A 15
B 12
C 6.8
D 16.2
E none of these

Answer

8. (12 / 5) x 3 =
A 7.2
B 9.5
C 12
D 2.23
E none of these

9. Add
7345 + 4231
A 11576
B 36732
C 23456
D 13425
E none of these

.
.
.
.
.
.
.
.

Figure A

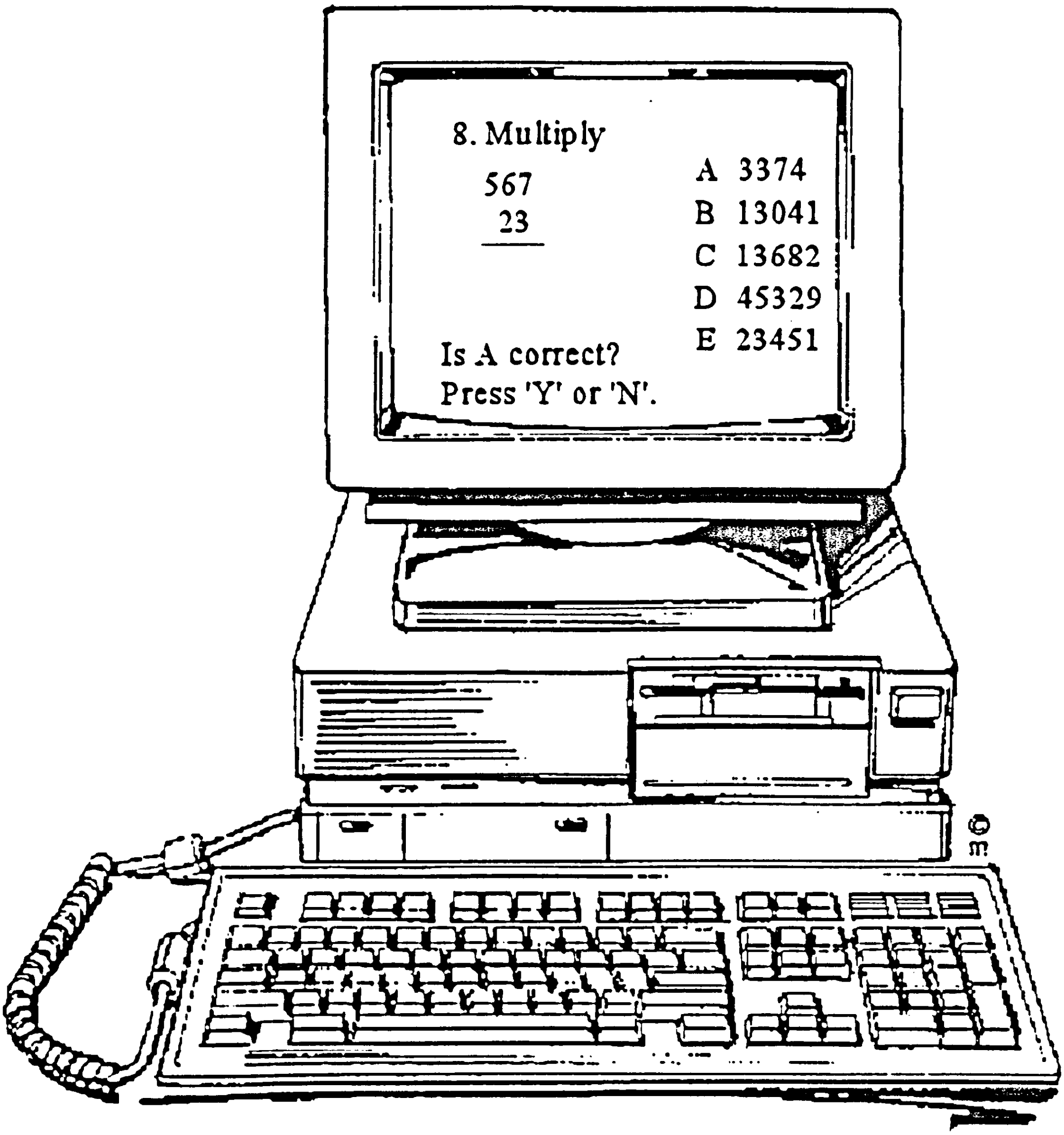
8. Multiply
567
23
A 3374
B 13041
C 13682
D 45329
E 23451
Is A correct?
Press 'Y' or 'N'.

Figure B.