Computerised ACER Advanced Test BL: Analysis of Equivalency, Test Anxiety, and the Effects of Input Device Using New Zealand University Participants

A thesis presented in partial fulfilment
of the requirements for the degree
of Master of Arts in Psychology
at Massey University

Michelle A. Gosse

1996

This thesis is dedicated to my father, Brian Robert Gosse.

O Trinity of love and power,
Our brethren shield in danger's hour;
From rock and tempest, fire and foe,
Protect them wheresoe'er they go:
Thus evermore shall rise to thee
Glad hymns of praise from land and sea.

Abstract

Study 1 examined the effects of the computerised format of the ACER Advanced Test BL (ACER-BL) on the test scores and anxiety of undergraduate participants, compared with the traditional paper-and-pencil format. Forty-one students were assigned to either a computer or paper-and-pencil treatment group using a stratified random design. Participants sequentially completed a general background questionnaire, the ACER-BL, an anxiety questionnaire, the ACER-BL, and a final anxiety questionnaire, with a 10 minute test-retest period between the ACER-BL administrations. There were no significant differences in ACER-BL score, and subsection scores, between the 2 treatment groups on either administration. The internal consistency reliability of each formats was moderate to high, and there was a high test-retest reliability for each format. While the mean scores for each treatment group were higher for the second test administration compared with the first, this result only reached significance for the computerised group. Gender, Undergraduate Year, and Typing Ability significantly influenced test score, although these failed to remain significant when treatment group was included in each analysis. These results suggest that the computerised version of the ACER-BL is equivalent to the paper-and-pencil version. Generally, there was no significant difference in reported test anxiety measures between the treatment groups, with mean reported anxiety indicating "slight anxiety." These anxiety results suggest little influence of test format on test anxiety.

Study 2 examined the influence of input device (keyboard, numeric pad, and mouse) on ACER-BL scores and test anxiety of undergraduate participants. Using stratified random assignment, 90 subjects were tested on all three input devices using a one factor repeated measures design. Each participant sequentially completed a general background questionnaire, the ACER-BL, an anxiety questionnaire, the ACER-BL, an anxiety questionnaire, the ACER-BL, and a final anxiety questionnaire, with a 10 minute delay between each ACER-BL administration. There was no significant main effect of input device on test score, and there was no significant order effect for input device. Between-subjects analyses indicated a significant increase in mean test score across administrations for the keyboard and numeric pad, but no significant change in mean scores with the mouse. These results were also reflected in the analyses of mean input response time. While there

was no significant effect of any measured participant characteristic on input device scores, mathematical ability and undergraduate year each had a significant influence on mean scores in the first ACER-BL administration. Participants with higher mathematical ability or more years at university had significantly higher mean test scores than participants with less mathematical ability or first year undergraduates respectively. While mean reported anxiety on all test anxiety measures decreased over the ACER-BL administrations, all mean reported anxiety indicated "slight anxiety." These anxiety results suggest little influence of input device on test anxiety.

The lack of test-retest comparisons between the computerised and paper-and-pencil formats of a test was discussed along with the need for future computerised testing research to use participants from the general population.

Acknowledgments

I would like to thank my thesis supervisor, Dr. Ross St. George, for his detailed comments on each stage of this thesis, and for proofreading earlier drafts of this thesis.

Special thanks go to two people. First, to Cedric Croft of the New Zealand Council of Educational Research for enabling computerised versions of the ACER Advanced Test BL to be written. Second, to Dr. Steve Humphries for the many hours spent debugging the programs, and for statistical help.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction to Automated Psychological Testing

Historical Background

Many United States universities created computer centres in the 1950s, and the first practical application for psychometricians was computer analysis and scoring of research data bases (Fowler, 1985). The use of computer technology in applied psychology settings began in the 1960s, when the Minnesota Mayo Clinic operated the first computer-assisted psychological test, in the form of the Minnesota Multiphasic Personality Inventory (MMPI). MMPI items administered on IBM cards, that were marked by a patient, were read by a scanner connected to a computer (Fowler, 1985). The Sixteen Personality Factor Questionnaire (16 PF) and the Rorshach were other personality tests initially computerised in the 1960s. However, the initial computerisation of these last two tests was limited to scoring and interpretation of test results, also called computer-based test interpretation (CBTI). Europe acquired CBTI after the United States, as Europe acquired computer technology after the US, and because Europe had relatively lower numbers of clinical psychologists and testing itself was less popular. The first European use of CBTI occurred in Switzerland, where the Fowler MMPI system was translated into a number of European languages.

As noted above, the computerisation of personality tests was originally limited to test scoring and interpretation, (Bartram & Bayliss, 1984), and early CBTI systems required the use of testing centres with large mainframes to score and analyse answer sheets (Fowler, 1985). Thus, the initial automation of psychological tests resulted in only partially computerised tests (Fowler, 1985). However, from the onset of CBTI, researchers were interested in computer-based test administration and scoring, in other words computerised psychological testing (CPT).

Initially automated test administration utilised machines other than computers, although attempts were made to standardise test administration. Clinical psychology applications of this 1960s computer technology were limited to purpose-built automated devices (Bartram & Bayliss, 1984). Lang (1969) describes an automated systematic desensitisation procedure where the computer application was limited to driving the

audiotape devices for stimulus presentation. This type of computerised assessment continued into the 1970s. For example, Kleinmuntz and McLean (1968) describe the use of a computer to adaptively administer the MMPI to clinical clients. The program initially presented clients with 5 items from each of the 15 MMPI subscales, then continued to administer items on subscales where the T-scores of the client did not fall into the normal range for those subscales. The feasibility of this program was tested using participants who completed the entire paper-and-pencil MMPI item set and had their answer sheets read into the computer. At the time of publication, Kleinmuntz and McLean (1968) had not tested their proposed automated procedure for MMPI administration. Elwood and Griffin (1972) describe an automated Wechsler Adult Intelligence Scale (WAIS) where a punched-paper tape reader ran tape decks that presented questions verbally and recorded participants' verbal answers and a Teletype machine that printed out participant response data, such as total number of correct responses.

In 1974, Lushene, O'Neal, and Dunn (cited in Fowler, 1985) developed a computerised format of the MMPI, which administered, scored and interpreted participant responses, and which was comparable with the traditional paper-and-pencil format of the test. The 1970s also saw the first, large-scale, practical application of computerised psychological testing (CPT) and score interpretation. In the early 1970s, the computerised Psychiatric Assessment Unit (PAU) was developed in Utah, which administered a battery of tests to psychiatric clients and was found to produce patient assessments superior to traditional assessment modes, such as higher internal consistency and cheaper patient reports (e.g. Klingler, Miller, Johnson, & Williams, 1977, cited in Fowler, 1985).

The form of test-taker response has also changed over time, due to computer technology developments such as the ability of participants to respond on attached keyboards. In the early days of computerised psychological testing, psychology clients marked cards that were "read" into the computer (Fowler, 1985). Optical scanners have had a long history in computerised testing, where test takers marked special answer sheets with carbon pencil and the scanner used light to detect the coordinates of these pencil marks (Burke & Normand, 1987). These marks were then translated to data, using special software, which was then stored on an output medium such as magnetic tape. The optical

scanner has been improved by linking a computer to the scanner, so that test taker results could be analysed and reported immediately after the scanning process, thus reducing the amount and cost of the software required.

In 1977, some PAU researchers developed Psych Systems, the first company marketing hardware and software for psychiatric testing, covering all stages of testing from administration to interpretation (Fowler, 1985). The second company to market such testing software and hardware was CompuPsych, which marketed desktop microcomputer-based tests, although Fowler (1985) does not state when CompuPsych entered the interactive testing market.

Elithorn and Telford (1969, cited in Bartram & Bayliss, 1984) provided one of the first studies on the application of computer technology to ability testing, automating a maze test to determine problem-solving strategies of participants. However, most research on automated tests has occurred with personality tests rather than for ability tests. This was due to the higher complexity involved in automating tests that include pictures, such as the WAIS. The theoretical basis of such CBTI personality assessment was found in the rising interest in actuarial prediction, based on the 1950s work of Hathaway and Meehl (Fowler, 1985). This work provided the basis for the decision rules involved in generating computerised personality reports.

The major problems that limited the entry of computers into applied psychological fields were: (a) the high cost of microcomputers; (b) the necessity for undedicated computers to timeshare thus causing inconsistency in timing during test administration; and (c)the fact that computer development and maintenance typically interfered in timed data collections (Kennedy, Wilkes, Dunlap, & Kuntz, 1987). For example, undedicated computers could not administer speeded tests, or tests with speeded items, because these computers used timesharing and thus control of timing on these computers could not be guaranteed. Two main factors led to the increase in fully or partially computerised tests (Bartram & Bayliss, 1984). First, microcomputers have become widely used, both in business and in the home, increasing the feasibility of using computerised tests. Second, the cost of skilled personnel has increased while the cost of computers has decreased. Thus,

using a computerised test decreases the cost involved in psychological testing, as less psychologist time is required for the administration and scoring of computerised tests.

Generally, studies comparing the computerised format of a test with its traditional administration format have found no significant differences between test administration formats, for example the different test formats have produced very similar mean test scores and comparable test-retest reliabilities (Bartram & Bayliss, 1984). Studies of computerised personality tests, such as the Eysenck Personality Inventory, have found that the computerised format of a test is equivalent to its paper-and-pencil format. This equivalence is demonstrated by high correlations for scale scores between test formats, and by no significant differences in mean scale scores. However, the psychological tests that have been computerised have tended to be power tests rather than speed tests, and contain test items that are wholly textual, requiring forced-choice or multiple-choice formats. With developments in computer graphics quality, computer vocal output (and input) facilities, central processor power, and hard drive storage capacity, the utility benefit of computers to psychology is ever expanding.

Disadvantages and Advantages of Computerised Psychological Tests

Many concerns about the initial CPT software have continued to be raised against later systems. It has been suggested that the computer may dehumanise assessment, although studies have found the majority of participants undergoing computerised psychiatric testing (rather than ability testing, intelligence testing, and so forth) prefer the computerised test formats to traditional paper-and-pencil format (e.g. White, 1983, cited in Fowler, 1985). Researchers appear to agree that test takers easily establish rapport with CPT systems, and that this appears to stem from user-friendly assessment software (Bartram & Bayliss, 1984). Negative concerns about psychological testing may simply result from negative investigator concerns and not from actual participant attitudes. Staff preparation and education appear to alleviate many negative staff attitudes towards CPT (e.g. Klonoff & Clark, 1975, cited in Burke & Normand, 1987).

There are 4 main advantages in using computerised tests rather than pencil-and-paper tests (Bartram & Bayliss, 1984). Firstly, an automated test costs little to use, whereas the equivalent paper-and-pencil format requires appropriately trained staff to administer and score. Also, when the computer or terminal is not in use for assessment, it can be used to perform other office tasks, such as word processing. Thus, automated testing is cost-effective. While adaptive testing, where test item difficulty is a function of test taker answering, has been the CPT area where the greatest cost-saving has been identified, nonadaptive CPT software also provides reliability, speed, and economic benefits to test users (e.g. Space, 1981, cited in Burke & Normand, 1987). However, these benefits can be negated by the use of inadequate, excessively expensive, or unnecessary hardware.

Secondly, automated tests enable testing environments to be easily controlled, whereas traditional paper-and-pencil tests inherently involve tester-testee personality interactions (Bartram & Bayliss, 1984). Thus automated tests enable better standardisation of the testing environment, although the equipment used for CPT should strongly adhere to the equipment used in the validation of the CPT itself. For example, if the validation involved colour VGA screens, then all testing should use colour VGA screens (Skinner & Pakula, 1986).

Thirdly, automated testing increases test presentation speed, prevents transcription errors, and can provide score interpretations immediately following the test administration (Bartram & Bayliss, 1984). In other words, CPT-generated results - such as reports - are inherently more reliable than human-generated reports, including reports generated for ability tests (e.g. Myers, Schemmer, & Fleishman, 1983, cited in Burke & Normand, 1987). This increase in reliability may be due, at least in part, to the inability of test takers to give responses out of the range for the test item, or to choose multiple responses where only one response is indicated (Doherty & Thomas, 1986, cited in Rosenfeld, Booth-Kewley, & Edwards, 1993). The constancy of CPT administration also increases the efficiency of testing. For example, researchers can test more participants with computerised tests than with conventional-format tests, within the same time period. A related advantage is that, with CPT, the test taker cannot mismatch test and answer booklets, so all test items will be answered with the pertinent test responses (Byers, 1981, cited in French, 1986). However,

a related disadvantage is that CPT software can consistently produce invalid reports if data is only partly correct or is faulty, for example if the standardisation process for scores is incorrect (Burke & Normand, 1987). Obviously, this criticism also applies to paper-and-pencil testing.

Fourthly, automated testing provides flexibility, both in test choice and in item selection within tests (Bartram & Bayliss, 1984). As computers are capable of storing, working upon, and merging, very large files, raw data and test statistics such as norms, reliability, and validity can be easily and quickly accessed, updated, and reappraised.

Computerised testing also provides avenues of data collection that are unavailable, or simply very tedious to collect, with paper-and pencil test formats, such as item response latency. This type of ancillary data can be used in conjunction with actual test responses to determine participant response patterns (Skinner & Pakula, 1986). For example, response latency is becoming more popular for examining individual differences, although it has been limited to studies involving simple tasks (e.g. Jensen & Munro, 1979, cited in Hofer & Green, 1985). Long response latencies on particular ability test items may help to indicate those items that have higher discrimination rates between population groups. However, response latencies can also be created by incidental test behaviour, such as the participant falling off their chair (Skinner & Pakula, 1986).

Rafaeli and Tractinsky (1989) examined three areas of visual cues and response time relating to CPT software: (a) whether response time should be measured; (b) whether response times should be limited by the software; and (c) whether special cues about time should be visible on test taker monitors. Their study found that accuracy and speed performances were highly correlated, as the time information displays provided during test administration increased the probability of correct responses and reduced the average response time of test takers. Fekken and Jackson (1988) examined the utility of four test item response models, two of which were based only upon item characteristics, the other two based on item and individual characteristics, in predicting the responses of individuals to personality test items. While the social desirability characteristic of items strongly predicted which items participants would change on retesting, response latency was the strongest predictor of which item response changes between testing administrations, probably due to

this variable indicating those items that test takers found most difficult. This decision difficulty may influence item instability. Thus, time-related testing issues seem important, and require more study. Of course, this information would be tedious or impossible to collect under paper-and-pencil administration conditions.

Given these existing advantages of CPT and the increasing developments in computer technology and psychometric test theory, it appears that CPT will be progressively superior in many respects to conventional paper-and-pencil testing.

Chapter 2: Issues in Computerised Testing

General Overview

The issue of who is qualified to use CPT software has become pertinent to psychologists (Fowler, 1985). In 1953 the American Psychological Association (APA) published its ethical standards that established the particular qualifications and experience that would-be test purchasers required in order to purchase specific tests (cited in Fowler, 1985). However, test publishers sold clinical tests to physicians as well as to psychologists, and 1960s CPT mail-in assessment distributors supplied reports to psychiatrists and psychologists as both professional groups had the qualifications and experience required to purchase the corresponding traditional paper-and-pencil formats of such tests (Fowler, 1985). On-line testing distributors practiced a similar policy.

In 1966, the APA published standards for automated test systems, endorsing the provision of automated psychological test services to those individuals and organisations that only used such tests under the active supervision of appropriately qualified and trained personnel (cited in Fowler, 1985). The catalyst for this concern was the development of mail-in computerised personality testing in the 1960s. While the APA had an existing policy against mail-order testing, an APA committee decided that mail-in testing was different to mail-order testing. Essentially, the APA distinguished between the client-professional interaction of mail-order testing and the professional-professional interaction of mail-in testing. An example of mail-in testing is the Position Analysis Questionnaire, a job analysis questionnaire that is used by a psychologist and then mailed to the USA for analysis. Mail-order testing is not currently available in New Zealand, although the American company Caliper are presently attempting to introduce mail-order instruments such as personality tests, which are designed to be administered by people other than psychologists and which are sent to the USA for analysis. In response to mail-in testing, the APA developed an interim set of guidelines in 1966, covering automated test scoring and interpretation (APA, 1966).

These standards have not lessened the concern among psychologists regarding psychologist access to on-line services, and concern about non-qualified and untrained

individuals accessing mail-order CPT software (Fowler, 1985). These interim standards were then followed by references to computerised testing in 1974, 1975, and 1985 APA publications on standards for psychological and educational testing, culminating in a 1986 APA publication devoted to guidelines for computerised test administration, scoring, and interpretation (Committee on Professional Standards and Committee on Psychological Tests and Assessment). The purpose of these guidelines was to apply existing APA testing standards to CPT. The proliferation of computerised psychological tests also led to the British Psychological Society to comment on computerised tests (Standing Committee on Test Standards, 1984).

The 1986 APA guidelines for computerised tests arise from the 1981 APA ethical principles, the 1977 APA provider standards, and the 1985 APA testing standards (cited in Committee on Professional Standards and Committee on Psychological Tests and Assessment, 1986). The guidelines are categorised into: (a) user responsibilities, for example the psychologist using the test has personal responsibility for their own use of computerised testing or test interpretation; (b) guidelines specific to computerised testing and interpretation, for example the testing environment must aid optimal test performance of test takers; (c) developer responsibilities, for example psychologists developing and standardising computerised tests must use appropriate scientific procedures and adhere to the relevant APA standards; (d) guidelines for developers of computerised testing services, for example the hardware and software of the test must not frustrate or impair test taker performance.

The British Psychological Society, via the Standing Committee on Test Standards, published 12 notes on computerised psychological testing in 1984, that are in the form of comments rather than guidelines. Essentially, the Committee point out that computerised testing is likely to increase, and that the test taker-computer interaction is likely to be qualitatively different from the interaction between paper-and-pencil test and test taker. As the computerised test may thus differ from the paper-and-pencil format, the Committee suggest that statistical comparisons, such as inter-mode reliability coefficients, must be computed to demonstrate the comparability of the computerised test to its paper-and-pencil format.

CPT competence requires practice and responsibility standards, as well as professional consensus, published in professional standards (Hofer & Green, 1985). Testing must be performed competently, as any publicised abuse of CPT may cause the whole area of CPT to be viewed negatively, both by the general public and by professionals. Professional standards on CPT would also provide judges and lawyers with a standard against which to compare any malpractice or negligence cases involving CPT, although many CPT issues are subsumed under general psychological practice standards. Standards specifically covering CPT software have been published from the early 1980s, for example the Guidelines for Use of Computerized Testing Services (Colorado Psychological Association, 1982, cited in Hofer & Green, 1985).

Copyright Concerns

A major concern for CPT developers is breach of copyright due to the piracy of CPT software (French, 1986). As most of the ongoing revenue from traditionally administered testing arises from the sale of item answer sheets, and as CPT requires no answer sheets, developers essentially have two options regarding income. First, CPT developers can charge a one-off fee for the software, and clearly express in the registration information the particular conditions which apply, for example that only one installation on one computer is allowed per registration number, or they can provide floppy disks that keep a record of the number of test administrations, and then auto-erase once a certain number of administrations is reached. The main problem associated with auto-erase disks is the potential for these disks to accidentally auto-erase too soon. Thus, for the protection of software users, hard drive installation of CPT software is the best option.

Privacy of Test Results

Suggestions have been made that the amount of data storage space on hard drives, CD-ROMS, and so forth, creates a larger potential for the abuse of participant privacy, for example unauthorised persons accessing test results (French, 1986). However, ensuring participant privacy for computerised tests little differs from ensuring privacy on traditionally

administered tests, only the mode of storage differs. The problem of ensuring computer data cannot be accessed by unauthorised persons has become a more pressing issue since a number of homes and businesses have computers linked to modems, and thus can electronically access some computer systems via telephone lines. If the data are stored on a computer that is not linked to a network, such as a stand-alone PC, then the only possible unauthorised access is via the local "logon" options on that particular computer. It is possible both to password protect the computer itself, or password protect certain files, so the greatest potential for abuse arises with networked computers, especially when these can be accessed remotely via a modem connection. Thus, the safest option is for participant data to only be held on computers without network and/or modem attachments.

Validity Issues

One major point is that practitioner requirements for demonstrating the validity of computerised tests also apply to traditional paper-and-pencil tests, so practitioner ability in CPT requires no special skills peculiar to using that particular test format (Hofer & Green, 1985).

The bulk of criticism on CPT software has been directed towards the validity of CPT reports (Fowler, 1985), and the most basic requisite of any report is validity (Hofer & Green, 1985). Up to 1984 there was little published research on the validity of CPT reports generated by computerised personality tests (Lanyon, 1984). Generally, researchers agree that the interpretation accuracy and utility of computerised reports is difficult, due to methodological constraints such as the lack of access clinicians have to the rules by which software engineers generate these reports (Fowler, 1985). When the traditional paper-and-pencil format has already been empirically demonstrated to be both reliable and valid this task is less difficult. However, depending on the software company, the software engineer may not be a psychologist. In this situation, the person writing the computerised test is not qualified to administer or interpret psychological tests, yet writes programs that perform both tasks.

While, in the 1970s, researchers have criticised CPT software as being grounded in clinical lore, rather than empirical research (Fowler, 1985), this situation is becoming less common due to greater efforts by researchers to publish data relating to the reliability and validity of computerised psychological tests. The validity of CPT reports must continue to be demonstrated in order to establish they are providing useful and accurate data (Hofer & Green, 1985).

The acceptance of the need to demonstrate CPT validity is shown by the large number of published articles addressing this issue. A number of suggestions have been made to improve the reliability and validity of computerised tests. Validation studies must sample a representative group of users or, ideally, potential users of the particular computerised test (Snyder, Widiger, & Hoover, 1990). These participants must not systematically differ from nonparticipants, especially in general demographic characteristics such as gender and race. Also, the participant population used in the validation study must be a sample spread across the behavioural domains measured by the test, for example participants should differ in ability level for achievement test validation. Moreland (1985) suggests that the CPT administrators in validation studies must be representative of the actual test administrators in applied settings. The reason for this suggestion is that the validity and reliability of applied psychological testing arises from the practitioner using the test, so practitioners must be qualified both to interpret resulting CPT reports and to decide whether any CPT report is appropriate for each particular client, taking test norms and validation research into account (Hofer & Green, 1985).

Focussing on the computerised test itself, CPT practitioners must know how the individual items and test subscales are used, including the order in which information is presented (Conoley, Plake, & Kemmerer, 1991) or, in other words, the individual test scale, or scale combinations, from which the test taker scores are derived (Hofer & Green, 1985). The actual test taker scores on each appropriate scale in a validation study must also be provided to test users (Hofer & Green, 1985). The empirical evidence underlying test interpretation is another requirement in judging CPT validity. Examples of adequate empirical data are estimates of consistency and confidence (Hofer & Green, 1985), the

discriminant validity of test scores (Moreland, 1985), and probability information on predictive or classification items (Conoley et al., 1991).

Regarding test reports, CPT developers must provide the empirical evidence underlying test interpretation (Hofer & Green, 1985), the discriminant validity of test interpretations (Moreland, 1985), and - for each interpretative statement - whether it is derived from research or expert judgment (Conoley et al., 1991). If the interpretative statements are derived from judgment, the identity of the experts must also be provided. Finally, the report must link response interpretations to the appropriate scales (Hofer & Green, 1985).

Equivalency Between Computerised and Paper-and-Pencil Test Formats

This section provides the rationale for Study 1 of this thesis, which compares the paper-and-pencil format of the ACER Advanced Test BL with a computerised format of this test.

Studies comparing computerised tests with their traditional administration formats have tended to find high equivalence correlations between these two formats (Bartram & Bayliss, 1984). While there appears to be little reason to suspect that the computerised format of a paper-and-pencil test will have lower reliability, it is questionable whether the norms developed for the traditional format would also automatically apply to the computerised format. This is the equivalency issue in CPT. For example, is the VGA colour presentation of a test equivalent to paper-and-pencil test presentation? Are either of these presentations equivalent to a monochrome computerised version of the test? These examples of possible threats to the equivalence of different formats of a test are extraneous variables inherent to the administration phase of testing.

Hofer and Green (1985) have criticised CPT administration for containing non-test factors that may influence test results, and the reports generated from these. First, CPT reports tend to be assigned greater face credibility than are reports originating from paper-and-pencil testing, due to the higher objectivity assigned to computers over humans. Thus, psychologists are less likely to be critical consumers of computerised tests. Second,

professional reviewers of CPT software are normally not provided with computer algorithm data, such as decision rules, thereby restricting the detail and value of their reviews (e.g. Mitchell, 1984, cited in Skinner & Pakula, 1986). It is probable that professional reviewers will be unable to detect the influence of non-test factors.

The reliability, validity, and normative data of the traditional test format cannot be simply generalised to include the computerised format, as participants will receive one score on the traditional format and another score on the computerised format (Burke & Normand, 1987). Norm data must be collected on the computerised format to determine the equating formulae used to generate conventional statistics, such as cutting scores, which can then be applied to the computerised format. Demonstrating equivalence between the traditionally administered test and the CPT format is not simply a matter of showing high test score correlations between the two formats; there must also be evidence that test score frequency distributions are almost identical and that there are only minor changes to test taker rank between formats.

The types of data and data statistics (such as norms) originating from traditional test administration that can be applied to the data arising from computerised administration of the test depends on the nature of the score differences for each test format *and* if an equating method can be applied (Hofer & Green, 1985). Users of CPT software must know what equivalence data exists for every test used, and how this data influences the generalisability of data from traditional administration to the computer-based administration results. If the two modes of testing are wholly equivalent, then data obtained from the traditionally administered test mode can be used to interpret the data from the CPT format. If rank order differences occur in the computerised format, no data derived from the traditional format can be used to interpret the CPT-based data because a change in construct between the two test formats has occurred. However, this suggestion seems rather extreme if only slight rank order differences between the two formats occur. If the data between formats has metric or mean changes and equating formulae exist, CPT-based data can be interpreted from the traditional administration data (Hofer & Green, 1985). In this case, if there is no equating formulae then normative data based on traditional administration cannot

be used and validity data based on traditional administration may also not be appropriate for use.

Thus, comparisons of CPT-based and traditional-based data must report distribution data, and test score correlations if the study uses a within-participants paradigm. The probability of demonstrating equivalence is higher with a multiple-choice power test of fixed length, where there is little format change between test formats (Allred & Green, 1984, cited in Burke & Normand, 1987). Thus, ability tests should have higher equivalency between computerised and paper-and-pencil formats than personality tests.

Chapter 3: Human-Computer Interaction

The "person-machine" interaction inherent in computerised testing has concerned researchers from the onset of computerised psychological testing (Bartram & Bayliss, 1984). Developers of any computerised test must ensure that human factors issues are adequately addressed (French, 1986). Skinner and Pakula (1986) have identified three major factors that appear to influence the person-machine interaction: structure, process and function factors. Structure factors are those directly related to human-computer interaction, such as computer input and output modes. Process factors are the quantity and quality of user involvement in system design, such as assessment of organisational resistance to computerisation. Function factors involve the use of the computer in the applied (e.g. occupational) environment. The most problematic issues are those relating to function, as staff may disagree on the appropriateness of each type of computer use. An important point here is that CPT is a system, where the test taker uses the hardware and software at the same time. The purpose of separating hardware and software considerations is only to simplify the following discussion. In any applied setting, the hardware and software required for CPT must be considered simultaneously as each must complement the other.

Hardware Considerations

Increasing the user-friendliness of the hardware interface involves selecting input and output devices that are acceptable to the user, for example computer mice (Stevens, 1983). Carr, Wilson, Ghosh, Ancill, and Woods (1982) suggest that computerised testing must be developed with the proposed test taker population in mind. Regarding CPT hardware, the use of special types of input devices, such as touch-screens, increase the utility of computerised testing with such demographic groups as the elderly and physically handicapped. In other words, the hardware interface must feel 'natural' to the user, so that input and output devices are not novel to the typical user (Stevens, 1983). With the advent and decreasing cost of computer soundcards that enable the computer to communicate to the user in stereo sound, there is the ability of tests to be administered aurally. Software already exists that can "understand" user vocal commands. Perhaps future computerised

testing will involve vocal input and output, thereby increasing the potential of computers to test the physically impaired.

Hardware specifications also arise from the anticipated numbers of clients who will be tested within known time constraints, the number of proposed assessment/test administration sites, the graphic capabilities required, and so forth (Burke & Normand, 1987). Thus, the user-computer interface in automated testing has become a primary design and ergonomics concern, addressing types of automated tests and types of user (Bartram & Bayliss, 1984).

Software Considerations

The main interaction between person and computer occurs between the user and the software, so optimal hardware use requires that the software interface is user-friendly (Stevens, 1983). It is probably for this reason that most of the literature on the person-computer interaction deals with software, with each author presenting a different model of the human-software interaction (e.g. Norman, 1984). As one major goal of CPT software is user and test taker acceptance, CPT software must be designed to eliminate, or at least reduce, potential problems regarding the CPT acceptance of test takers (Burke & Normand, 1987).

The human-computer interaction model proposed by Norman (1984) and outlined below is a cognitive one, reflecting the contribution cognitive psychology has made to CPT. However, the work of Booth (1991) provides a simpler introduction to cognitive factors in the human-computer interaction. Booth (1991) has identified two cognitive factors inherent to CPT: tool factors and task factors. Tool factors are how the test taker uses the computer software, and are typically limited to response entry and movement between questions. Task factors are how the test taker uses the computer system to meet test demands, and can be direct or indirect. Direct task factors involve problem solving specifically related to the test questions, and indirect factors include situation-specific factors such as the order in which test items may be answered. Both tool and task factors must be successfully combined by the test taker into a strategy appropriate to the test.

To increase the likelihood of test takers to generate such adaptive strategies, and to increase test taker trust in the person-computer interface, the computerised psychological test should have three types of software consistency (Booth, 1991). First, the sequence of information should be consistent, for example test items should be presented in the same order for all test takers. However, by definition, this first type of consistency is not possible with computerised adaptive testing (CAT). Second, the placement of information should be consistent, for example each test item must fit completely on one screen. Lastly, if colour is used then this use should be consistent, for example all test items should be presented in the same colour.

Norman's (1984) four stage model of human-computer interaction provides a clear means of identifying software features that increase the user-friendliness of CPT, and provides a more detailed description of how cognitive factors relate to the human-computer interface. The four stages are: (a) intention; (b) selection; (c) execution; and (d) evaluation, although it is recognised that people do not move through these stages in a smooth sequence.

The intention stage contains two facets, the ability of the system to know the user's intentions and the software support to the user to help them form appropriate intentions (Norman, 1984). Here, intentions are defined as cognitive specifications of action that initiate and guide subsequent activity. To meet these user needs, the software must provide help for incorrect user responses, for example by displaying an error message, and provide the user with information on their current status in the test and what options are available to them at that point, for example showing the user which question the cursor position relates to and the answering options for that question, and the entry required to move from the selected question to the desired question.

The selection stage also contains two facets: the method that must be used to complete the task and the system commands that are used to accomplish this (Norman 1984). System support for the user in the selection stage comes from on-line help, such as prompts. The execution stage is where the user specifies an action to the computer, and this is performed either by naming the appropriate command, such as occurs in programming, or by indicating which of the actions provided by the software is to be performed, such as by

typing in the multiple-choice answer in a test. For these three stages, the user must be able to understand the software, and should not have to guess the inputs expected by the software (Stevens, 1983).

In the evaluation stage the user is informed whether their selected action (from the execution stage) is correct or incorrect (Norman, 1984). Applying this stage to a computerised test, the first process is to determine whether the action of the user is correct or incorrect, that is, whether the user's response to an item is within the range of allowable input options. For example, in a multi-choice test of 5 alphabetical answer options per question, a numeric input by the user would be an incorrect action as this input is out of range. There are three possible system responses to such an action: (a) allow the input and not prompt the user; (b) disallow the input but do not provide an error prompt to the user; and (c) disallow the input and provide an error prompt to the user. A second possible process is to provide immediate feedback to the user, for example by displaying a "correct" message if their answer to that test item is correct and displaying an "incorrect" message if that is not the case. Thus, during all four stages, the software should be helpful, for example by enabling the user to ask for help and responding adequately to such requests (Stevens, 1983).

Kearsley (1986) suggests general ways in which software can be improved. Screen design is the most salient feature of the software to the user, and screen displays should be uncrowded with text, use graphics instead of text where possible, and should contain titles and headings. Graphical emphasis such as highlighting text and multiple colour combinations should be used minimally, as these have the potential to lower display legibility. The user must be able to control the software, for example the user should be able to determine the pace by which they proceed through items on a computerised test, and users should be able to control the sequence through which they proceed through test items by being able to skip items. Obviously, this second consideration does not apply to CAT. The software must also be able to provide feedback on user responses, such as indicating which option the user has selected as an answer to an item, allowing users to change their responses to items. The program should not rely on case-sensitive input where the user must input an alphabetical character to a test item: both upper case and lower case

characters must be recognised. The software must be tested for expected and unexpected responses before release, for example if an alphabetical character is required as the input to a test question then the programmer must check that numeric, out-of-range alphabetical characters, and characters such as tabs are not accepted as valid input. The software must provide specific online help to the user that is constantly available and easy for the user to access. The help information must also be accurate and complete. An example here is a screen prompt on the command the user must enter to move between test items.

Visual Factors And CPT

Due to the salience of the screen display to the test taker in CPT, as this is the hardware device that presents the test to the test taker, visual factors involved in CPT must be examined in some depth. General display research will be examined first, then research on reading display-presented text, and finally research on CPT itself.

Lie and Watten (1994) conducted one experimental study and one clinical study on VDU work and eye strain. In the experimental study, experimental participants proof-read text presented on a green monochrome screen continually for three hours, and control group listened to the same text using headphones, entering detected errors on a keyboard. All participants were skilled text editors. There was a significant change in myopic direction for both eyes for the experimental group during the three hours, which did not occur in the control group. There was also a consistent difference in vergence and ciliary muscular ability after three hours of VDU work between the experimental and control groups, with all Zone of Clear Single Vision (ZCSV) changes for experimental participants over the three hours reaching statistical significance, but none of these changes reaching significance for the control group. The ZCSV measures accounted for 68% of the variance in reported optometric symptoms, such as dizziness and headache, between pre- and post-VDU work symptom measures for experimental participants. No such relationship was found for control participants.

In the clinical study, participants from the first study with severe symptoms were given optometric examinations and optical corrections in the form of spectacles (Lie &

Watten, 1994). The participants in this clinical study were given pre- and post-correction assessments consisting of subjective symptom reports and optometric measures. Post-correction assessment was conducted at the six month point. There were significant improvements on both the optometric measures and reported symptoms. The results of these two studies suggest that prolonged VDU work causes optometric problems that are reflected in subjective complaints such as muscular pain in the neck, and that these problems can be reduced by vision correction. Watten, Lie, and Birketvedt (1994) conducted a field study of 45 workers whose jobs primarily involved interactive computer work, such as manual data entry. ZCSV measures were taken before participants started their daily work and after they had completed it. The results indicated that prolonged near-work significantly decreases the contraction capacity of the ciliary muscles, the capacity of the extraocular muscles, and the ability of the lenses to stretch. These physiological changes led to marked deterioration of vergence and accommodation.

Typically, current displays use colour, so it is appropriate to examine research on the effect of colour displays. Neri, Luria, and Kobus (1986) conducted a series of two studies on the effect of foreground and background colour pairings on target search reaction time on a colour graphics terminal. In the first study, the background colours were grey, red, yellow, green, and blue, the foreground (target) colours were red, orange, yellow, green, blue, and purple, and lighting conditions were no light, low level white, blue, and red. A target was never the same colour as the background. There was no significant effect of lighting on mean reaction time, mean reaction time was significantly faster for blue background than for green, yellow, or red background, and target colour significantly interacted with background colour on mean reaction time. The general interaction pattern was the farther apart the background and target colours, the faster the mean reaction time. For example, green and blue targets had the fastest reaction times on a red background. However, for the grey targets, the fastest reaction times were on the red and blue backgrounds, and the slowest were on the yellow and green. In the second study, low level white illumination was used, the background colours were blue, green, yellow, and black, and the target colours were red, orange, yellow, green, blue, purple, and grey. Again, a target was never the same colour as the background. However, in this study the background

colours were matched for brightness. There were no significant differences in mean reaction time between the background colours, and there were no significant differences in mean reaction time for target colour. These results indicate that chromatic contrast enhances visual detection on colour terminals, but brightness contrast is superior to chromatic contrast. Ambient lighting did not affect this result.

Hughes and Creed (1994) conducted two studies on the eye movements of participants familiar with avionic displays. In the first study, participants had to locate 5 aircraft variables in monochrome and colour slides of horizontal situation indicator (HSI) displays, under low and high complexity situations in a counterbalanced design. Complex displays contained an extra 4.3 symbols. There was significantly less eye scanning time and less individual fixations for the colour displays than for the monochrome displays for active waypoint detection (the first aircraft variable participants had to locate) but no significant differences for colour condition for the other 4 variables. These results indicate that target discrimination was lower for monochrome displays than for colour displays. The second study essentially replicated the first, using the same participants, except the waypoint was the only aircraft variable and the complex displays contained an extra 11 symbols. Significantly higher average gaze duration and significantly more fixations occurred in the complex displays. There were no significantly different eye movement patterns between the two colour conditions. The results of these two studies suggest that colour information is more important than spatial code for targets when the spatial location is unknown, but there is no colour advantage when target location is predicted by other visual cues.

Thus, the general finding is that colour does play a role under some reaction time conditions. Of more importance to CPT is the function of display colour in reading text. Belmore (1985) compared reading time and comprehension of undergraduate student participants on paper presentation and computer presentation of passages from reading texts, where presentation format was counterbalanced. Most participants had no familiarity with computers. In the paper presentation condition, each passage was typed onto one sheet of paper. In the computer presentation condition, a monochrome screen was used and no passage more than two screens in length. Participants were able to scroll between the screens of a passage. Presentation order had no effect on reading time, but the computer

condition produced longer reading times than the paper condition, although this was only significant for the first four passages. Comprehension was significantly different between the presentation modes, and order of presentation mode was also significant. Participants had a significantly shorter reading time and significantly higher comprehension for the paper condition than for the computer condition, although this result did not occur when the computer condition was presented after the paper condition.

In a series of two studies, Creed, Dennis, and Newstead (1987) examined the accuracy of proof-reading prose presented on a VDU, using undergraduate students as participants. In the first study, participants proof-read text on a green monochrome VDU, on a photograph, and on paper, and the order of text presentation format was counterbalanced. Proof-reading accuracy was highest for the paper condition and lowest for the VDU condition, and these results were significant. The second study was designed to test the hypothesis that reduced accuracy in the VDU condition was due to participant difficulty in reading the computer text. This second study consisted of a 2x2 experimental design - single or dual column by VDU or paper presentation - and the order of experimental conditions was again counterbalanced. Proof-reading accuracy was significantly higher for the paper condition, and time to complete proof-reading was significantly faster for the paper condition, but there was no significant difference for column type. The results of these two studies suggest that the legibility of monochrome text is lower than that of text presented on paper.

Horton and Lovitt (1994) compared the reading comprehension of learning disabled and normal secondary students for paper text and computer text passages, using an equivalent samples design. The type of computer monitor used was not described. The results suggest that reading comprehension was higher for the computerised text, for both types of student, however the significance level of this finding was not reported in the study. Rosenfeld, Doherty, Vicino, Kantor, and Greaves (1989) found that monitor resolution and colour displays had little influence on survey responses to a workforce attitudes instrument. Thus, the literature offers mixed support for the idea that computer-presented text is more difficult to read than paper-presented text. However, the two studies that suggest computer-presented and paper-presented texts are essentially equivalent in legibility were

conducted more recently. Perhaps the participants in the Horton and Lovitt (1994) and Rosenfeld et al. (1989) studies simply had greater familiarity with computers than the participants in the Belmore (1985) and Creed et al. (1987) studies. The influence of computer familiarity on test taker performance is the subject of the following section.

Test Taker Familiarity With Computers

Overview

Historically, much of the research dealing with the differences in person-computer interaction as a function of computer familiarity has examined computer software other than psychological tests, such as programming languages, text editors (Allwood, 1986). As such software involves factors, such as complex input decisions, that are not present in CPT, an application of this research to this present study will not be performed.

The impact of test taker familiarity with computers is an important issue as most legal claims regarding applied psychological testing have centred on test bias against women, ethnic minorities, and handicapped (Hofer & Green, 1985). This bias may also exist regarding computerised testing, as the availability of computers may be a function of job level for work-based computer familiarity, and personal income level for home computer familiarity. The proposition that people familiar with computer use may have an advantage over those individuals unfamiliar with computers seems to be a valid concern, especially if equipment unfamiliarity is one variable influencing participant test anxiety.

Hofer and Green (1985) suggest that computer familiarity is a function of age, gender, ethnicity, and socioeconomic status interactions, so poorer performance of certain population groups on CPT software may simply be a direct result of computer unfamiliarity rather than group membership per se. However, little research has been directed at the issue of CPT disadvantaging certain population groups, such as the elderly, so any conclusions are essentially speculative. Research on possible CPT biases against certain group members must be directed towards locating specific factors, such as unfamiliarity with computers, as such research focussing on general factors, such as race, will probably provide no answers as any bias may not be linear for all members of a population group.

Research on Test Taker Familiarity With Computers

This subsection provides the rationale for determining participant computer, mouse, numeric pad, and typing ability in Study 1 and Study 2 of this thesis.

It is possible that test takers unfamiliar with computers have low computer self-efficacy, causing lower CPT performance in these individuals compared to expert computer users. Torkzadeh and Koufteros (1994) administered the Computer Self-Efficacy Scale to undergraduate student participants enrolled in a computer course at two stages: immediately before course exposure and immediately following course completion. There were significant increases in self-efficacy factor scores between the two self-efficacy test administrations, suggesting that targeted computer training significantly improves the self-efficacy of computer users.

It appears to be appropriate to provide CPT test takers with practice on both equipment and procedure to focus test takers on test item content (Hofer & Green, 1985). This process can be achieved by allowing test takers to practise on sample test items until each test taker decides they are ready for the actual test administration. Lushene, O'Neil, and Dunn (1974, cited in Hofer & Green, 1985) found that initial participant computer anxiety can be eliminated by providing test takers with adequate practice. However, issues that then arise are the operationalisation of "adequate practice," and how such practice time differs both between participants for the same test and between tests for the same participant.

Test practice for standardised tests is a controversial area, with many studies examining the effects of practice on American student Scholastic Aptitude Test (SAT) scores. Messick and Jungeblut (1981) reviewed studies on SAT coaching and found that while many were methodologically flawed (such as lacking a control group), coaching increased student performance on two SAT subscales, although coaching causes only a slight performance increase on these subscales. However, few of these studies have addressed the particular test item characteristics, and thus appear to ignore the finding of Vernon (1954, cited in Powers, 1986) that some items are more susceptible to practise and

coaching effects. There are two strategies that address practice and coaching effects on tests: the provision of practice tests and familiarisation materials to all prospective test users; and test construction that eliminates items susceptible to practise and coaching. Powers (1986) reviewed ten studies that examined either special test preparation or participant practice within the test, finding that the more complex the test item, the more susceptible it was to coaching, test practice, or test preparation.

Beaumont (1985a) used two studies to compare the performance of undergraduate students on a computerised digit span task. The first study had three input response conditions - computer keyboard, numeric pad, and light pen - and the results showed a significant effect for response device, but not for digit span direction (i.e. forwards or backwards). The keyboard condition produced the highest mean digit span score and the light pen condition produced the lowest mean digit span score. The second study had four experimental conditions: (a) computer presentation with touch screen input response; (b) computer presentation with keyboard input response; (c) computer presentation of stimuli with verbal response; and (d) conventional test administration (verbal) and response (verbal). The results showed a significant effect for experimental condition and for digit span direction, but no interaction between these factors. The verbal-verbal condition produced the highest mean digit span score, and the computer-touch screen condition produced the lowest mean digit span score. The computer-keyboard condition produced a higher mean digit span score than the computer-verbal condition. These results indicate that the lower the participant familiarity, or operating ability in the light pen condition, with a test input device, the lower the participant score.

In a second related study, Beaumont (1985b) compared the response latencies of undergraduate students on a simple, self-paced computerised continuous performance task. There were four input response conditions: keyboard; numeric pad; light pen; and touch screen. No participants were familiar with touch-screens or light pens, although participants had varying familiarity with keyboards and calculators (similar to a numeric pad). Familiarity was entered as a variable. The mean response latencies for input device were significantly different. In order of increasing mean latency, the input conditions were: (a) touch screen; (b) keyboard; (c) numeric pad; and (d) light pen. No input error analysis was

conducted. These results suggest that the test results of naive computer users are influenced by the type of input device they use.

One method that has been used to aid the performance of test takers unfamiliar with a keyboard is to physically cover all keys not required during test administration, to reduce keyboard complexity (e.g. Styles, 1991). Styles noted that the provision of a keyboard cover reduced embarrassment and anxiety in child participants unfamiliar with a keyboard, and also prevented those participants familiar with computers from accessing particular keys such as the escape key. The other advantage of the cover was the elimination of accidental key pressing, for example by participants leaning on the keyboard.

Little research appears to have been conducted into the factors inherent in the use of computer mice as input devices. This may be a function of the fact that many studies on computer input devices were conducted before the widespread use of mice, or simply because many computerised psychological tests do not appear to currently include a mouse as an input option. However, as new computers are now supplied with a mouse and as popular software, such as word processing packages, are now relying heavily on the use of mice, it is anticipated that mouse-based psychological testing will become the industry standard.

In a series of two studies, Bedford (1994) examined the spatial factors involved in translating the movements of a pen on a digitising tablet onto the screen representation of the pen movements. The studies used undergraduate students as participants, and the task was to use the pen to move a character into one of nine randomly selected spatial locations. The computer familiarity of participants was not described. Participants were provided with practice sessions, and were obstructed from seeing their hands and the pen. In the first study, participants were assigned to one of two experimental conditions: (a) a left-right pen movement moved the character twice this distance on the screen (X condition); or (b) both the left-right and top-bottom pen movements were doubled on the screen (XY condition). In the second study, participants were assigned to one of two conditions: (a) no alteration to pen movement magnitude (N condition); or (b) a top-bottom pen movement moved the character twice this distance on the screen (Y condition). For the N and XY conditions, where symmetry existed between the two directions of pen movement and the movement of

the character on the screen, participants encountered little difficulty translating the movement of the pen to the movement of the character on the screen. However, while participants in the Y condition learned that only this pen direction had to be reduced in magnitude, participants in the X condition tried to apply the magnitude change to the top-bottom direction as well. These results suggest that the spatial dimensions of X and Y are dependent.

Computer mouse movements are similar to the pen task described above in the Bedford (1994) studies, thus the findings of these studies are important to psychological tests that involve mouse input by test takers. Mouse movements in any direction must have the same magnitude translations to cursor movement to reduce the difficulty for test takers positioning the mouse on the response option they select. Experienced mouse users are familiar with the fact that the distance moved by the cursor on a screen is larger than the actual mouse movement, a visual-motor skill that must be learned by test takers that are unfamiliar with computer mice. Any alterations in this movement translation in one direction could disadvantage experienced mouse users, and will seriously disadvantage naive test takers. Oltman (1994), in a study where most participants were naive mouse users, found that participants in minority ethnic groups were not more disadvantaged than White participants, and females were not more disadvantaged than males, on tests of reading and mathematics. This results held for simple items, where the participant simply selected a multichoice option, as well as more complex items that required more mouse movement and more mouse clicks.

In summary, test taker practice appears to be a factor that influences scores on computerised tests. Although the number of work and home computers is increasing, as is the number of occupations that require employees to use computers, there are still certain groups that have lower access to a computer, such as people on lower incomes. For this reason, the computer familiarity of each CPT test taker must be assessed, especially when the test is an aptitude one. The studies also suggest that the input device, such as a keyboard, that test takers use may influence their test scores. Study 2 of this thesis is designed to determine if the input device used influences participant test scores and anxiety on aptitude tests.

Chapter 4: Comparisons of Computerised and Paper-and Pencil Formats

Overview

The changes in the item presentation and response formats between traditional pencil-and-paper format and computer-based format may cause differences in participant test scores between these modes, in other words item parameter variance (Leary & Dorans, 1985). Historically, studies of within-test item context have addressed four areas: item arrangement; item order interactions; repositioning of item sections (e.g. Faggen & Peck, 1981, cited in Leary & Dorans, 1985); and item parameter invariance (e.g. Yen, 1980). Monk & Stallings (1970) found that item rearrangement did not significantly influence test scores, test reliabilities, or individual item difficulty on paired geographical power tests constructed from the same item pool, with each pair containing a different arrangement of the same items, with no attempt to arrange items in order of difficulty. Munz & Smouse (1968) examined the interaction between item order and test taker performance and anxiety, and found that the difficulty sequence of test items interacts with test taker anxiety, and it is this interaction that influences test scores.

While these context effects have been demonstrated for over 30 years, some inconsistent results have been produced, and persistently replicable results are quite general: easy-to-hard item arrangements produce higher test taker scores on speeded tests than does the hard-to-easy arrangement; random item or item section rearrangement does not affect test taker scores on power tests; and test taker scores on aptitude tests seem to be more affected by item arrangement than are achievement test scores (Leary & Dorans, 1985). Noticeably, little of this research has any direct relevance to test taker score variance between computerised and paper-and-pencil test administrations.

Many questionnaires used in psychological research have a randomised format, where the construct(s) under investigation are hidden, although grouping items together is another format that has been used (Schriesheim, Kopelman, & Solomon, 1989). However, there is little research comparing the effects of different questionnaire formats. Schriesheim et al. (1989) conducted a series of three experiments to determine the format effects of paper-and-pencil tests on test reliabilities and validity, using undergraduates as participants.

The first study used a counterbalanced design to compare the effects of randomised and grouped questionnaire formats, and found that both formats were roughly equal, neither having high internal consistency, although the grouped format had a slightly higher discriminant validity. The second study examined the influence of format on the stabilities of scale scores, and found little advantage of format for stability, method bias, or discriminant validity over time. The third study examined the effects of format on respondent description accuracy, and found little difference between formats on this dependent variable. Thus, this series of studies suggests that questionnaire format has little impact on questionnaire findings, although questionnaires that group items measuring the same constructs may decrease the quality of measurement.

As test score appears to be influenced by the structure of paper-and-pencil tests, it thus appears likely that performance variance will occur between computerised and paper-and-pencil test formats. If the computer presents the items one at a time, or a few at a time, the test taker may not know how many items are included in the test, and thus the test taker may focus more on each item in the computerised format, leading to more deliberation on answering items than would occur in the traditional format of the test (Hofer & Green, 1985). This concern mainly relates to unspeeded tests and personality tests, rather than to speeded or ability tests, which demand less participant deliberation because of time constraints and/or participant knowledge that each item has a specific number of correct responses.

To determine if participant responding and scores are different on the computerised format of a test compared to the paper-and-pencil format of the test, a review of the literature on comparisons between these two test formats is necessary. The importance of this comparison is suggested by studies that simply compare differences in paper-and-pencil formats. For example, altering the answer-sheet format itself, without any alterations to actual item presentation format, appears to influence test taker scores on some aptitude tests. Boyle (1984) examined the influence of answer sheet format on participant mean scores on each of the 7 General Aptitude Test Battery (GATB) subtests. These subtests consist of multi-choice items and require the use of answer sheets. Three answer sheet formats were used: (a) National Computer Systems (NCS), where each multichoice letter is

situated immediately above its corresponding response circle; (b) Opscan, where vertical rectangles are used instead of circles; and (c) Data Research Services (DRS), which has the multichoice letters positioned immediately above horizontal rectangles. All these formats are able to be scored by machine. There was no consistent influence of answer sheet mode on participant score, although the answer sheet format and subtest interaction was significant. For the Names Comparison and Tool Matching tests, participants responding on the NCS format had a significantly lower mean score than participants responding on either Opscan or DRS formats, and these were the only two speeded tests. Thus, if simply altering the answer sheet influences test taker scores, altering the entire administration format - where both test presentation and answering formats are altered - would be expected to have a strong influence on test taker scores.

Research on Personality Tests and Tests of Affect

Regarding the responding patterns on personality tests, while some researchers (e.g. Evan & Miller, 1969) have found computerised formats of personality tests produce greater honesty and lower defensiveness in participant responses than do the paper-and-pencil formats, other researchers have not found a systematic social desirability response set difference between these two modes of administration (e.g. Booth-Kewley, Edwards, & Rosenfeld, 1992). Some research on the MMPI (e.g. Biskin & Kolotkin, 1977, cited in Hofer & Green, 1985) has shown differences between the traditional and computerised formats of this personality test, which may be a function of differences in participant item omission between these format modes. Thus, it appears that the particular population group under assessment and the specific assessment context interact to influence the generalisability of paper-and-pencil based statistics, such as norms, to the computerised format of the test (Skinner & Pakula, 1986).

Schuldberg (1990) examined the influence of repeated testing and CPT on undergraduates' responding on the MMPI. While some studies have found significant differences between MMPI formats, this finding is not shown in other studies (e.g. Dahlstrom, Welsh, & Dahlstrom, 1972, cited in Schuldberg, 1990), and little research has

addressed the individual differences involved in format effects (Schuldberg, 1990). Basically, Schuldberg (1990) proposes that these inconsistent test-retest results are due to response inconsistency between testing administrations. These response inconsistencies can be due to individual attributes, the test item itself, or some combination of these two factors. When the test format changes between administrations, the interactions become more complex. For example, test format may then interact with time, person factors, and item variables, especially when the test uses multi-choice items. Thus, inconsistent item responding may involve a number of responding styles.

Schuldberg (1990) categorises inconsistent responding on the MMPI into two major areas: (a) systematic inconsistent responding; and (b) unsystematic inconsistent responding. Systematic inconsistent responding is caused by format or time effects, and involves response shifting in a specific keyed direction on particular items over two tests. Systematic inconsistent responding is thus directional and probably related to specific item characteristics. On the other hand, unsystematic inconsistent responding is the total number of response shifts a test taker makes on all test items over two test administrations. This means that unsystematic inconsistent responding is directional shifting only when all possible types of response shift are counted separately.

When research on stability of the test across time only addresses instability on the total number of items, or on repeated items in the same assessment, the instability index is both unsystematic and nondirectional (Schuldberg, 1990). This type of instability index thus has the disadvantage of treating all shifted responses as the same. If the test taker is given two different test formats on two occasions, the unsystematic inconsistency value must be comprised of both repeated testing effects and test format change effects. That is, these two effects are confounded.

Schuldberg (1990) administered the CPT and paper-and-pencil formats of the MMPI to participants, counterbalancing format presentation. The results showed that the only significant change in true and false responding rates occurred for participants in the paper-and-pencil format when this was the second administration condition. There was an average response change on just over 94 items (Schuldberg, 1990), higher than the response shift predicted (67 items) for repeated testings using the same format (Fekken & Holden, 1987,

cited in Schuldberg, 1990). Thus, repeated testing of participants using two different MMPI formats caused higher response inconsistently than did repeated testing using one format (Schuldberg, 1990).

Participants tended to shift their responses on all items across administration times (Schuldberg, 1990). This result suggests that item shifting and item consistency variations were due more to participant characteristics than to actual item properties. Response shifting occurred more frequently on the MMPI maladjustment items than on the neutral items, and this effect was not moderated by the order of test format presentation. Inconsistent responding tended to be a general trend, occurring on all test item types.

Thus in the Schuldberg (1990) study, inconsistent responding was not due to changes in test format or to repeated testing, rather it suggested participants had a careless, deviant, or inconsistent attitude towards the MMPI. Also, response changes did not appear related to item characteristics; participants who altered their responses tended to alter them over many items, although item content correlated with response inconsistency, as shown by the higher rate of response shifting on maladaptive items. It is therefore important for researchers to use counterbalanced designs to compare inconsistency variables across both time and format.

Sanitioso and Reynolds (1992) administered paper-and-pencil and computerised forms of the Eysenck Personality Inventory (EPI) and the Adjective Check List (ACL) to undergraduate student participants. A partial counterbalanced design with a one week test-retest delay was used. Half the participants completed the paper-and-pencil EPI and the computerised ACL in the first testing session, and the computerised EPI and the paper-and-pencil ACL in the second testing session, and for the other participants this session order was reversed. However, the EPI was always administered before the ACL. Regarding the EPI, the Lie scale mean score and the Impulsivity mean score were significantly higher for the computerised format than for the paper-and-pencil format. EPI subscale scores were highly positively correlated between EPI formats, suggesting high test-retest reliability between the formats. Regarding the ACL, 20 of the 27 subscale scores were significantly higher for the computerised format compared to the paper-and-pencil format. The test-

retest reliabilities between the formats ranged from .47 to .84. There was also a significant difference in ACL profiles between the two formats.

Lukin, Dowd, Plake, and Kraft (1985) used a Latin Squares design to compare psychology undergraduate student performance on computerised and paper-and-pencil formats of the Therapeutic Resistance Scale (TRS), the State-Trait Anxiety Inventory (STAI), and the Beck Depression Inventory (BDI). All participants completed both formats, with half the participants completing the paper-and-pencil formats first and half completing the computerised formats first. The test-retest interval was one week. Before test administration all participants underwent an assessment interview resembling a counselling interview and completed post-test measures consisting of a semantic differential instrument and questions about the testing experience. Test presentation order was not counterbalanced. There were no significant differences in test score on the administration format and time factors, and there was no significant difference in semantic differential score between administration formats. Regarding participant preference in administration format, over 84% of participants preferred the computerised formats.

Glaze and Cox (1991) administered a computerised format and a paper-and-pencil format of the Edinburgh Postnatal Depression Scale (EPDS) to 29 women who had given birth within 6 months of test administration. Test format administration was counterbalanced with no test-retest delay. Computer format scores were not significantly different to the paper-and-pencil format scores, and 67% of participants preferred the computerised format with 21% expressing no preference.

Aptitude Test Research

Research suggests that some test items do not generalise easily from pencil-and-paper format to CPT format (Hofer & Green, 1985). Some literature suggests that this may be due to test taker responding. Hoffman and Lundberg (1976) found the paper-and-pencil and computerised formats of multiple-choice and true-false test items were equivalent, although items requiring matching produced different test responding behaviour and significantly lower participant scores for the computerised test formats. Greaud and Green

(1984) found large differences between participant total scores on traditional and computerised formats of a speeded test for simple arithmetic ability. Time to register a response was an important component of item answer time, and significantly affected participant scores. Part of the problem may be a lack of standardisation in the CPT test administration procedure. Hofer and Green (1985) suggest that high participant performance can be aided by the provision of a quiet and comfortable room, rest periods, a clear and glare-free monitor, obvious response equipment such as a keyboard, and uniform short time delays between test items.

One of the potential problems of using a repeated measures design, involving human participants and an ability test, is the potential for participants to remember which test items gave them difficulty and then purposely seek out the answers to these items so they can answer these items correctly in the second test administration. Kennedy, Wilkes, Dunlap, and Kuntz (1987) administered 11 tests, all of which had previously demonstrated utility in repeated measures experiments, to 25 American undergraduates using an ABAB paradigm where the paper-and-pencil test format always preceded the corresponding CPT format. The means, standard deviations, and intertrial correlations achieved stability, most of which occurred within 20 minutes of practice, suggesting that CPT was as stable as traditional-format tests.

Lee, Moreno, and Sympson (1986) suggest that the time allocated to testing, test difficulty, test demands on test taker cognitive processes, and the presence or otherwise of a human test administrator, and interactions between these factors may influence score variance between test formats, although it is difficult to understand why the first three factors would automatically differ between test formats. Lee et al. (1986) compared the mean score of military recruits completing the paper-and-pencil Arithmetic Reasoning subscale of the Armed Services Vocational Aptitude Battery (ASVAB) with the mean score of those completing the computerised format of this test. The mean score of recruits on the computerised format was significantly lower than that of recruits on the paper-and-pencil format; of the 30 test items, 21 were more poorly answered on the computerised format, 3 were more poorly answered on the paper-and-pencil format, and 6 showed no score variation between test formats. However, this study had two main methodological flaws: a

counterbalanced design was not employed, and the relationships of item content to intertest score variation were not examined.

Huba (1988) administered computerised and paper-and-pencil formats of the Western Personnel Test (WPT), a short test of general ability, to individuals applying for clerical and white collar jobs. Participants were administered both test formats, using a counterbalanced design. Forms A and B of the WPT were used, and form administration was also counterbalanced. For example, participants who completed a paper-and-pencil format of Form A first then completed the computerised Form B format. The test-retest delay was not mentioned. There were no significant differences between the participant groups on demographics such as age, gender, or amount of education. There was no significant difference in mean WPT score between the test formats, and there were no significant differences in WPT Form A and Form B scores between the test formats.

Van de Vijver and Harsveld (1994) administered paper-and-pencil and computerised formats of the GATB to applicants for a military academy. The computerised GATB was presented on an EGA colour monitor and participants entered answers using extended keyboard keys, with the remainder of the keyboard shielded. Half the participants completed the paper-and-pencil format and half completed the computerised format, with participants between the groups matched for age, gender, and general intelligence (based on Berenschot Intelligence Test scores). After GATB completion, participants completed a questionnaire on the computerised format. There were significant differences in mean GATB subtest score between the two formats, with computerised format participants producing significantly higher scores on the name comparison, computation, and tool matching subtests and significantly lower scores on the three-dimensional space, vocabulary and form matching subtests. There was no significant difference in mean scores on the arithmetic reasoning subtest. The questionnaire results indicated that computer format participants found some problems with the deformation of lines and figures on the screen and only 9% disliked working with a computer. Three-quarters of computer condition participants had at least some experience with computers.

Greaud and Green (1986) compared university student performance on shortened computerised and paper-and-pencil formats of the Numerical Operations (NO) and Coding

Speed (CS) subtests of the ASVAB. These two subtests are speeded tests of clerical ability. Test format and test type were counterbalanced, however the subtests were not completed under speeded conditions, so participants continued until they finished each test and their test completion times were recorded. As completion time increased, the reliability coefficients of the computerised NC test decreased. Mean NO and mean CS scores were significantly higher for the computerised formats than for the paper-and-pencil formats, apparently because participants were significantly faster at completing the tests when they were administered on the computer.

Federico (1992) used a within-subjects design to compare computerised and paper-and-pencil formats of a semantic knowledge test on front-line Soviet platforms, such as facts on weapons systems and counterjamming procedures. Participants were Navy pilots and radar intercept officers, and both test formats were constructed from the same item database. Participants were administered both test formats, using a counterbalanced design, and there was no test-retest delay. There were no significant differences in split-half reliability or internal consistency between the two test formats and no significant differences in these reliability measures between the two test formats. There was no significant difference for the test scores between the two test formats. However, the discriminant validity for mean flight hours was higher for the computerised test than for the paper-and-pencil format.

Reardon and Loughead (1988) compared the performance of college student participants on paper-and-pencil and computerised formats of the Self-Directed Search (SDS), a career assessment test, using a counterbalanced design. Participants were randomly assigned to the testing conditions. There were three data collection sessions over a two-week period. In the first testing session participants completed either the paper-and-pencil form or the computerised form of the SDS, then completed a demographic questionnaire. In the second testing session participants completed the alternative SDS form and a Comparative SDS Rating form, which was a semantic differential instrument. The computerised SDS was completed significantly faster by participants, but there were no significantly different mean scores on each of the SDS subscales between the two forms.

There were no significant differences in Comparative Rating Form endorsements between the two SDS forms, although 86% of participants preferred the computerised form.

Kapes and Vansickle (1992) compared the paper-and-pencil and computerised formats of the Harrington-O'Shea Decision-Making System (HDS), a vocational guidance test. Participants were undergraduate students, and the study used a counterbalanced design with a two-week test-retest period. The median test-retest coefficients were significantly higher for the computerised format, although there were no significant differences in mean score between the HDS formats. There was no significant difference between mean score on first HDS administration and mean score on second administration, although the second administration of the HDS was completed significantly faster by participants.

A computer-based test requires some type of response in order for the test taker to move from one item to another (Hofer & Green, 1985). If the test taker changes their mind about their answer to one or more questions, amendments may not be allowed once the test taker has entered and verified their answer. However, some test takers do not answer tests in ascending item order and try to remember a test overview as they answer questions, and the later test item content may provide them with the answers to previously unanswered items. If the CPT designer prevents retracing, the computerised format of the test may have higher psychometric merit than the traditional format, although normative and validity data from the traditional format may not be then generalisable to the computerised format. Spray, Ackerman, Reckase, and Carlson (1989) compared three traditional paper-and-pencil Marine Corps Communication-Electronics School tests with their respective computerised formats. The results indicated that the computerised tests were equivalent to the paper-and-pencil tests. The suggested reason for this finding was the ability of participants using the computerised tests to retrace, examine previously answered items, and change answers. Studies that do not provide this flexibility to test takers (e.g. Divgi & Stoloff, 1986, cited in Spray et al., 1989) also do not demonstrate equivalence between test formats.

Lunz and Bergstrom (1994) constructed a calibrated item bank that matched the content specifications for a medical technology certification test. A computerised adaptive test (CAT) and a paper-and-pencil test were both constructed from this item bank and administered to participants as part of their certification process. Participants completed

both test formats, and order of format presentation varied between medical technology programs. There were three paper-and-pencil conditions, determined by item difficulty at the start of the test (easy, medium, or hard), and paper-and-pencil participants were randomly assigned to these conditions. There were four computer conditions, determined by level of test taker control over the test. These four conditions were: (a) skip, where participants could choose the items to answer; (b) review, where participants answered all items as they were presented but could review and change item responses after test completion; (c) defer, where participants had to answer all presented items, but could defer answering items until the end of the test; and (d) none, where participants had to answer all items as they were presented, and had no skip, review, or defer options. Again, computerised condition participants were randomly assigned to the computer subconditions. There was no significant effect of test difficulty manipulation, the computerised test formats and the paper-and-pencil formats had similar reliabilities, and participants in the skip condition had a significantly higher mean ability score than participants in the no control condition, although no other differences in mean score between the computer conditions was significant.

Thus, an overview of research on CPT indicates mixed support for the idea that test takers of a computerised test will produce the same scores or profiles as if they had completed a paper-and-pencil version of that test. One of the main problems with understanding how the computerised and paper-and-pencil versions of a test are related is that many of the studies in this area have had severe methodological flaws, the three major weaknesses being: (a) no matching of participants between test format conditions; (b) the inability of participants on the computerised tests to omit, skip, or review test items, as is possible on most paper-and-pencil tests; (c) no counterbalancing of test format administrations; and (d) no test-retest comparisons between test formats. In response to these issues, for both Study 1 and Study 2 of this thesis participants were matched between experimental conditions. Participants in both Studies were able to omit, skip, and review test items. As Study 1 was a between-subjects design, addressing the test-retest issue, counterbalancing was not used. However, for Study 2, where there was a within-subjects experimental manipulation involving input device, counterbalancing was used.

As well, existing research contains a number of problems, one of which is the lack of measurement of participant computer familiarity. The influence of this variable on test taker scores on computerised tests is essentially unknown. Even the typing ability of participants has not been measured, another variable could reasonably be expected to influence participant scores due to the use of the Qwerty keyboard in computerised aptitude tests. For these reasons, both Study 1 and Study 2 of this thesis measure participant familiarity with computers, computer numeric pads, and computer mice, and participant typing ability.

CPT Research Involving Test Anxiety Measures

Test-related anxiety has been demonstrated to decrease test taker performance (Sarason, 1984). It has been suggested (Lushene, O'Neil, & Dunn, 1974, cited in Burke & Normand, 1987) that adequate practice on the computer eliminates participant CPT anxiety, and that any anxiety typically results from specific computer test procedures, namely lack of instruction clarity and unfamiliarity with computer hardware (Hedl, O'Neil, & Hansen, 1973).

Llabre, Clements, Fitzhugh, Lancelotta, Mazzagatti, and Quinones (1987) compared the performance and test anxiety of college students on verbal reasoning items from the California Short-Form Test of Mental Maturity (CMM). All participants were enrolled in a developmental reading course. Participants were randomly assigned to the computer and paper-and-pencil CMM conditions, and received one administration of the test. Participant state test anxiety was measured before CMM administration, using a revised format of the Test Anxiety Scale (TAS-R). Participants in the computerised condition received significantly lower test scores and reported significantly higher state test anxiety than participants in the paper-and-pencil condition. However, only 23% of participants reported occasional or frequent use of a computer, thus computer familiarity may have interacted with CMM test performance and state test anxiety, although this interaction was not examined in the Llabre et al. (1987) study.

Ward, Hooper, and Hannafin (1989) randomly assigned undergraduate student participants to either a paper-and-pencil or computerised administration of a class

examination. Each participant completed the examination once. After examination administration, each participant completed a questionnaire on test anxiety and attitudes on reviewing and skipping items, with computer format participants also completing a questionnaire section on attitudes towards computerised testing. There was no significant difference in mean examination scores between the examination formats, and there was no significant difference between the two conditions on attitudes towards skipping and reviewing items. However, the computer format participants had significantly higher mean anxiety scores than the paper-and-pencil format participants, and computerised format participants had negative attitudes towards computer testing, with 75% of these participants responding that computerised tests were more difficult than paper-and-pencil tests. However, participant familiarity with computers was, again, not examined in this study.

Chin, Donn, and Conry (1991) compared the anxiety levels and science test scores of secondary school students who sat either a paper-and-pencil, multiple-choice science achievement test or a computerised format of this test. Participants were randomly assigned to the experimental conditions, with no attempt to match on demographic characteristics such as gender. The mean test score was significantly higher for the computer condition than for the paper-and-pencil condition, the score distributions on the two test formats were unequal, and the internal consistency reliability of the computerised format was lower than that of the paper-and-pencil format. There were no significant differences in test anxiety between the two groups, and this result was not modified when the computer experience of the participants was included in the analysis. However, test format varied across the two test presentation mode conditions, with the paper-and-pencil format containing two items per page and the computerised format containing only one item per screen.

In a series of two studies, Dimock and Cormier (1991) compared paper-and-pencil test performance to computerised test performance, using undergraduate students as participants. In the first study, participants were administered both parallel forms of a verbal reasoning test, one form in traditional paper-and-pencil format and the other in a format where items were presented on individual cards. A counterbalanced design was used, although this resulted in each condition containing only 5 participants. Mean card condition score was significantly lower than mean booklet condition score, and mean test completion

time was significantly shorter for the card format, although both these effects interacted with presentation order and were limited to the first test administration. The results of this first study suggest that presenting test items singularly decreases participant test scores, but this effect is counteracted by practise on test format. In the second study, participants naive to computers were administered both forms of the verbal reasoning test, one form in individual card presentation, as in the first study, and the other form was in computerised format with one question per screen presented. A counterbalanced design was used, resulting in 9 participants per condition. Participant state anxiety was also measured before and after each test administration. Mean computer condition score was significantly lower than mean card condition score, and mean test completion time was significantly shorter for the computer format, although both these effects interacted with presentation order and only held for the first test administration. The anxiety results were ambiguous.

George, Lankford, and Wilson (1992) administered the Computer Anxiety Rating Scale (CARS), the BDI, and the STAI to undergraduate students, with half the participants receiving the traditional paper-and-pencil forms and the others receiving computerised formats. Test presentation was counterbalanced within the format conditions. Mean BDI score and mean STAI state anxiety score were significantly higher for computer condition participants than for participants in the traditional format condition. BDI score was significantly positively correlated with computer anxiety level for participants in the computer condition but was not correlated with computer anxiety level for participants in the traditional format condition. These results suggest that computerised tests induce more negative affect in test takers than do their paper-and-pencil formats, and that this negative affect originates from computer anxiety. However, the computer familiarity of participants in this study was not examined.

The main problem with these studies on test anxiety is that only general measures of test anxiety were used, such as the STAI. As the research consistently suggests that administering a test using a computer format causes greater test anxiety than the paper-and-pencil format, a more sensitive measure of test anxiety is required. In other words, these studies indicate that test anxiety is greater, but do not pinpoint the area of test presentation or administration that causes this test anxiety. Thus, currently we do not know if test

anxiety arising from CPT is general or specific. For example, are participants more anxious before the computerised test administration than during the test administration? Are they more anxious at the start of the test administration proper than towards the end of the administration process? As a need for more specific measures of test anxiety is required, the questionnaires in Study 1 and Study 2 of this thesis contain anxiety questions that are directly tied to different facets of test administration, and these use a Likert scale to increase the objectivity of the anxiety measures.

Feedback and Test Taker Performance

There is little research on the issue of providing item feedback to test takers when the tests have correct answers, such as ability tests, and this issue is more pertinent now that CPT is increasing in popularity (Wise, Plake, Pozehl, Barnes, & Lukin, 1989). One potential advantage of providing feedback is to enable the test taker to monitor their performance, although a related potential disadvantage is that negative feedback would increase test taker anxiety, and thus impair test taker performance. Item feedback research has provided inconsistent findings, for example while Morris and Fulmer (1976) found that feedback reduced test taker anxiety on undergraduates sitting an exam, Rocklin and Thompson (1985) found that feedback aided undergraduates sitting an easy test, but not those students taking a difficult task. Studies finding negative effects of feedback on test taker performance have specifically found that the feedback increases participant test anxiety, increases time taken to complete the test, and decreases answer accuracy (e.g. Strang & Rust, 1973).

The study by Wise et al. (1989), using undergraduate participants and a computer administered test of algebra ability, found that item feedback and running score total increased test anxiety in participants who were given difficult items for their first 5 items, although there was no corresponding decrease in their test performance. Item feedback with no running score total appeared to have no effect on anxiety or performance. Thus, the effect of feedback on test takers remains unclear. For this reason, as well as to increase the comparability of test scores between test administrations, no feedback was provided to participants in Study 1 and Study 2 in this thesis.

Hypotheses

This thesis consists of two studies. Study 1 is based on the equivalency issues raised in Chapter 2, and compares a computerised speeded aptitude test with its traditional (paper-and-pencil) format. The hypothesis, based on the previous research comparisons on computer and paper-and-pencil test formats (refer Chapter 4), is that the computerised test will be equivalent to the traditional format. Due to the finding that test anxiety may be a function of test format, the anxiety of participants in this first study will be examined using both specific and general anxiety measures, and these anxiety measures will be compared between the two treatment groups.

Study 2 is an examination of factors that have been proposed to influence performance on computerised aptitude tests - for example, age; gender; familiarity with computers - and to determine if these hypothesised influence of these factors are modified by the actual input device used: keyboard; numeric pad; or mouse. A detailed summary on issues related to these factors is included in Chapter 3. It is proposed that the input device used will influence ACER-BL scores, dependent upon demographic characteristics, and that this will be reflected in the anxiety measures for each device. This suggestion results in two further hypotheses: (a) the familiarity of a participant with a particular input device will - through the mechanism of lowering test anxiety - increase test performance; and (b) an order effect of input devices will occur. That is, the test performance on one input device may be influenced by previous test administration (i.e. practise) on a different input device.

To increase the potential for test-taker anxiety in a situation where the outcome of the test does not affect the test-taker, for example where the test taker's results will not cause them to be rejected for a job, a speeded multiple-choice aptitude test was used. It was assumed that participants would be motivated to get the highest score possible for them on the test due to their desire to appear intelligent to the experimenter (i.e. social desirability set). A realistic testing situation was used, where participants were not allowed to talk until the completion of their testing session. Thus, the nature of the test and the administration conditions of the test were designed to increase external validity.

Chapter 5: Study 1: Reliability Comparisons Between Paper-and-Pencil Format and Computerised Format of the ACER Advanced Test BL

The main purpose of this initial study was to demonstrate that the computerised format of the ACER Advanced Test BL was as reliable as the traditional paper-and-pencil format. A second aim was to compare the anxiety of participants who completed paper-and-pencil versions of the ACER Advanced Test BL with participants who completed the computerised version.

Method

Participants

This study involved 41 Massey University internal undergraduate volunteers. Participants were recruited from undergraduate lectures in the Massey University faculties of Science, Computer Science, Social Science, and Business Studies. Assignment to each of the two treatment groups (paper-and-pencil testing versus computerised testing) was performed using a stratified random design, with age and gender matched between the two treatment groups. Nineteen participants were assigned to the paper-and-pencil group and 22 participants to the computerised group. All participants who turned up to the testing sessions completed the study, therefore participant attrition does not affect the results.

Apparatus

ACER Advanced Test BL Rationale

The ACER Advanced Test BL (ACER-BL) was the psychological test used in this present study. There were a number of reasons for selecting this particular test. Firstly, the ACER-BL is currently in use in New Zealand and Australia. A computerised format of this test could provide another administration format for current ACER-BL users. Thus, a computerised format of the ACER-BL has high practical applicability. Secondly, a test with New Zealand norms should be used to allow comparison of participant scores with norm samples containing larger numbers of test takers. This would allow the determination of

whether individual participant scores, and the score spread, were typical of the age and educational level of the average New Zealander, as shown in the ACER-BL norm supplement. The ACER-BL has New Zealand norms, including norms derived from university students only.

Thirdly, the ACER-BL has a tightly standardised administration procedure in the manual, which is relatively easily applied to a computerised testing situation. Thus the computer and paper-and-pencil administration groups were extremely similar in administration procedure. Finally, the ACER-BL contains only 30 items, each of which is multi-choice, and is a speeded test with a maximum completion time of 15 minutes. As participant boredom and fatigue were considered to be extraneous factors that could influence the results of this study, and length of test could influence these two factors, the use of a short test administration time should minimise fatigue and boredom.

Paper-and-Pencil Questionnaire Construction

Three paper-and-pencil questionnaires, sampling participant background and demographics, and participant test anxiety, were created for use in this study. The questionnaires contained only Likert scale and multiple-choice items for speed and ease of completion (Questionnaires 1.1, 1.2, and 1.3, refer Appendix B).

Questionnaire 1.1, hereafter called the Participant Characteristics Questionnaire, measured the general participant characteristics: (a) demographics, such as age, general family income; (b) educational background, such as year of undergraduate study, level of mathematical ability; and (c) computer familiarity, such as amount of computer use, typing ability. Apart from the university student identification number and age questions, where participants simply filled in their information, all other questions were restricted choice with participants circling the option that applied to them. Questionnaire 1.1 was designed to ascertain general subject characteristics that could influence test performance, especially on the computerised format of the ACER-BL. For example, general family income was included because American researchers (e.g. Hofer & Green, 1985) suggest that people from lower income families do not have as much access to a computer as people from higher incomes, and thus people from the lower incomes would be expected to perform more

poorly on a computerised test simply because of a lack of computer familiarity. However, to increase questionnaire sensitivity to detect participant differences in computer familiarity, questions on the frequency of computer use, and frequency of use of computer input hardware such as mouse and keyboard, were also included.

Questionnaire 1.2, hereafter called the Anxiety Questionnaire, sampled factors associated with the first administration of the ACER-BL for all participants. The Questionnaire concentrated on test anxiety *and* factors that could reasonably be expected to influence test anxiety. Regarding the test anxiety questions (Questions 1 to 5), these were all presented using a Likert scale format, with every point on the scale behaviourally anchored to minimise intra-item subjectivity in participant responding. Most of the scales contained the same behavioural anchors and the same number of Likert scale items to minimise inter-item subjectivity in participant responding. As participant test anxiety could change over time and across test completion, anxiety was measured for a number of specific test areas, such as pre-test administration anxiety, anxiety during test administration, and anxiety on specific test factors such as type of test question.

To increase the power of subsequent multivariate analyses, the first 5 questions of the Anxiety Questionnaire were divided into three conceptual categories: (a) general sequential anxiety; (b) specific sequential anxiety; and (c) anxiety on specific question type. Category (a) contains the first 3 items of the Anxiety Questionnaire, and was designed to tap the types of anxiety measured by general anxiety questionnaires, such as the STAI state scale. Category (b) contains item 4 of the Anxiety Questionnaire, anxiety on the 3 subsections of the ACER-BL. This second category was designed to tap changes in test anxiety as a function of test completion. Category (c) contains item 5 of the Anxiety Questionnaire, anxiety on specific ACER-BL question types such as analogy items. This third category was designed to tap item-dependent anxiety on the ACER-BL.

Regarding the factors that could influence test anxiety, one questionnaire item sampled seven test components that were either general, such as the test time limit, or specific to test format, such as anxiety related to entering an incorrect response due to unfamiliarity with a computer keyboard. This item set used a forced choice response option. One open-ended question enabled participants to nominate changes that could possibly

decrease test anxiety, and this item was included to detect those factors influencing test anxiety that were not included in the forced choice item. Finally, two questions on participant familiarity with psychological tests in general, and computerised psychological tests specifically, were also included. Test anxiety was measured retrospectively and subjectively, a standard procedure for measuring participant test anxiety.

Questionnaire 1.3, hereafter called the General Questionnaire, sampled factors associated with the second ACER-BL administration, using the same questions as Questionnaire 1.2, but with an extra 5 questions at the end of the questionnaire. These additional questions were mainly concerned with identifying the test administration upon which the participant identified the most test anxiety (tapped two ways) and also if the participant felt eyestrain had occurred during the test administrations (tapped two ways).

Computerised ACER Advanced Test BL

A computerised in-house format of the ACER-BL was created with the permission of the New Zealand copyright holders (New Zealand Council For Educational Research - NZCER). The program was written in Turbo Pascal and administered on standard IBM-compatible computers in the Massey University Department of Psychology computer laboratory. The computerised format was designed to be as similar as possible to the paper-and-pencil test in order to minimise the possibility that any differences in participant test scores on the two treatment groups were caused by differences in test layout. For example, the computerised format showed more than one question per screen, just as the paper-and-pencil format shows more than one question per page.

However, while the paper-and-pencil test is black type on white paper, colour was utilised in the computerised format of the test. The reason for this is that Turbo Pascal does not allow the use of font styles such as italics and bolding unless it is written in graphics mode, and the computerised ACER-BL was written in text mode. Text mode was used as this provided the text format normally displayed on computer screens, for example in MS-DOS. In order to maintain the similarity in text presentation between the paper-and-pencil and computerised ACER-BL tests, three The only changes in ACER-BL format between these two test modes were: (a) normal text was dark grey on a very light grey background

for the computerised format; (b) the computerised format substituted black text for bolding and brown text for italics; and (c) the computerised format substituted keyboard entry for writing.

The same three Questionnaires were also used with the participants in this treatment group.

Procedure

General Procedure

Prospective participants volunteered for this study, which was conducted outside their lecture times on campus in a computer laboratory. The experimenter contacted each prospective participant to book testing times suitable for each participant. The number of participants in each session ranged from 2 to 8 over eight testing sessions.

The paper-and-pencil group (P group) and computer group (C group) participants were tested in the same room, and most of the testing sessions involved simultaneous testing of participants in both groups. In a between-subjects design, participants were assigned to either the P group or the C group. As there is no parallel form of the ACER-BL, and as the split-half procedure for speeded tests is problematic, participants in each treatment group were administered the same test *and* test format twice.

Upon entry to the computer laboratory, each participant was provided with an information sheet and consent form. Participants were advised they would be completing a speeded aptitude test and questionnaires about the test, although at this stage participants were not informed of the actual reasons for testing, the number of ACER-BL administrations, and the number and purpose of the questionnaires they would complete. Upon completion of the consent form, participants were seated so they would not be able to see other participants' responses to the ACER-BL test. This was achieved by seating every participant in front of a computer, two participants per computer bench, with one unused computer separating each participant on the same bench. P group participants simply sat in front of a computer that was switched off, with enough bench room to comfortably complete the paper-and-pencil ACER-BL and the questionnaires. Participants then

completed the Participant Characteristics Questionnaire. This questionnaire was collected prior to starting the treatment group procedures.

In order to maximise external validity so generalisations to the test taker populations could be made, it was necessary that participants should feel they were in an actual testing situation. This realistic testing situation was created in two ways: (a) the ACER-BL manual instructions were strictly adhered to, with no talking allowed from the first test administration to the debriefing; and (b) the ACER-BL is a multi-choice aptitude test, so each participant knew that they could answer any or all items incorrectly resulting in a decreased score compared with other participants. While participants were informed their scores were confidential to the researcher, it was hoped that participants would be motivated to perform highly due to individual factors such as pride. However, while the influence of motivating factors upon participant performance was beyond the scope of this present study, and thus was not tested, this influence was held constant for both test administrations.

P Group Procedure

Paper ACER-BL test booklets were provided to the P group participants. Participants were instructed to complete the age, date, and name sections of the ACER-BL, substituting their university student identification number for their name. The oral standardised ACER-BL test-manual instructions were provided to participants. Participants were then asked to start the test, and were timed according to the standardised test instructions.

Once all participants had completed the first administration of the ACER-BL the Anxiety Questionnaire was administered. Participant completion of this questionnaire occurred within 10 minutes, so the time delay between finishing the initial ACER-BL administration and starting the second ACER-BL administration was held constant at 10 minutes for all participants. For the second ACER-BL administration, participants were, again, instructed to complete the demographic information but to ignore the practice questions. Again, the P group participants were timed. Once the participants completed the second administration they were the General Questionnaire to complete. After completing this last questionnaire, participants were debriefed and thanked.

C Group Procedure

The ACER-BL computerised test was started for participants in the C group. Participants were instructed to complete the age, date, and university student identification number sections of the aptitude test. The oral test instructions were provided to participants, with additional instructions (such as "to make a correction, simply overtype the incorrect letter in the [answer] bracket"). Participants were then asked to start the test. Participants in this treatment group were not timed by the experimenter as the computerised test contained a timer. Once the time limit was reached, the program automatically exited from the ACER-BL to a screen page that thanked participants for undertaking the test.

Once all participants had completed the first administration of the ACER-BL the Anxiety Questionnaire was administered. Again, the time delay between finishing the initial ACER-BL administration and starting the second ACER-BL administration was held constant at 10 minutes for all participants. For the second ACER-BL administration, participants were instructed to complete the demographic information but to ignore the practice questions. Once participants completed the second administration they were given the General Questionnaire to complete. After completing this last questionnaire, participants were debriefed and thanked.

Results

All results were analysed using SPSS PC Version 5.0. All variables were entered in numeric format. Repeated measures ANOVAs were performed using the SPSS MANOVA wsfactors command with multivariate analysis that was free of the univariate analysis test assumptions. Unless otherwise stated, chi-squares were performed on dichotomised variables.

General Participant Characteristics

Table 5.1 shows the demographic statistics for participants. The majority of participants were aged 20 years or younger and were of New Zealand European descent, 63% were female, and 63% came from families with an annual income of $NZD40,000 or greater.

Table 5.1. Demographics of participants, by treatment group.

Demographic Variable	P Group		C Group	
	N	%	N	%
Age				
18 to 20 years:	13	68	16	73
21 years and above:	6	32	6	27
Gender				
Male:	7	37	8	36
Female:	12	63	14	64
Ethnicity[a]				
NZ European:	17	89	20	91
Other:	2	11	2	9
General Family Income				
Up to $40,000 pa:	9	47	6	27
Over $40,000 pa:	10	53	16	73

[a]No participants identified themselves as having New Zealand Maori, Samoan, or Tongan ethnicity.

Table 5.2 shows general characteristics of participants. In order to provide sufficient participant numbers per cell for later statistical analysis, these variables were dichotomised. Sixty-eight percent of participants reported normal vision, 68% of participants had at least

bursary level mathematical ability, 54% had at least first year university statistical ability, 66% had at least bursary level English ability, and 46% had at least some programming ability. Forty-nine percent of participants were first year university undergraduates, with 39% of participants majoring in a social science.

Table 5.2. General characteristics of participants, by treatment group.

General Characteristic	P Group		C Group	
	N	%	N	%
Eyesight				
Uncorrected vision:	13	68	15	68
Corrected vision:	6	32	7	32
Mathematical ability				
Up to 6th Form:	7	37	6	27
Bursary onwards:	12	63	16	73
Statistical ability				
Up to Bursary:	9	47	10	45
First year university onwards:	10	53	12	55
English ability				
Up to 6th Form:	6	32	8	36
Bursary onwards:	13	68	14	64
Programming ability				
None reported:	11	58	11	50
Any experience:	8	42	11	50
Year at university				
First year undergraduate:	11	58	9	41
Second and third year:	8	42	13	59
Faculty Major				
Social Science:	9	47	8	36
Other than Social Science:	8	42	14	64
Not specified:	2	11		

Table 5.3 shows the computer background of participants. Again, most of the variables were dichotomised due to low participant numbers in specific cells. Fifty-four percent of participants reported touch-typing ability, 39% reported frequent numeric pad use, 63% reported frequent mouse use, and 76% reported using a computer at least once a week.

Table 5.3. Computer abilities reported by participants, by treatment group.

Computer Ability	P Group		C Group	
	N	%	N	%
Typewriter/keyboard ability				
Touch-typist:	11	58	11	50
Can't touch-type:	8	42	11	50
Numeric pad use				
Frequent:	4	21	12	55
Infrequent:	15	79	10	45
Mouse use:				
Frequent:	12	63	14	64
Infrequent:	7	37	8	36
Computer use				
At least 4 times per month:	14	74	17	77
Under 4 times per month:	5	26	5	23

The nominal variables of Gender, Ethnicity, Eyesight, and the university subject Major variable, were analysed using chi-square (refer Tables 5.1 and 5.2). Thirty-seven percent of P group participants and 36% of C group participants were male, a nonsignificant difference, $X^2(1, \underline{N} = 41) = 0.00$, $\underline{p} = .97$. Eighty-nine percent of P group participants and 91% of C group participants were New Zealand Europeans, a nonsignificant difference, $X^2(1, \underline{N} = 41) = 0.02$, $\underline{p} = .88$. Sixty-eight percent of participants in both the P and C groups reported normal vision, a nonsignificant difference, $X^2(1, \underline{N} = 41) = 0.00$, $\underline{p} = .99$. Forty-seven percent of P group participants and 36% of C group participants had a social science major, a nonsignificant difference, $X^2(1, \underline{N} = 39) = 1.07$, $\underline{p} = .30$ All other participant characteristics, such as mathematical ability and computer familiarity, were sampled using interval scales, and independent t-tests were conducted to determine between-group differences. The results of the t-test analyses are summarised in Table 5.4, showing no significant differences between the two experimental groups on these 11 participant characteristics.

Table 5.4. Characteristics reported by participants, by treatment group. Results of t-test analyses.

Characteristic		P Group	C Group	t-value	P (2-tailed)
Age	M	21.26	20.41	0.74	.47
	SD	4.47	2.87		
Family Income	M	2.47	2.64	0.82	.42
	SD	0.61	0.66		
Mathematical ability	M	3.58	3.91	0.91	.37
	SD	1.17	1.15		
Statistical ability	M	4.16	3.95	0.40	.69
	SD	1.34	1.81		
English ability	M	3.89	3.77	0.43	.67
	SD	0.88	0.92		
Programming ability	M	1.26	1.64	0.57	.57
	SD	2.02	2.13		
Year at university	M	1.58	1.86	1.13	.27
	SD	0.77	0.83		
Typing ability	M	3.16	3.27	0.26	.80
	SD	1.57	1.24		
Numeric pad familiarity	M	2.21	1.82	1.26	.21
	SD	0.98	1.01		
Mouse familiarity	M	1.47	1.41	0.30	.76
	SD	0.77	0.59		
Computer familiarity[a]	M	2.53	2.64	0.20	.85
	SD	1.74	1.81		

[a]This analysis was performed after recoding the options for this question to produce an interval scale.

Comparison of Participant Results with ACER-BL New Zealand Norms

The NZCER sampled students from Form 7, Teachers College, and University to provide the national norms for the New Zealand revision of the ACER-BL (Reid, Croft, Gilmore, & Philips, 1986). There were 1083 participants in their university sample, of which 781 were female, covering 10 university departments. The majority of their participants were first and second year undergraduates studying education and psychology papers. As shown in Table 5.5, P and C group scores for the first administration of the ACER-BL were similar to the university norm sample, with higher mean scores, lower standard deviations, and lower standard errors of measurement.

Table 5.5. Comparison of NZCER university students norm sample scores with the first ACER-BL administration scores for P and C group participants.

Statistic	NZCER Sample	P Group	C Group
Mean	19.87	20.87	20.23
SD	4.62	3.79	4.13
SE of measurement	2.5	1.0	1.5

ACER Advanced Test BL Internal Consistency

The internal consistency of each ACER-BL administration for this present study was analysed for the P and C treatment groups using coefficient alpha, and the results are presented in Table 5.6. The NZCER reported a KR21 reliability of .71 for their university norm sample (Reid et al., 1986).

Table 5.6. Internal consistency reliabilities for each ACER-BL administration, by treatment group.

Test administration	P Group	C Group
Test administration #1		
Alpha:	.73	.80
Number of items with zero variance:	3	5
Test administration #2		
Alpha:	.78	.70
Number of items with zero variance:	1	5

The first ACER-BL administration for C group participants had the highest internal consistency, however, the second ACER-BL administration for the C group participants had the lowest internal consistency.

Independent t-tests were conducted to determine P and C treatment group differences in test score for each ACER-BL administration in this present study. There was no significant difference between the P and C group mean test scores for the first ACER-BL administration, 20.87 and 20.23 respectively, t (39) = 0.51, p = .61. There was also no significant difference between the P and C group mean test scores for the second ACER-BL administration, 21.42 and 22.45 respectively, t (39) = 0.90, p = .37. Thus, the changing reliabilities shown in Table 5.6 was not reflected in these mean test score comparisons between the P and C groups.

ACER Advanced Test BL Test-Retest Reliability

The test-retest reliability of scores between each ACER-BL administration was analysed for each treatment group. As noted in the Method section, the test-retest interval was 10 minutes. The test-retest reliability for the P group was .93, and .87 for the C group. Reid et al. (1986) do not report a test-retest reliability for their norm samples.

Paired t-tests were conducted to determine test-retest score changes for the P and the C groups. The P group mean test score increased from 20.87 to 21.42, a nonsignificant difference, t (18) = 1.62, p = .12. However, the C group mean test score increased from 20.23 to 22.45, a significant increase, t (21) = 4.95, $p < .001$.

Subsection Analysis of ACER Advanced Test BL Scores

The 30 questions of the ACER-BL test were categorised into 3 subsections: (a) the first 10 questions of the test, representing the questions on the first test booklet page; (b) the second 9 questions, representing the questions on the second test page; and (c) the last 11 questions, representing the questions on the last test page. This categorisation was performed because a practice effect might occur with repeated testing. Any practice effect could result in an increased test score for the second ACER-BL administration. One outcome would be that participants who were unable to complete the speeded test on the first administration would complete most or all of the test in the second administration. This practice effect would be reflected in increased participant scores, especially for subsection (c) if participants used a sequential answering strategy. An analysis of participant scores for each subsection by ACER-BL administration and by treatment group was performed (Table 5.7).

Table 5.7. Mean correct items by subsection for each ACER-BL administration, by treatment group.

Test administration	P Group Mean Score	C Group Mean Score
Test #1		
Subsection (a)	8.61	9.07
Subsection (b)	6.71	6.66
Subsection (c)	5.55	4.50
Test #2		
Subsection (a)	8.50	9.05
Subsection (b)	6.45	7.16
Subsection (c)	6.47	6.25

Table 5.7 shows that participant scores tended to drop across ACER-BL subsections for both treatment groups on both ACER-BL administrations. Collapsing these subsection means across treatment group and ACER-BL administration, the means for the three ACER-BL subsections were 8.82, 6.76, and 5.67 respectively, a significant decease in subsection scores, F (2, 39) = 89.39, p <.001. This result is consistent with participants sequentially completing the ACER-BL and running short of time.

There was a significant interaction for Subsection Score by ACER-BL Administration for the P group (see Table 5.7), F (2, 17) = 3.67, p < .05. Participant scores did not drop as much across the test subsections on the second administration, and a paired t-test showed that participants scored significantly higher on the second ACER-BL administration compared with the first administration on subsection (c), t (18) = -2.72, p = .01.

A significant interaction was also found for Subsection Score by ACER-BL Administration for the C group, F (2, 20) = 15.94, p < .001. . As for the P group, participant scores in the C group did not drop as much across the test subsections on the second administration. Significantly higher second ACER-BL mean scores occurred for both subsection (b), t (21) = 2.53, p = 02, and subsection (c), t (21) = 6.03, p < .01. These results suggest that many participants in both the P and C groups used a simple sequential answering strategy, rather than targeting questions of a particular type. Visual analysis of the elapsed time for each item shows that most C group participants used this type of answering strategy.

Using the means in Table 5.7, a repeated measures analysis determined that there was no main effect for Treatment Group, F (1, 39) = 0.03, p = .87, and no significant interaction between Treatment Group and Subsection, F (2, 38) = 2.81, p = .07. That is, there were no significant differences in subsection scores between the P and C groups for either ACER-BL administration.

Effect of Treatment Group by ACER-BL Administration Interactions on Total Test Score

For participant mean test score on the first and second ACER-BL Administrations there was a significant interaction with ACER-BL Format, $F(1, 39) = 8.36$, $p < .01$. As Figure 5.1 shows, there was improvement in mean test score for both format groups, and this improvement was greater for the C group.

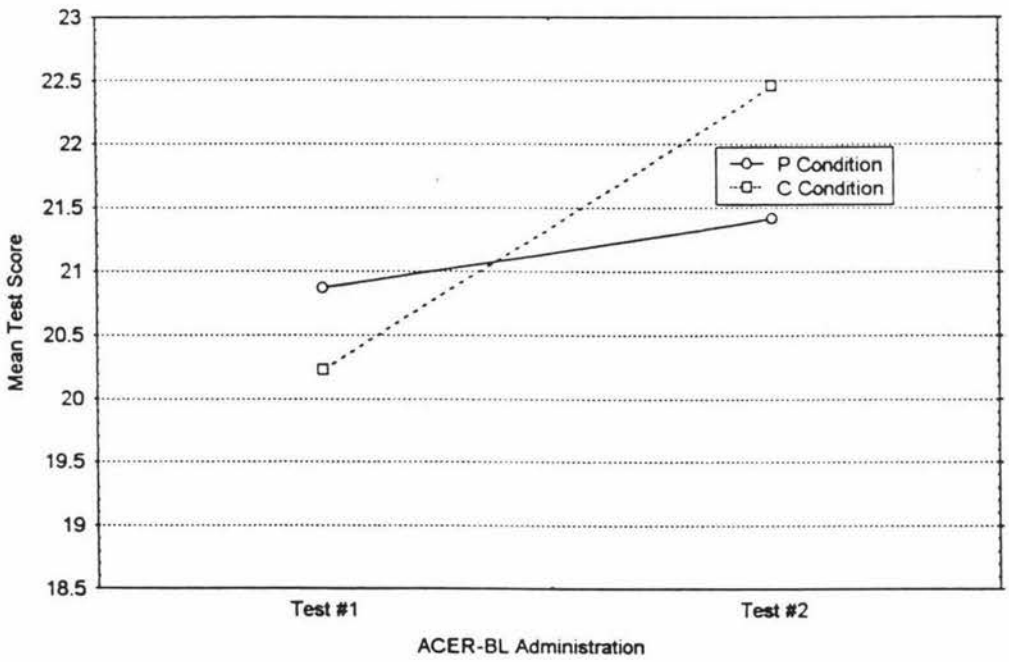

Figure 5.1. Mean test score for each ACER-BL administration and treatment group.

Effect of Participant Characteristics Interactions on Total Test Score

For participant mean score on the first and second ACER-BL administrations there was a significant interaction between ACER-BL Administration and Gender, $\underline{F}$ (1, 39) = 4.20, $\underline{p} < .05$. As Figure 5.2 shows, there was improvement in mean test score for both genders and this improvement was greater for males. Although the three-way interaction between Practise, Gender, and ACER-BL Format failed to reach significance, $\underline{F}$ (1, 37) = 0.59, $\underline{p} = .45$, Figure 5.3 shows that C group males had a lower mean score on the first ACER-BL administration compared to C group females, an opposite result to that predicted in the CPT literature.

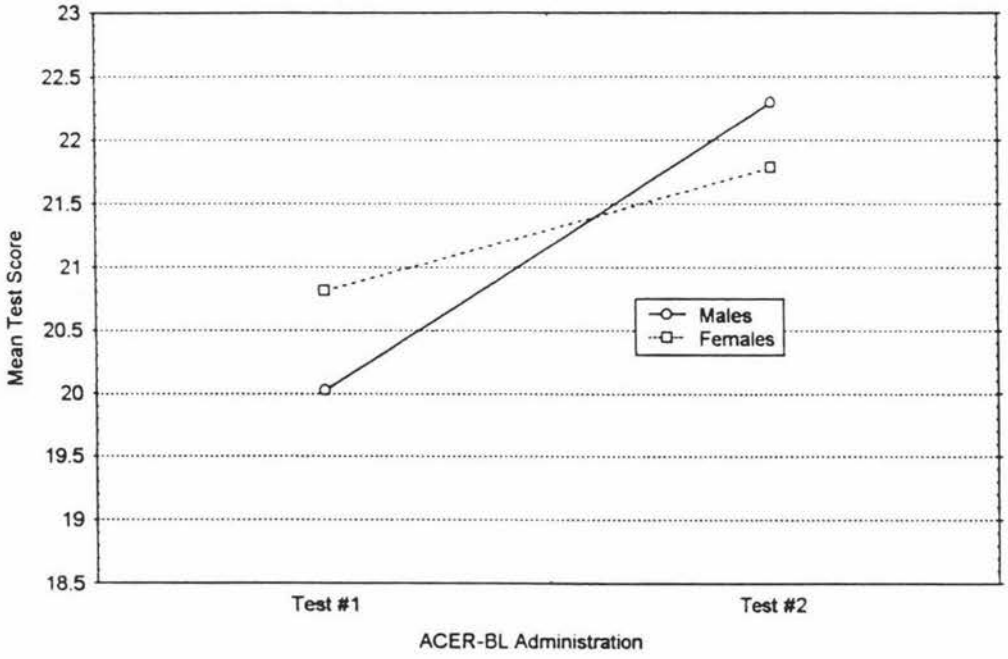

Figure 5.2. Mean test score for each ACER-BL administration, by gender.

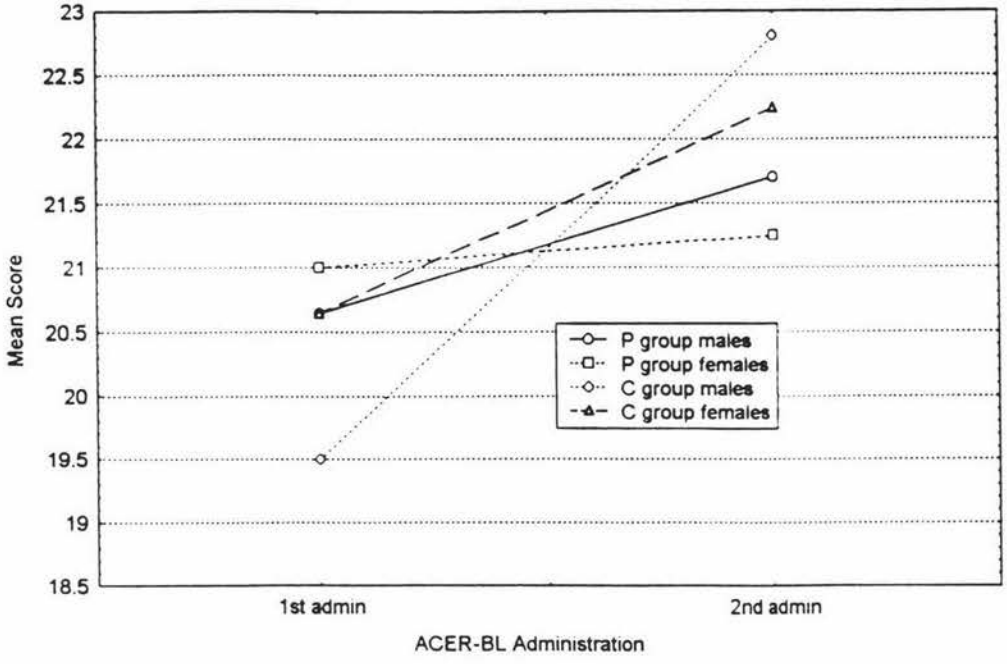

Figure 5.3. Mean test score for each ACER-BL administration, by gender and treatment group.

There was a significant interaction between Practise and Undergraduate Year, $\underline{F}$ (1, 39) = 4.83, $\underline{p} < .05$. As shown in Figure 5.4, there was improvement in mean test score for both undergraduate divisions, with first year undergraduates achieving the highest mean score for both ACER-BL administrations and showing the lowest practise effect. The three-way interaction between Practise, Undergraduate Year, and ACER-BL Format failed to reach significance, $\underline{F}$ (1, 37) = 2.49, $\underline{p} = .12$.

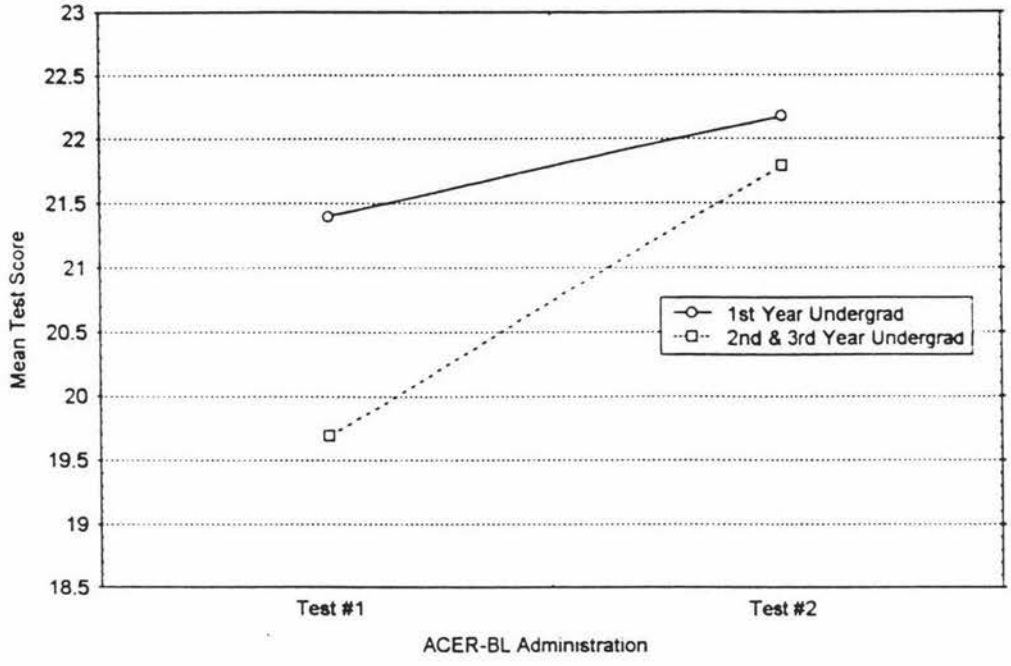

Figure 5.4. Mean test score for each ACER-BL administration, by year of undergraduate study.

There was a significant interaction between Practise and Typing Ability, $\underline{F}$ (1, 39) = 5.19, $\underline{p} < .05$. As Figure 5.5 shows, there was improvement in mean test score for both typing groups, and this improvement was greater for participants who could not touch type. The three-way interaction between Practise, Typing ability, and ACER-BL Format failed to reach significance, $\underline{F}$ (1, 37) = 2.31, $\underline{p} = .14$. Figure 5.6 shows that C group participants who could not type had a lower mean score on the first ACER-BL administration compared to C group typists. No other participant characteristics influenced mean test score, although the trends are shown in Figures C1 to C10 (refer Appendix C). Due to the high number of participant subject majors, no analysis by major was performed.

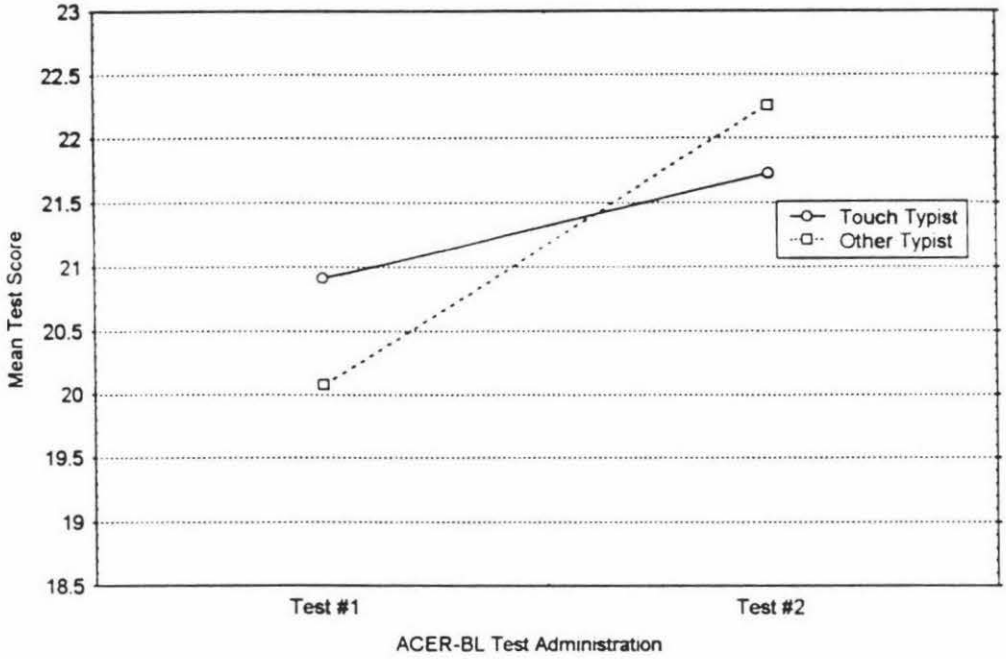

Figure 5.5. Mean test score for each ACER-BL administration, by typist grouping.

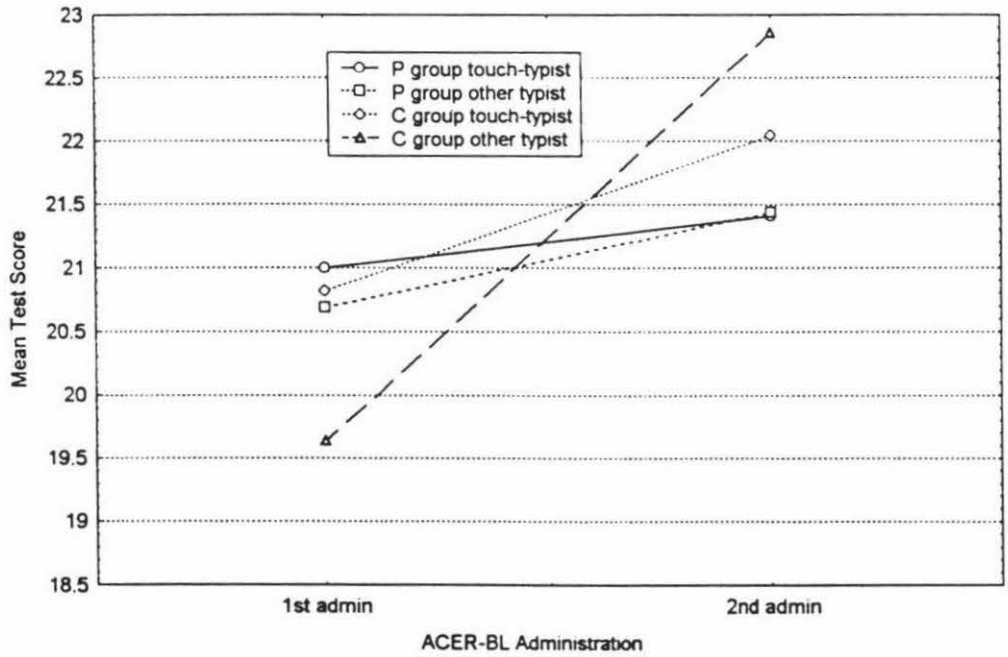

Figure 5.6. Mean test score for each ACER-BL administration, by typing ability and treatment group.

Test score on the first ACER-BL administration was analysed as a function of previous experience with similar questionnaires (Question 10 of the Anxiety and General

Questionnaires, refer Appendix B). There was no significant difference in ACER-BL mean score between participants with experience on similar tests and participants without this experience, t (39) = 0.43, p = .67.

Chi-square analyses were conducted to compare the two treatment groups on Questions 15 and 16 of the General Questionnaire (Appendix B). Eighty-nine percent of P group participants and 86% of C group participants reported no eyestrain during ACER-BL administration, a nonsignificant difference, $X^2(1, N = 41) = 0.09$, p = .76. Seventy-nine percent of P group participants and 86% of C group participants reported no eyestrain for either ACER-BL administration, a nonsignificant difference, $X^2(2, N = 41) = 4.07$, p = .13.

Test Anxiety Analyses

A repeated measures ANOVA design was used to determine the changes in participant test anxiety for the first 5 questions of the test anxiety measures (refer Appendix B, Anxiety Questionnaire and General Questionnaire). For general ACER-BL test anxiety, represented by Questions 1 to 3, there was no significant three-way interaction between General Anxiety, ACER-BL Administration, and ACER-BL Format, F (2, 38) = 0.58, p = .57. There was a significant interaction between General Anxiety and ACER-BL Administration, F (2, 38) = 6.06, p = .005, and a significant main effect for General Anxiety, F (2, 38) = 9.05, p = .001. As Figure 5.5 shows, anxiety dropped once participants had completed the ACER-BL practice questions, and increased during ACER-BL administration, and these types of anxiety were lowest on the second ACER-BL administration.

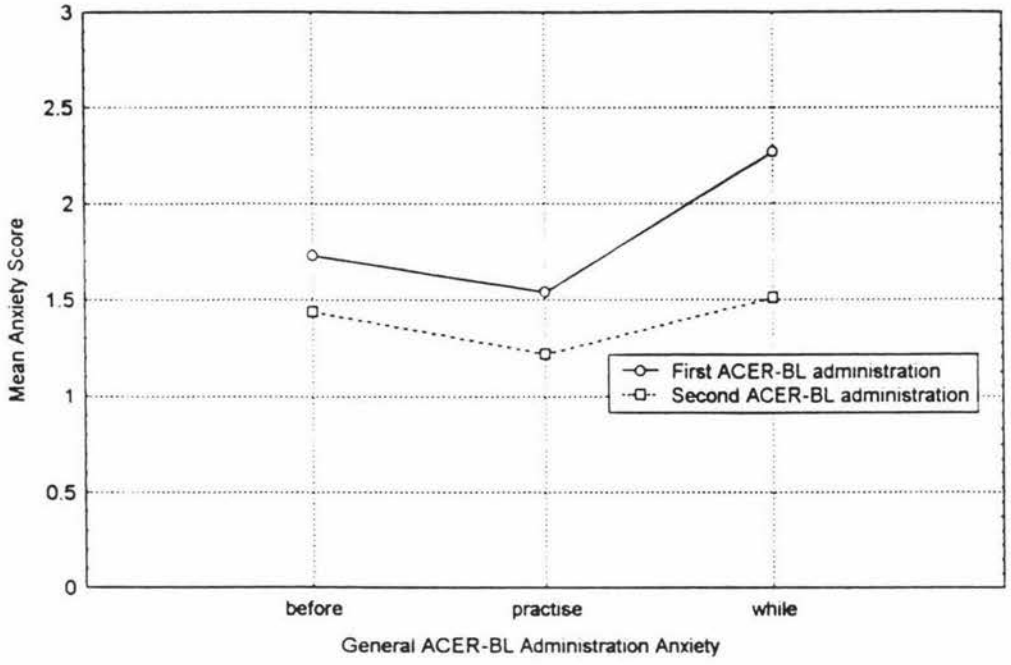

Figure 5.7. Mean general anxiety scores by ACER-BL administration, scores for treatment groups combined.

Regarding ACER-BL subsection anxiety (Question 4), there was no significant three-way interaction between Subsection Anxiety, ACER-BL Administration, and ACER-BL Format, F (2, 36) = 0.87, p = .43. There was a significant interaction between Subsection Anxiety and ACER-BL Administration, F (2, 36) = 3.69, p < .04, and a significant main effect for Subsection Anxiety, F (2, 36) = 20.83, p < .001. As shown in Figure 5.6, test anxiety increased over the subsections, with the second ACER-BL administration having the smaller increase in anxiety over subsections and the lower mean subsection anxiety scores.

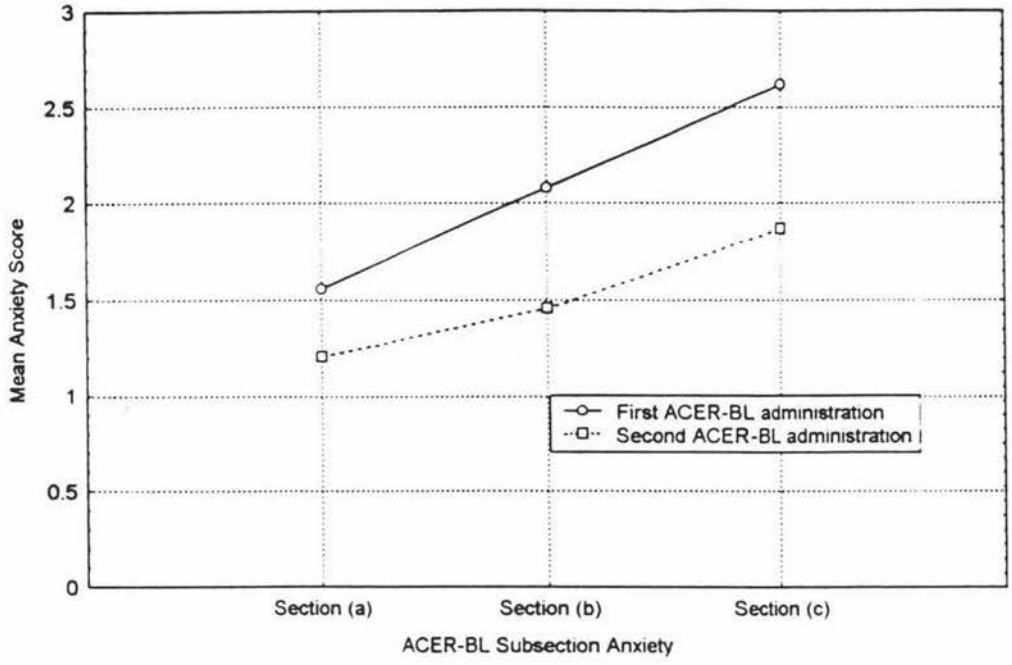

Figure 5.8. Mean subsection anxiety scores by ACER-BL administration, scores for treatment groups combined.

Regarding ACER-BL question type anxiety (Question 5), there was a significant three-way interaction between Question Type Anxiety, ACER-BL Administration, and ACER-BL Format, F (3, 34) = 3.82, $p < .02$ (refer Figure 5.7). There was no significant interaction between Question Type Anxiety and ACER-BL Format, F (3, 34) = 1.50, p = .23, no significant interaction between Question Type Anxiety and ACER-BL Administration, F (3, 34) = 0.86, p = .47, and no main effect for Question Type Anxiety, F (3, 34) = 1.30, p = .29.

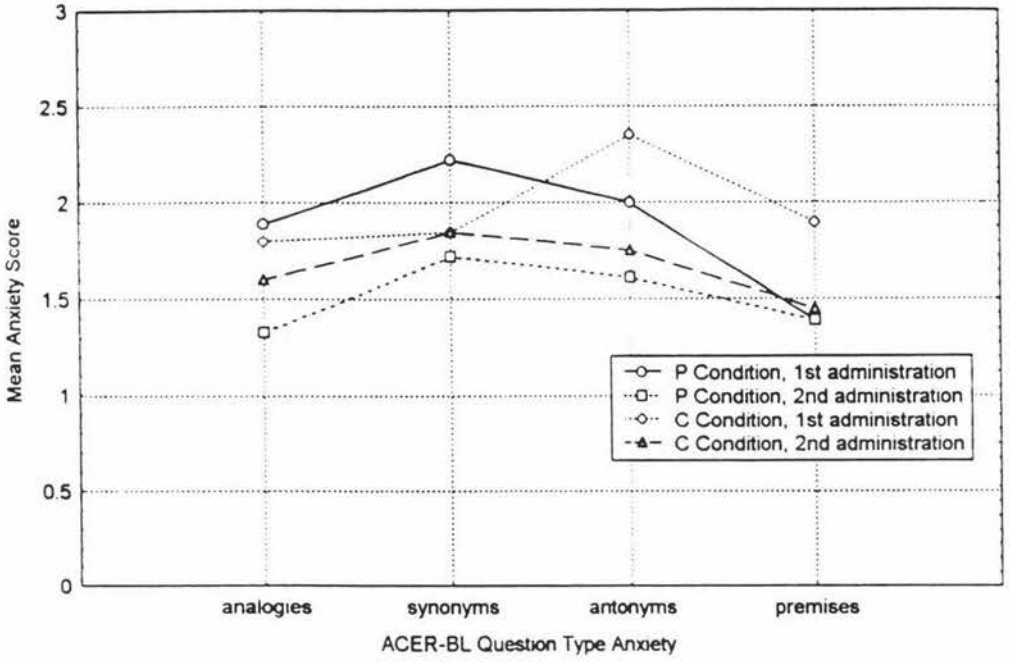

Figure 5.9. Mean question type anxiety scores by ACER-BL administration and treatment group.

There was no significant difference between the treatment groups on participants' perceptions of how ACER-BL format influenced their test anxiety for the first ACER-BL administration, t (39) = 1.10, p = . 28, or for the second administration, t (39) = 0.41, p = .68. For both ACER-BL administrations, participants in both treatment groups reported a mean indicating "no influence" of test format.

Chi-square analyses were conducted to compare the two treatment groups on Questions 13 and 14 of the General Questionnaire (Appendix B). Eighty-four percent of P group participants and 64% of C group participants reported they felt most anxious on the first ACER-BL administration, a nonsignificant difference, $X^2(2, N = 41) = 2.26$, p = .32. Eighty-four percent of P group participants and 64% of C group participants reported they felt least anxious on the second ACER-BL administration, a nonsignificant difference, $X^2(2, N = 41) = 2.26$, p = .32.

Results Summary and Brief Conclusions

The P and C group mean scores and standard deviations were compared to the NZCER university norm sample to determine the generalisability of results to the general New Zealand university population. As shown in Table 5.5, the mean test scores for P and C group participants was only slightly higher than the respective NZCER norm sample, while the standard deviations and standard errors of measurement were lower than those reported for the university norm sample. These comparisons suggest that the participants in this present study were comparable to the general New Zealand university student population.

The internal consistency of the computerised format was higher than the paper-and-pencil format for the first ACER-BL administration but lower for the second ACER-BL administration (Table 5.6). However, the lowest internal consistency measure was .70, suggesting that both test formats have moderate to high internal consistency. The NZCER reported a KR21 internal consistency reliability of .71 for their New Zealand university norm sample (Reid et al., 1986). While the internal consistency reliability in Study 1 used the coefficient alpha measure rather than KR21, these two internal consistency measures provide essentially the same result (Cronbach, 1990). The test-retest reliability of the computerised format was lower than that of the paper-and-pencil format, although both ACER-BL formats demonstrated high reliability. No test-retest reliability was reported by Reid et al. (1986) for their New Zealand norm samples.

There were no significant differences in test score between the P and C groups, as demonstrated by analyses of the total ACER-BL scores and the ACER-BL subsection scores. These results suggest that the lower reliability for the computerised format had little practical effect on participant test performance, although the reason for this lower reliability is not clear. These results also suggest that the performance of university students on aptitude tests is not influenced by the format of the test itself, whether paper-and-pencil or computer. This conclusion is supported by the fact that participants did not significantly differ between the two treatment groups on any measured characteristics, such as gender and computer familiarity, and thus the two subject populations were matched.

A practise effect on the ACER-BL occurred for both the P and C groups, with the C group showing a greater practise effect. One possible explanation for this could be that participants had more experience with paper-and-pencil format tests than with computerised formats. However, the practise effect was not due to differential test practise, as previous test experience, measured by Question 10 of the Anxiety Questionnaire (refer Appendix B), did not explain the difference. Three participant characteristics also interacted with ACER-BL administration scores; Gender, Undergraduate Year, and Typing Ability (Figures 5.2, 5.4, and 5.5). There was no significant three-way interaction effect for any of these variables when treatment group was also entered (Figures 5.3 and 5.6), although this may be due to lack of power resulting from small participant numbers in each treatment group.

There was no main effect of treatment group for either the general or subsection ACER-BL test anxieties, suggesting that a computerised aptitude test does not lead to higher temporal anxiety than that encountered on paper-and-pencil aptitude tests (refer Figures 5.7 and 5.8). These figures indicate that the mean scores for temporal ACER-BL anxiety remained at the "slight" level, where participants felt their test performance was not affected by their test anxiety, for all 5 measures. These results were reflected in the finding that participants in both treatment groups reported "no influence" of test format on their test anxiety. While there was a three-way interaction between ACER-BL question type anxiety, treatment group, and ACER-BL administration, again the mean anxiety scores represented "slight" anxiety that was not perceived to influence ACER-BL performance. Regardless of treatment group, the majority of participants reported their highest test anxiety on the first ACER-BL administration, and the lowest anxiety on the second administration.

Chapter 6: Study 2: Comparisons Between Input Device for the Computerised Format of the ACER Advanced Test BL

The main purpose of this second study was to determine if input device - keyboard (K), numeric pad (N), or mouse (M) - influenced ACER Advanced Test BL score. The two other aims were to compare the anxiety of participants, and test completion time, across these 3 input conditions.

Method

Participants

This study involved 93 Massey University internal undergraduate volunteers. Participants were recruited from undergraduate lectures in the Massey University faculties of Science, Computer Science, Social Science, and Business Studies. Assignment to each of the six treatment groups was performed using a stratified random design, with age and gender matched between testing groups. The number of participants in each treatment group, defined by the order of input device administration, was as follows: (a) KNM group, 16 participants; (b) KMN group, 14 participants; (c) NKM group, 14 participants; (d) NMK group, 15 participants; (e) MKN group, 16 participants; and (f) MNK group, 15 participants. Two participants completed the study but did not receive all input administrations due to batch file failure, and one participant did not complete the first input device administration due to computer failure, therefore the 3% participant attrition rate does not affect the results.

Apparatus

Paper-and-Pencil Questionnaire Construction

To increase the equivalence of the participant test anxiety measures between Study 1 and this present study, the Questionnaires from Study 1 (refer Appendix B) were used with minor modifications. The four Questionnaires used in this present study are contained in Appendix D. Questionnaire 2.1, hereafter called the Participant Characteristics

Questionnaire, is identical to the Participant Characteristics Questionnaire used in Study 1. Questionnaires 2.2 and 2.3, hereafter called the First Anxiety and Second Anxiety Questionnaires respectively, are identical to the Anxiety Questionnaire in Study 1, apart from Question 6 which was modified to tap the input format used as opposed to the administration format. Questionnaire 2.4, hereafter called the General Questionnaire, is identical to the General Questionnaire used in Study 1, apart from modification to Questions 6, 12, 13, 14, 15, 16, and the inclusion of Question 17 so that eyestrain was tapped three ways instead of two.

Computerised ACER Advanced Test BL

To maximise the test score equivalency between Study 1 and this present study, the same Turbo Pascal ACER-BL program was used for the keyboard input phase. For the numeric pad and mouse input programs, the only modifications made to the Turbo Pascal source code were the enabling of the appropriate input device and disabling of other input devices and, for the numeric pad version, the multichoice options were given as numeric choices rather than alphabetical choices. Again, the compiled programs were administered on standard IBM-compatible computers in the Massey University Department of Psychology computer laboratory.

As participants were required to complete all ACER-BL computerised versions, MS-DOS batch files were written so that the version sequence was administered automatically to all participants, thus eradicating administration sequence error. There was one batch file for each version sequence, so a total of 6 batch files were used.

Procedure

General Procedure

Prospective participants volunteered for this study, which was conducted outside their lecture times in the computer laboratory. The experimenter contacted each prospective participant to book testing times suitable for each participant. The number of participants completing the study, in each session, ranged from 2 to 11 over thirteen testing sessions.

All participants were tested in the same room, and most of the testing sessions involved simultaneous testing of participants in different groups. In a between-subjects design, participants were assigned to one of the six treatment groups, based on the sequence in which the input devices were used: (a) KNM group; (b) KMN group; (c) NKM group; (d) NMK group; (e) MKN group; and (f) MNK group. As there is no parallel form of the ACER-BL, and as the split-half procedure for speeded tests is problematic, participants in each group were administered the same test three times.

Upon entry to the computer laboratory, each participant was provided with an information sheet and consent form. Participants were advised they would be completing a speeded aptitude test and questionnaires about the test, although at this stage participants were not informed of the actual reasons for testing, the number of ACER-BL administrations, and the number and purpose of the questionnaires they would complete. Upon completion of the consent form, participants were seated so they would not be able to see other participants' responses to the ACER-BL test. This was achieved by seating every participant in front of a computer, two participants per computer bench, with one unused computer separating each participant on the same bench. Participants then completed the Participant Characteristics Questionnaire. This questionnaire was collected before starting the treatment group procedures.

The procedure then followed the C Group procedure of Study 1, apart from the following amendments. First, once the first administration of the ACER-BL was completed, the First Anxiety Questionnaire was administered. Second, once the second administration of the ACER-BL was completed, the Second Anxiety Questionnaire was administered. Third, once the final administration of the ACER-BL was completed, the General Questionnaire was administered. After completing this last questionnaire, participants were debriefed and thanked. Participants completed each Anxiety Questionnaire within 10 minutes, so the time delays between the first and second, and second and third, ACER-BL administrations were held constant at 10 minutes for all participants.

Results

All results were analysed using SPSS PC Version 5.0. All variables were entered in numeric format. Repeated measures ANOVAs were performed using the SPSS MANOVA wsfactors command with multivariate analysis that was free of the univariate analysis test assumptions. Unless otherwise stated, chi-squares were performed on dichotomised variables.

General Participant Characteristics

Table 6.1 shows the demographic statistics for participants. As in Study 1, the majority of participants were aged 20 years or younger and were of New Zealand European descent, and 63% came from families with an annual income of $NZD40,000 or greater. Fifty-three percent of participants were female.

Table 6.1. Demographics of participants, by treatment group.

Demographic Variable	KNM group		KMN group		NKM group		NMK group		MKN group		MNK group	
	N	%	N	%	N	%	N	%	N	%	N	%
Age												
18 to 20 years:	11	69	10	71	10	71	10	67	10	63	10	67
21 years and above:	5	31	4	29	4	29	5	33	6	38	5	33
Gender												
Male:	8	50	6	43	7	50	6	40	7	44	7	47
Female:	8	50	7	50	7	50	9	60	9	56	8	53
Missing:			1	7								
Ethnicity												
NZ European:	14	88	12	86	12	86	15	100	13	81	13	87
Other:	2	13	2	14	2	14			2	13	2	13
Missing::									1	6		
General Family Income												
Up to $40,000 pa:	7	44	4	29	7	50	6	40	4	25	5	33
Over $40,000 pa:	9	56	10	71	7	50	9	60	12	75	10	67

Table 6.2 shows general characteristics of participants. In order to provide sufficient participant numbers per cell for statistical analysis, these variables were dichotomised. Sixty-eight percent of participants reported normal vision, 56% of participants had at least

bursary level mathematical ability, 39% had at least first year university statistics ability, 58% had at least bursary level English ability, and 40% had at least some programming ability. Fifty-six percent of participants were first year university undergraduates, with 34% majoring in a social science.

Table 6.2. General characteristics of participants, by treatment group.

General Characteristic	KNM group		KMN group		NKM group		NMK group		MKN group		MNK group	
	N	%	N	%	N	%	N	%	N	%	N	%
Eyesight												
Uncorrected vision:	11	69	10	71	7	50	11	73	9	56	13	87
Correct vision:	5	31	4	29	7	50	4	27	7	44	2	13
Mathematical ability												
Up to 6th Form:	8	50	7	50	7	50	5	33	6	38	6	40
Bursary onwards:	8	50	7	50	6	43	10	67	10	63	9	60
Missing:					1	7						
Statistical ability												
Up to Bursary:	12	75	11	79	9	64	7	47	8	50	7	47
First year university onwards:	4	25	3	21	4	29	8	53	8	50	8	53
Missing:					1	7						
English ability												
Up to 6th Form:	7	44	3	21	6	43	7	47	9	56	5	33
Bursary onwards:	9	56	11	79	7	50	8	53	7	44	10	67
Missing:					1	7						
Programming ability												
None reported:	10	63	8	57	10	71	9	60	8	50	9	60
Any experience:	6	38	6	43	4	29	6	40	8	50	6	40
Year at university												
First year undergraduate:	10	63	10	71	6	43	7	47	7	44	10	67
Second and third year:	6	38	4	29	8	57	8	53	9	56	5	33
Faculty Major												
Social Science:	6	38	4	29	6	43	4	27	7	44	4	27
Other than Social Science:	10	63	8	57	8	57	11	73	9	56	11	73
Missing:			2	14								

Table 6.3 shows the computer background of participants. Again, the variables were dichotomised due to low participant numbers in specific cells. Forty-eight percent of participants reported touch-typing ability, 28% reported frequent numeric pad use, 60% reported frequent mouse use, and 72% reported using a computer at least once a week.

Table 6.3. Computer abilities reported by participants, by treatment group.

Computer Ability	KNM group		KMN group		NKM group		NMK group		MKN group		MNK group	
	N	%	N	%	N	%	N	%	N	%	N	%
Typewriter/keyboard ability												
Touch-typist:	8	50	7	50	6	43	7	47	9	56	6	40
Can't touch-type:	8	50	7	50	8	57	8	53	7	44	9	60
Numeric pad use												
Frequent:	4	25	5	36	3	21	4	27	6	38	3	20
Infrequent:	12	75	9	64	11	79	11	73	10	63	12	80
Mouse use												
Frequent:	12	75	5	36	9	64	9	60	9	56	10	67
Infrequent:	4	25	9	64	5	36	6	40	7	44	5	33
Computer use												
At least 4 times per month:	11	69	11	79	11	79	12	80	10	63	10	67
Under 4 times per month:	5	31	3	21	3	21	3	20	6	38	5	33

The nominal level variables of Gender, Ethnicity, Eyesight, and the university subject Major variable, were analysed using chi-square. The percentage of male participants in each treatment group ranged from 40% to 50%, a nonsignificant gender difference, $\mathbf{X}^2(5, \underline{N} = 89) = 0.45$, $\underline{p} = .99$. The percentage of New Zealand European participants in each treatment group ranged from 86% to 100%, a nonsignificant ethnicity difference, $\mathbf{X}^2(5, \underline{N} = 89) = 2.32$, $\underline{p} = .80$. The percentage of participants with normal vision in each treatment group ranged from 50% to 87%, a nonsignificant eyesight difference, $\mathbf{X}^2(5, \underline{N} = 90) = 5.75$, $\underline{p} = .33$. The percentage of participants with a social science major ranged from 27% to 44%, a nonsignificant difference, $\mathbf{X}^2(5, \underline{N} = 88) = 1.89$, $\underline{p} = .86$. All other participant characteristics, such as mathematical ability and computer familiarity, were sampled using interval scales, and two-way ANOVAs were conducted to determine between-group differences. The results of the ANOVAs are summarised in Table 6.4, showing no significant differences between the six experimental groups on these 11 participant characteristics.

Table 6.4. Mean characteristics reported by participants, by treatment group. Results of ANOVA analyses.

Characteristic	KNM group	KMN group	NKM group	NMK group	MKN group	MNK group	F-value	p
Age	21.63	20.93	21.77	20.93	22.31	20.60	0.23	.95
Family Income	2.44	2.57	2.38	2.40	2.56	2.47	0.16	.98
Mathematical ability	3.44	3.64	3.23	4.00	3.94	3.80	0.55	.74
Statistical ability	3.00	3.57	2.77	4.33	4.25	4.13	1.74	.14
English ability	3.81	3.79	3.46	3.60	3.50	3.73	0.25	.94
Programming ability	1.44	1.71	1.43	2.13	1.63	1.20	0.25	.94
Year at university	1.69	1.36	1.79	1.80	1.75	1.33	1.15	.34
Typing ability	3.25	3.64	3.50	3.27	3.31	3.73	0.36	.88
Numeric pad familiarity	2.31	2.00	2.00	2.40	2.13	2.73	1.01	.42
Mouse familiarity	1.31	1.79	1.36	1.47	1.63	1.47	1.00	.42
Computer familiarity[a]	2.81	2.71	2.50	2.27	2.88	2.87	0.28	.92

[a]This analysis was performed after recoding the options for this question to produce an interval scale.

ACER-BL Test Score Analyses

There was no significant interaction between Treatment Group and ACER-BL Administration for mean ACER-BL scores, F (10, 168) = 0.39, p = .95 (refer Figure 6.1), although there was a significant main effect for Administration, F (2, 83) = 33.09, $p < .001$. This pattern of results also occurred for test completion time, with no significant interaction between Treatment Group and Administration for mean ACER-BL scores, F (10, 168) = 0.63, p = .78, and a significant main effect for Administration, F (2, 83) = 804.08, $p < .001$.

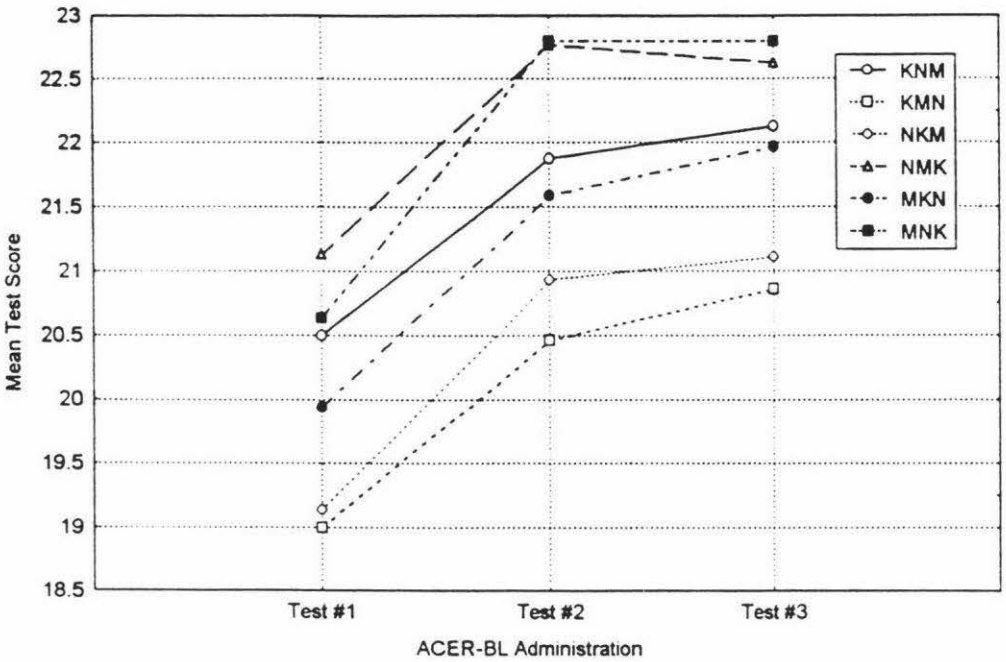

Figure 6.1. Mean test score for each ACER-BL administration, by treatment group.

Figure 6.2 shows the mean test scores for each ACER-BL administration by input device. There was a significant difference in keyboard mean scores across the three ACER-BL administrations, F (2, 87) = 5.62, p = .005. This was due to the significant increase in ACER-BL scores between the first and third ACER-BL administrations, t (58) = 3.27, p = .002. There was a significant increase in ACER-BL scores between the first and second ACER-BL administrations for the numeric pad, t (58) = 2.24, p = .03. There was no significant difference in mouse mean scores across the three ACER-BL administrations, F (2, 87) = 1.55, p = .22.

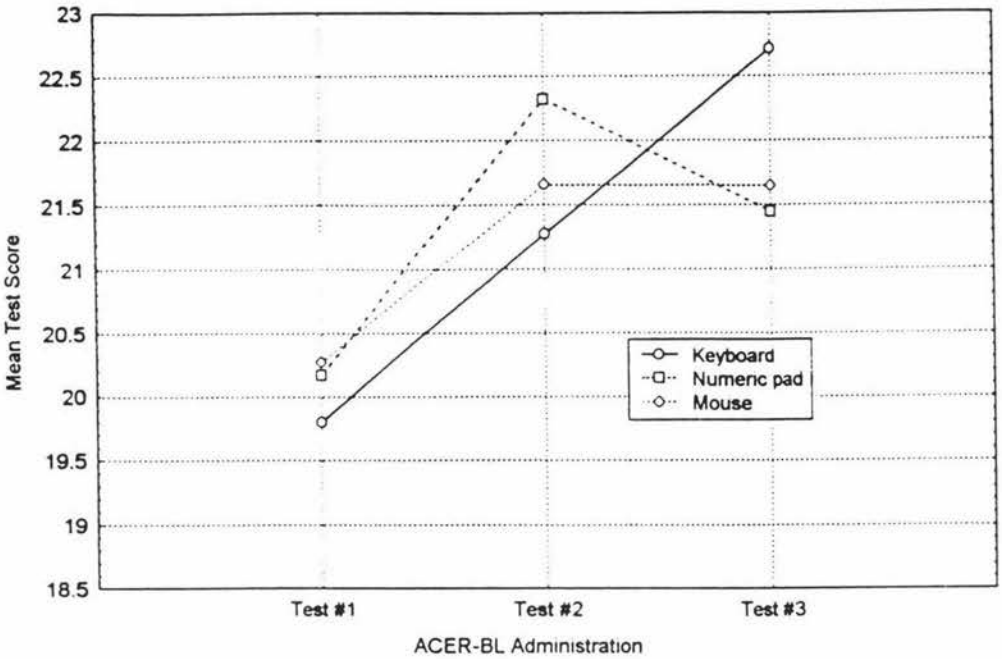

Figure 6.2. Mean test score for each input device, by ACER-BL administration.

Analysis of Order Effects

The effect of order (i.e. carryover) of input device on ACER-BL score was examined. On the second ACER-BL administration, the mean keyboard score was 20.93 when preceded by numeric pad input on the first ACER-BL administration, and 21.59 when preceded by mouse input, $\underline{F}$ (1, 28) = 0.32, $\underline{p}$ = .58. On the second ACER-BL administration, the mean numeric pad score was 21.88 when preceded by keyboard input on the first ACER-BL administration, and 22.80 when preceded by mouse input, $\underline{F}$ (1, 29) = 0.52, $\underline{p}$ = .48. On the second ACER-BL administration, the mean mouse score was 20.46 when preceded by keyboard input on the first ACER-BL administration, and 22.77 when preceded by numeric pad input, $\underline{F}$ (1, 28) = 3.76, $\underline{p}$ = .06. None of these three analyses reached statistical significance.

The effect of order of input device was also examined using test completion times. On the second ACER-BL administration, the mean keyboard completion time was 5 minutes and 34 seconds when preceded by numeric pad input on the first ACER-BL administration, and 5 minutes and 20 seconds when preceded by mouse input, $\underline{F}$ (1, 28) = 0.11, $\underline{p}$ = .75. On the second ACER-BL administration, the mean numeric pad completion time was 5 minutes

and 1 second when preceded by keyboard input on the first ACER-BL administration, and 5 minutes and 26 seconds when preceded by mouse input, F (1, 29) = 0.28, p = .60. On the second ACER-BL administration, the mean mouse completion time was 5 minutes and 6 seconds when preceded by keyboard input on the first ACER-BL administration, and 5 minutes and 35 seconds when preceded by numeric pad input, F (1, 27) = 0.36, p = .56. None of these three analyses reached statistical significance.

Input Device Analyses

As there was no order effect on ACER-BL scores or test completion times and, as treatments were counterbalanced, a one factor repeated measures ANOVA was used to compare the ACER-BL scores across the three input devices. There was no significant effect of input device for ACER-BL mean score, F (2, 88) = 0.18, p = .83, with mean ACER-BL scores of 21.27, 21.34, and 21.18 for the keyboard, numeric pad, and mouse respectively. A one factor repeated measures ANOVA also indicated no significant effect of input device on ACER completion time, F (2, 88) = 0.09, p = .92, with mean ACER-BL completion times of 6 minutes and 12 seconds, 5 minutes and 58 seconds, and 6 minutes and 8 seconds for the keyboard, numeric pad, and mouse respectively.

The influence of participant characteristics on the input device test scores was analysed using two factor repeated measures ANOVAs, with the dichotomised characteristics entered as grouping variables. Major was excluded from analysis due to the large number of majors reported across subjects, resulting in a total of 14 nonsignificant ANOVAs (refer Appendix E).

Ecological Validity Analyses

To determine the ecological validity of the three input devices, independent t-tests were used to determine the influence of participant characteristics on test score on the first ACER-BL administration. Participants who had completed at least Bursary mathematics had a significantly higher ACER-BL mean score than participants who had less education in mathematics, t (87) = 2.69, p = .008. Participants with at least Bursary mathematics

understanding had significantly higher mean ACER-BL scores on the keyboard and numeric pad, t (28) = 2.57, p = .02, and t (26) = 2.65, p = .01 respectively, although there was no significant difference between the groups on the mouse, t (29) = 0.69, p = .50. Participants who were second or third year undergraduates had a significantly higher ACER-BL mean score than first year undergraduates, t (88) = 2.11, p = .04. However, there were no significant differences in ACER-BL mean score for the keyboard, numeric pad, or mouse, t (28) = 1.91, p = .07, t (27) = 1.59, p = .12, and t (29) = 0.04, p = .97 respectively. The other 11 dichotomised participant characteristics did not interact with ACER-BL mean score.

Test score on the first ACER-BL administration was analysed as a function of previous experience with similar questionnaires (Question 10 of the Anxiety and General Questionnaires, refer Appendix D). There was no significant difference in ACER-BL mean score between participants with experience on similar tests and participants without this experience, t (64.9) = 0.35, p = .73.

Questions 15, 16, and 17 of the General Questionnaire (Appendix D) were examined using frequency statistics. Seventy-three percent of participants reported feeling no eyestrain during ACER-BL administration, and 65% percent reported no difference in eyestrain across the three ACER-BL administrations.

Test Anxiety Analyses

A repeated measures ANOVA design was used to determine the changes in participant test anxiety for the first 5 questions of the test anxiety measures (refer Appendix D, First Anxiety Questionnaire, Second Anxiety Questionnaire, and General Questionnaire). For general ACER-BL test anxiety, represented by Questions 1 to 3, there was a significant interaction between General Anxiety and ACER-BL Administration, F (4, 85) = 4.41, p = .003, and a significant main effect for both General Anxiety, F (2, 87) = 22.56, p < .001, and Administration, F (2, 87) = 31.19, p < .001. As Figure 6.3 shows, anxiety dropped once participants had completed the ACER-BL practice questions, and increased during ACER-

BL administration, and these types of anxiety were lowest on the second and third ACER-BL administrations.

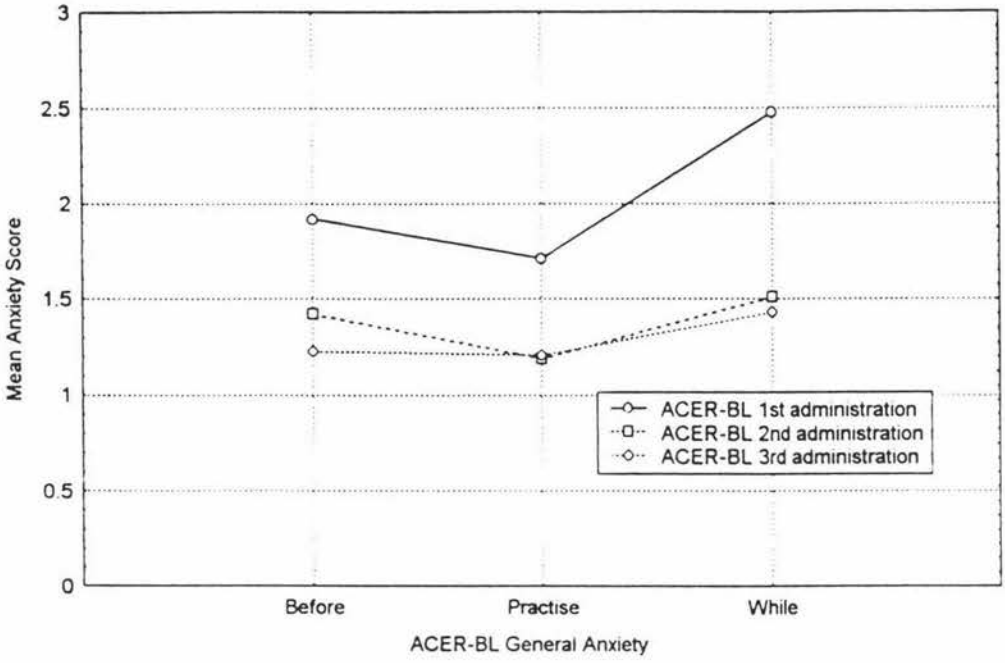

Figure 6.3. Mean general anxiety scores by ACER-BL administration, scores for treatment groups combined.

Regarding ACER-BL subsection anxiety (Question 4), there was a significant interaction between Subsection Anxiety and ACER-BL Administration, F (4, 84) = 6.41, p < .001, and a significant main effect for both Subsection Anxiety, F (2, 86) = 20.56, p < .001, and for Administration, F (2, 86) = 34.40, p < .001. As shown in Figure 6.4, test anxiety increased over the subsections, with the second and third ACER-BL administrations having the smaller increase in anxiety over subsections and the lower mean subsection anxiety scores.

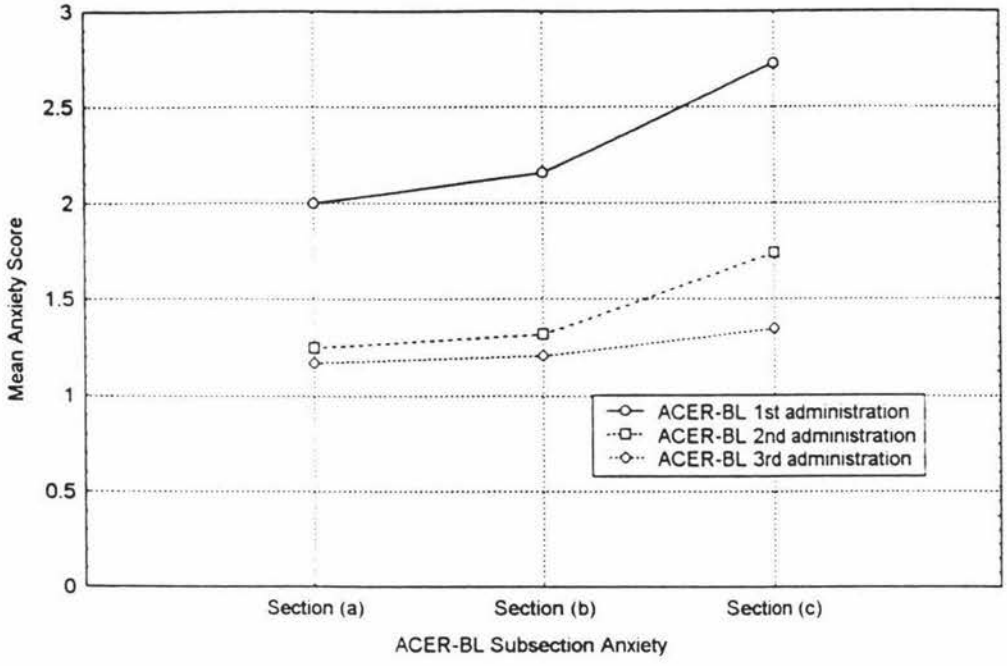

Figure 6.4. Mean subsection anxiety scores by ACER-BL administration, scores for treatment groups combined.

Regarding ACER-BL question type anxiety (Question 5), there was no significant interaction between Question Type Anxiety and ACER-BL Administration, $\underline{F}$ (6, 84) = 0.67, $\underline{p}$ = .67. There were significant main effects for both Question Type Anxiety, $\underline{F}$ (3, 87) = 7.01, $\underline{p} < .001$, and for Administration, $\underline{F}$ (2, 88) = 27.72, $\underline{p} < .001$. As Figure 6.5 shows, participants reported the highest mean anxiety for the synonym items, and the lowest mean anxiety for the analogy and premises items on all three ACER-BL administrations, although question type anxiety decreased for all items over ACER-BL administrations.

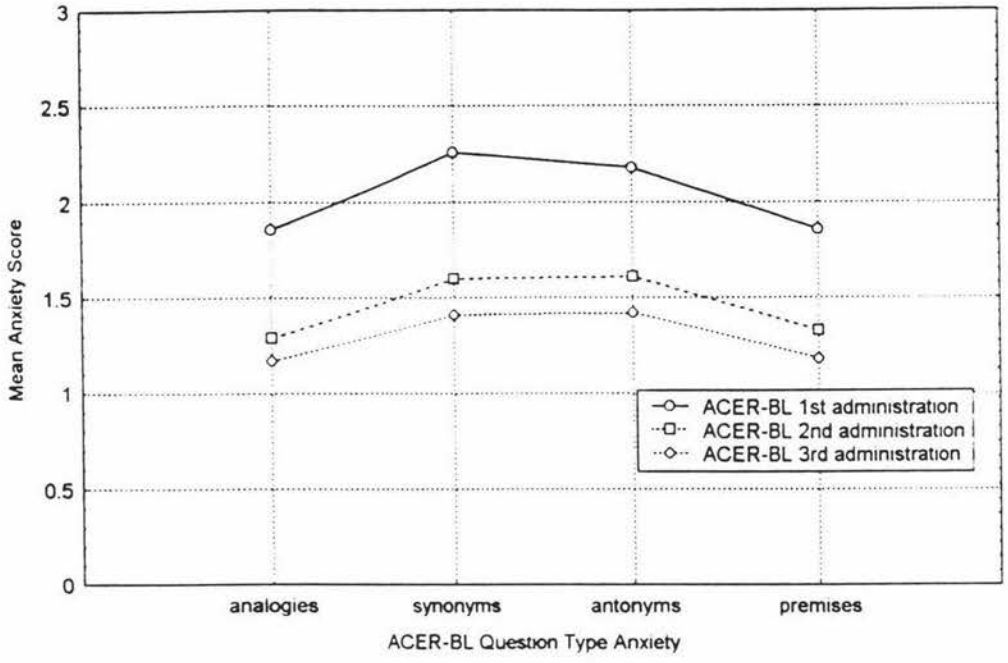

Figure 6.5. Mean question type anxiety scores by ACER-BL administration, scores for treatment groups combined.

Questions 13 and 14 of the General Questionnaire (Appendix D) were examined using frequency statistics. Fifty-eight percent of participants reported feeling the highest test anxiety on the first ACER-BL administration, and 48% percent reported feeling the lowest test anxiety on the third administration. Thirteen percent of participants reported no change in anxiety levels across ACER-BL administrations.

Results Summary and Brief Conclusions

There were no mean test score, or mean test completion time, differences between the six treatment groups across ACER-BL administration. All groups had higher mean scores on the second test administration, and on the third administration to a lesser extent (refer Figure 6.1), and decreased test completion times across the three ACER-BL administrations. When ACER-BL mean test scores on input device were analysed over test administration (refer Figure 6.2), a between-subjects analysis, keyboard-based and numeric pad-based mean scores showed a significant increase across ACER-BL administrations, but there was no significant change in mean mouse-based score. There was no order effect for

input device, suggesting that increases in test performance over ACER-BL administration was purely due to practise effects that were independent of input device, and this result was not influenced by participant characteristics such as gender and age.

Test takers complete the ACER-BL once in the applied setting, so the influence of participant characteristics on ACER-BL score for the first test administration was analysed. Mathematical ability influenced ACER-BL score, with participants of Bursary mathematics ability or higher having higher mean ACER-BL scores overall, and on the keyboard and numeric pad devices, than participants with less mathematical ability. Undergraduate year also influenced ACER-BL score, with second and third year undergraduates producing higher mean ACER-BL scores than first year undergraduates, although there was no difference between these groups when input device was entered into the analysis. One reason for the higher performance of second and third year undergraduates could be that their test taking experience is greater than that of first year undergraduates. While there was no influence of experience with similar tests on ACER-BL mean scores, perhaps familiarity with any academic testing situation confers an advantage in test performance on computerised aptitude tests.

All anxiety scores decreased over ACER-BL administration although the pattern of anxiety remained reasonably constant across administration. Test anxiety during ACER-BL administration was at least as high as test anxiety immediately prior to administration, with anxiety immediately following the completion of the ACER-BL practise questions having the lowest anxiety scores. Test anxiety during ACER-BL administration increased over ACER-BL subsection. Also, participants reported the highest anxiety for the synonym and antonym items and the lowest anxiety for the antonym and premise/conclusion items. However, all mean anxiety measures represented "slight anxiety" with no self-reported influence on test performance.

Chapter 7: General Discussion

Paper-and-Pencil and Computer Equivalence of the ACER-BL (Study 1)

Bartram and Bayliss (1984) suggest that there is no reason to assume that the paper-and-pencil and computerised formats of a test will differ in reliability, although there may be norm differences between the two formats. However, Burke and Normand (1987) recommend that the equivalence between a computerised test and it's traditional paper-and-pencil format must be demonstrated by high score correlations between the two formats, and by almost identical distributions in test score frequencies (see also Hofer & Green, 1985). For these reasons, detailed score comparisons between the computerised and paper-and-pencil formats of the ACER-BL, and between these formats and the New Zealand university norm sample, were performed in Study 1. The hypothesis was that the paper-and-pencil and computerised formats of the ACER-BL would be equivalent.

There were only minor differences between the ACER-BL university norm sample and the participants in Study 1, with the norm sample having a slightly lower mean score, higher standard deviation, and higher standard error of measurement than the P and C group participants for the first ACER-BL administration (refer Table 5.5). The internal consistency measures ranged from a low of .70 on the second ACER-BL administration for the computer format to a high of .80 on the first ACER-BL administration for the computer format. The NZCER reported an internal consistency result of .71 for their university norm sample (Reid et al., 1986). However, the NZCER used the KR21 internal consistency measure, whereas coefficient alpha was used for the Study 1 results. As the reliability measure used has little influence on the actual internal consistency result, these comparisons suggest that the internal consistency reliability of the P and C groups in Study 1 was at least as high as that reported for the university norm sample.

The test-retest reliability was .93 for the paper-and-pencil format and .87 for the computer format. The NZCER did not report a test-retest reliability for any of their New Zealand norm samples (Reid et al., 1986). An arbitrary test-retest delay of 10 minutes was used in Study 1 so that participants completed both test administrations in the same testing session, thereby reducing participant attrition. It has been suggested (Kline, 1993) that the

minimum test-retest period is three months in order to minimise the possibility that participants will remember their responses on the first test. Such a long test-retest delay would have caused unacceptably high participant attrition in Study 1, as participants were unpaid volunteers. Thus, the test-retest reliabilities reported in Study 1 are probably artificially high due to the low test-retest delay.

The equivalence of the paper-and-pencil and computerised formats of the ACER-BL was also analysed using subsection scores. For both treatment groups, and on both ACER-BL administrations, scores decreased over the subsections, and there were no significant differences in mean subsection scores between the two treatment groups. These mean score changes were greater for the first ACER-BL administration, and were associated with lower mean subsection scores compared with the second administration, a predictable result given that this test is speeded. These subsection results suggest that participants in both treatment groups used a sequential answering strategy to complete the ACER-BL, and a number of participants did not complete subsection (c) on the first ACER-BL administration due to the 15 minute time limit of the test. This conclusion is supported by the fact that the elapsed times for each item answered in the C group indicated a sequential answering strategy was used by the majority of these participants. There were no significant differences in ACER-BL mean score between the paper-and-pencil and computerised formats for either ACER-BL administration.

In summary, the results of Study 1 show no significant differences in mean score, show similar test-retest and internal consistency reliabilities, fulfilling Burke and Normand's (1987) criteria for demonstrating the equivalency of the test formats. Thus, the hypothesis that the paper-and-pencil and computerised formats of the ACER-BL would be equivalent is supported. These results also support the work of Kapes and Vansickle (1992), who demonstrated equivalency between the paper-and-pencil and computerised formats of a vocational guidance test using a test-retest method, and extend the general research on equivalency of aptitude tests (e.g. Huba, 1988, Van de Vijver & Harsveld, 1994). As all participants in Study 1 were able to skip, retrace, and change their answers, the equivalency finding for this study supports the results of Spray et al. (1989) and Lunz and Bergstrom (1994).

Practise Effects (Studies 1 and 2)

The analyses of participant characteristics between the P and C groups in Study 1 indicate that, on all 15 measured characteristics, the participants in each treatment group were equal. The difference in test scores between the two formats cannot be a function of differences in participant characteristics, and must therefore be due to the testing situation itself. Apart from minor, necessary, changes to test layout, the only difference between the two treatment groups was how the ACER-BL was administered.

Although the participants in both ACER-BL formats in Study 1 had higher mean scores on the second test administration, this result only reached significance for the C group. Unfortunately, few studies on computerised aptitude tests have performed test-retest comparisons between the paper-and-pencil and computerised formats of these tests, so little is known about the factors underlying practise effects, and what might explain the difference in practise effect for the two ACER-BL formats.

Despite the interaction between treatment group and ACER-BL administration in Study 1, there was no significant difference in mean test score between the P and C groups on the first administration. The practical implication of this is that on initial testing, which is how this aptitude test is normally administered, there may be no difference between the two test formats.

A number of participant characteristics appeared to influence the practise effects in Study 1. Males had a lower mean test score on the first ACER-BL administration, and a higher mean test score on the second administration, than females (Figure 5.2). Second and third year undergraduates had lower mean test scores on both administrations than first year undergraduates, although they showed a greater increase in mean test score across the administrations (Figure 5.4). Touch typists had a higher mean score on the first administration and a lower mean score on the second administration than other typists (Figure 5.5).

Three-way interactions were then conducted on these three variables. Although these interactions failed to reach significance, trends occurred for each interaction. Males in

the C group had the lowest mean score for the first ACER-BL administration and showed the greatest practise effect, producing the highest mean score for the second administration (Figure 5.3). This trend is interesting as researchers (e.g. Hofer and Green, 1985) in the CPT field have predicted that traditionally disadvantaged groups, such as females, would be disadvantaged by computerised testing as represented by the first ACER-BL administration scores. Non-touch typists in the C group had the lowest mean score for the first administration and showed the greatest practise effect, producing the highest mean score for the second administration (Figure 5.6). This trend supports the results of Beaumont (1985a), who found that unfamiliarity with an input device decreased participants' scores on a digit span task. As each cell in these three-way interaction averaged 5 participants, a larger sample size may show that these trends are significant. Also, as university students comprise a relatively homogenous sample, demographic effects on test score may be stronger in the general population.

Practise effects also occurred in Study 2, especially between the first and second administrations of the ACER-BL (refer Figure 6.1). Between-subjects analyses indicated that the keyboard and numeric pad were the input devices resulted in greater practise effects, with no practise effect occurring for the mouse (Figure 6.2), and there were no order effects for input device. Thus, the hypothesis that an order effect of devices will occur is not supported by the Study 2 results. Analyses of test completion times produced the same results. The implication of these results is that the significant practise effect for C group participants in Study 1 could be due entirely to the fact that the keyboard was used to answer the test. There were no significant interactions between participant characteristics, such as computer familiarity, and practise in Study 2, thus the hypothesis that familiarity with an input device will increase test performance was not supported.

For both Study 1 and 2, test score on the first ACER-BL administration was analysed as a function of previous experience with similar questionnaires (Question 10 of the Anxiety and General Questionnaires, refer Appendix B, and Question 10 of the Anxiety and General Questionnaires, refer Appendix D). General test familiarity did not significantly influence test score.

In summary, the factors underlying the practise effects found in Studies 1 and 2 appear to be reasonably complex and require further study. However, the implication of these results is that some computerised tests may disadvantage some population groups, such as males. There are a number of methods that could be used to determine factors influencing practise effects on computerised tests. First, a comparison of paper-and-pencil, keyboard, numeric pad, and mouse practise effects could be performed using larger sample numbers to determine if the small participant numbers caused the nonsignificant interaction results in Studies 1 and 2. Second, the physical manipulation of the input devices (including pencils) could be timed, and this measure could then be compared with the appropriate test completion time and test score. This would indicate the degree to which motor (and thus also cognitive) control of an input device interacts with the cognitive processing of test items. It is recommended that the participants for further studies are representative of the general population.

Test Anxiety (Studies 1 and 2)

There were no significant differences in reported test anxiety, by treatment group, for any of the temporal measures of ACER-BL anxiety in Study 1. These temporal measures are represented by Questions 1 through 4 of the Anxiety and General Questionnaires in Study 1 (refer Appendix B), and by Questions 1 though 4 of the First Anxiety, Second Anxiety, and General Questionnaires in Study 2 (refer Appendix D). For both the P and C groups, anxiety dropped once participants had completed the ACER-BL practice questions, and increased during ACER-BL administration, and these types of anxiety were lowest on the second ACER-BL administration (Figure 5.7). Test anxiety increased over the ACER-BL subsections, with the second ACER-BL administration having the smaller increase in anxiety over subsections and the lower mean subsection anxiety scores (Figure 5.8). This pattern of temporal anxiety changes also occurred in Study 2 (Figures 6.3 and 6.4).

These temporal anxiety results suggest that there are no significant differences in temporal anxiety between paper-and-pencil and computerised formats, thus general measures of anxiety should not distinguish between paper-and-pencil and computerised test

takers. This finding conflicts with the results of Llabre et al. (1987), Ward et al. (1989), and George et al. (1992), and supports the results of Chin et al. (1991). However, only 23% of participants in the Llabre et al. (1987) study reported familiarity with computers, and Ward et al. (1989) and George et al. (1992) did not measure the computer familiarity of their participants, suggesting that the computer experience of participants in these three studies was markedly different to that of participants in this thesis. Chin et al. (1991) measured the computer familiarity of participants and found, even when familiarity was included in the analysis, no significant difference between the paper-and-pencil and computer groups for general test anxiety.

While ACER-BL question type anxiety was not significantly different between the two treatment groups in Study 1, there was a significant qualitative difference (Figure 5.9). The first administration of the ACER-BL for the C group produced a different pattern of question type anxiety compared to the P group administrations and the second administration of the C group. This first C group anxiety result appears to be an anomaly, with the question type anxiety patterns for all ACER-BL administrations in Study 2 replicating those found for the P group in Study 1 (compare Figure 6.5 with Figure 5.9).

There were no significant differences between the P and C groups in Study 1 on their perceptions of the influence of test format on their test anxiety, with both groups reporting a mean indicating "no influence" of test format. For both studies, the majority of participants reported the highest anxiety for the first ACER-BL administration. The majority of Study 1 participants reported the lowest test anxiety for the second ACER-BL administration, compared to 48% of Study 2 participants reporting the lowest anxiety for the third administration. One possible reason for this difference could be that 13% of Study 2 participants reported no difference in anxiety levels across ACER-BL administrations.

Thus, altering test format produced no overall quantitative or qualitative changes in test anxiety. As mean test anxiety on each Likert item indicated slight anxiety, with no effect on test performance, for all treatment groups, this is an unsurprising result. As the anxiety measures were designed to be highly sensitive, one possible reason for the low anxiety reported is that the ACER-BL administration was simply not realistic to participants. Another possible reason is that university students are less likely to be anxious on aptitude

tests because they are in an ongoing testing situation by virtue of their studies. A replication of Study 1 using the general population may show that test format does significantly influence test anxiety.

Eyestrain Analyses

A number of research studies (e.g. Belmore, 1985) and field studies (e.g. Lie and Watten, 1994) suggest that computer-presented text may be more difficult to comprehend and may create more eyestrain problems than paper-presented text. There was no evidence that these problems occurred in Studies 1 and 2, with the majority of participants in both studies reporting no change in eyestrain across ACER-BL administrations, and no overall eyestrain.

Summary and Implications of Findings

- Equivalency was demonstrated between the paper-and-pencil and computerised formats of the ACER-BL. This suggests that the computerised ACER-BL can be used as an alternative testing format to the paper-and-pencil ACER-BL.
- Practise effects occurred only on the computerised administrations of the ACER-BL, and these appear to be a function of input device used and participant characteristics, such as gender. This suggests that some population groups, such as people who cannot touch-type, may be slightly disadvantaged by the use of the computerised ACER-BL. Another implication is that the mouse should be the input device to use on computerised tests, rather than the keyboard or numeric pad.
- Test completion times provided no extra information over test score analyses. This suggests that item completion time may be an unnecessary measurement for research on speeded tests.
- There were no differences in test taker anxiety between the paper-and-pencil and computerised formats. This suggests that the performance of computerised test takers will not be disadvantaged due to anxiety originating from the test format itself.

- The nonsignificant differences in test anxiety between the P and C groups on the ACER-BL-specific anxiety questions suggests that general measures of anxiety are adequate for researching computerised test anxiety.
- There were no differences in participant eyestrain between the paper-and-pencil and computerised formats of the ACER-BL. This suggests that computer-presented text is not unduly stressing for these test takers, at least for tests of relatively short duration.

References

Allwood, C.M. (1986). Novices on the computer: A review of the literature. *International Journal of Man-Machine Studies, 25*, 633-658.

American Psychological Association. (1966). Minutes of the annual meeting of the council of representatives. *American Psychologist, 21*, 1127-1147.

Bartram, D., & Bayliss, R. (1984). Automated testing: Past, present and future. *Journal of Occupational Psychology, 57*, 221-237.

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.

Bedford, F.L. (1994). Of computer mice and men. *Cahiers de Psychologie Cognitive, 13(4)*, 405-426.

Belmore, S.M. (1985). Reading computer-presented text. *Bulletin of the Psychonomic Society, 23(1)*, 12-14.

Booth, J. (1991). The key to valid computer-based testing: The user interface. *Revue Européenne de Psychologie Appliquée, 41(4)*, 281-293.

Booth-Kewley, S., Edwards, J.E., & Rosenfeld, P. (1992). Impression management, social desirability, and computer administration of attitude questionnaires: Does the computer make a difference? *Journal of Applied Psychology, 77(4)*, 562-566.

Boyle, S. (1984). The effect of variations in answer-sheet format on aptitude test performance. *Journal of Occupational Psychology, 57*, 323-326.

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

Carr, A.C., Wilson, S.L., Ghosh, A., Ancill, R.J., & Woods, R.T. (1982). Automated testing of geriatric patients using a microcomputer-based system. *International Journal of Man-Machine Studies, 17*, 297-300.

Chin, C.H.L, Donn, J.S., & Conry, R.F. (1991). Effects of computer-based tests on the achievement, anxiety, and attitudes of grade 10 science students. *Educational & Psychological Measurement, 51*, 735-745.

Committee on Professional Standards and Committee on Psychological Tests and Assessment. (1987). *Guidelines for computer-based tests and interpretations*. Washington, DC: American Psychological Association.

Conoley, C.W., Plake, B.S., & Kemmerer, B.E. (1991). Issues in computer-based test interpretive systems. *Computers in Human Behavior, 7*, 97-101.

Creed, A., Dennis, I., & Newstead, S. (1987). Proof-reading on VDUs. *Behaviour & Information Technology, 6(1)*, 3-13.

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

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

Elwood, D.L., & Griffin, H.R. (1972). Individual intelligence testing without the examiner: Reliability of an automated method. *Journal of Consulting & Clinical Psychology, 38(1)*, 9-14.

Evan, W.M., & Miller, J.R. (1969). Differential effects on response bias of computer vs. conventional administration of a social science questionnaire: An exploratory methodological experiment. *Behavioral Science, 14*, 216-227.

Federico, P. (1992). Assessing semantic knowledge using computer-based and paper-based media. *Computers in Human Behavior, 8*, 169-181.

Fekken, G.C., & Jackson, D.N. (1988). Predicting consistent psychological test item responses: A comparison of models. *Personality & Individual Differences, 9(5)*, 873-882.

Fowler, R.D. (1985). Landmarks in computer-assisted psychological assessment. *Journal of Consulting & Clinical Psychology, 53(6)*, 748-759.

French, C.C. (1986). Microcomputers and psychometric assessment. *British Journal of Guidance & Counselling, 14(1)*, 33-45.

George, C.E., Lankford, J.S., & Wilson, S.E. (1992). The effects of computerized versus paper-and-pencil administration on measures of negative affect. *Computers in Human Behavior, 8(2-3)*, 203-209.

Glaze, R., & Cox, J.L. (1991). Validation of a computerised version of the 10-item (self-rating) Edinburgh Postnatal Depression scale. *Journal of Affective Disorders, 22*, 73-77.

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

Hedl, J.J., O'Neil, H.F., & Hansen, D.N. (1973). Affective reactions toward computer-based intelligence testing. *Journal of Consulting & Clinical Psychology, 40(2)*, 217-222.

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

Hoffman, K.I., & Lundberg, G.D. (1976). A comparison of computer-monitored group tests with paper-and-pencil tests. *Educational & Psychological Measurement, 36*, 791-809.

Horton, S.V., & Lovitt, T.C. (1994) A comparison of two methods of administering group reading inventories to diverse learners: Computer versus pencil and paper. *Remedial & Special Education, 15(6)*, 378-390.

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

Hughes, P.K., & Creed, D.J. (1994). Eye movement behaviour viewing colour-coded and monochrome avionic displays. *Ergonomics, 37(11)*, 1871-1884.

Kapes, J.T., & Vansickle, T.R. (1992). Comparing paper-pencil and computer-based versions of the Harrington-O'Shea Career Decision-Making System. *Measurement & Evaluation in Counseling & Development, 25*, 5-13.

Kearsley, G. (1986). 33 ways to better software design. *Training & Development Journal, 40(7)*, 47-48.

Kennedy, R.S., Wilkes, R.L., Dunlap, W.P., & Kuntz, L.A. (1987). Development of an automated performance test system for environmental and behavioral toxicology studies. *Perceptual & Motor Skills, 65*, 947-962.

Kleinmuntz, B. & McLean, R.S. (1968). Computers in behavioral science: Diagnostic interviewing by digital computer. *Behavioral Science, 13*, 75-80.

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

Lang, P.J. (1969). The on-line computer in behavior therapy research. *American Psychologist, 24*, 236-239.

Lanyon, R.I. (1984). Personality assessment. *Annual Review of Psychology, 35*, 667-701.

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 Educational Research, 55(3)*, 387-413.

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

Lie, I., & Watten, R.G. (1994). VDT work, oculomotor strain, and subjective complaints: An experimental and clinical study. *Ergonomics, 37(8)*, 1419-1433.

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)*, 429-433.

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

Lunz, M.E., & Bergstrom, B.A. (1994). An empirical study of computerized adaptive test administration conditions. *Journal of Educational Measurement, 31(3)*, 251-263.

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89(2)*, 191-216.

Monk, J.J., & Stallings, W.M. (1970). Effects of item order on test scores. *Journal of Educational Research, 63(10)*, 463-465.

Moreland, K.L. (1985). Validation of computer-based test interpretations: Problems and prospects. *Journal of Consulting & Clinical Psychology, 53(6)*, 816-825.

Morris, L.W., & Fulmer, R.S. (1976). Test anxiety (worry and emotionality) changes during academic testing as a function of feedback and test importance. *Journal of Educational Psychology, 68(6)*, 817-824.

Munz, D.C., & Smouse, A.D. (1968). Interaction effects of item-difficulty sequence and achievement-anxiety reaction on academic performance. *Journal of Educational Psychology, 59(5)*, 370-374.

Neri, D.F., Luria, S.M., & Kobus, D.A. (1986). The detection of various colour combinations under different chromatic ambient illuminations. *Aviation, Space, & Environmental Medicine, 57*, 555-560.

Norman, D.A. (1984). Stages and levels in human-machine interaction. *International Journal of Man-Machine Studies, 21*, 365-375.

Oltman, P.K. (1994). *The effect of complexity of mouse manipulation on performance in computerized testing*. Princeton, NJ: Educational Testing Service, RR-94-22.

Powers, D.E. (1986). Relations of test item characteristics to test preparation/test practice effects: A quantitative summary. *Psychological Bulletin, 100(1)*, 67-77.

Rafaeli, S., & Tractinsky, N. (1989). Computerized tests and time: measuring, limiting and providing visual cues for response time in on-line questioning. *Behaviour & Information Technology, 8(5)*, 335-351.

Reardon, R., & Loughead, T. (1988). A comparison of paper-and-pencil and computer versions of the Self-Directed Search. *Journal of Counseling & Development, 67*, 249-252.

Reid, N., Croft, C., Gilmore, A., & Philips, D. (1986). *ACER Advanced Test BL-BQ New Zealand revision: Norms supplement*. Wellington: NZCER.

Rosenfeld, P., Booth-Kewley, S., & Edwards, J.E. (1993). Computer-administered surveys in organizational settings. *American Behavioral Scientist, 36(4)*, 485-511.

Rosenfeld, P., Doherty, L.M., Vicino, S.M., Kantor, J., & Greaves, J. (1989). Attitude assessment in organizations: Testing three microcomputer-based survey systems. *Journal of General Psychology, 116(2)*, 145-154.

Sanitioso, R., & Reynolds, J.H. (1992). Comparability of standard and computerized administration of two personality questionnaires. *Personality & Individual Differences, 13(8)*, 899-907.

Sarason, I.G. (1984). Stress, anxiety, and cognitive interference: Reactions to tests. *Journal of Personality & Social Psychology, 46(4)*, 929-938.

Schriesheim, C.A., Kopelman, R.E., & Solomon, E. (1989). The effect of grouped versus randomized questionnaire format on scale reliability and validity: A three-study investigation. *Educational & Psychological Measurement, 49*, 487-508.

Schuldberg, D. (1990). Varieties of inconsistency across test occasions: Effects of computerized test administration and repeated testing. *Journal of Personality Assessment, 55(1&2)*, 168-182.

Skinner, H.A., & Pakula, A. (1986). Challenge of computers in psychological assessment. *Professional Psychology: Research & Practice, 17(1)*, 44-50.

Snyder, D.K., Widiger, T.A., & Hoover, D.W. (1990). Methodological considerations in validating computer-based test interpretations: Controlling for response bias. *Psychological Assessment, 2(4)*, 470-477.

Spray, J.A., Ackerman, T.A., Reckase, M.D., & Carlson, J.E. (1989). Effect of medium of item presentation on examinee performance and item characteristics. *Journal of Educational Measurement, 26(3)*, 261-271.

Standing Committee on Test Standards. (1984). Note on the computerization of printed psychological tests and questionnaires. *Bulletin of the British Psychological Society, 37*, 416-417.

Stevens, G.C. (1983). User-friendly computer systems? A critical examination of the concept. *Behaviour & Information Technology, 2(1)*, 3-16.

Strang, 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.

Styles, I. (1991). Clinical assessment and computerized testing. *International Journal of Man-Machine Studies, 35*, 133-150.

Torkzadeh, G., & Koufteros, X. (1994). Factorial validity of a computer self-efficacy scale and the impact of computer training. *Educational & Psychological Measurement, 54(3)*, 813-821.

Van de Vijver, F.J.R., & Harsveld, M. (1994). The incomplete equivalence of the paper-and-pencil and computerized versions of the General Aptitude Test Battery. *Journal of Applied Psychology, 79(6)*, 852-859.

Ward, T.J., Hooper, S.R., & Hannafin, K.M. (1989). The effect of computerized tests on the performance and attitudes of college students. *Journal of Educational Computing Research, 5(3)*, 327-333.

Watten, R.G., Lie, I., & Birketvedt, O. (1994) The influence of long-term near-work on accommodation and vergence: A field study. *Journal of Human Ergology, 23*, 27-39.

Wise, S.L., Plake, B.S., Pozehl, B.J, Barnes, L.B., & Lukin, L.E. (1989). Providing item feedback in computer-based tests: Effects of initial success and failure. *Educational & Psychological Measurement, 49*, 479-486.

Yen, W.M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement, 17(4)*, 297-311.

Appendix A. Glossary of Technical Terms

CAT

Computerised Adaptive Testing. A specific form of CPT, where the test question administration is dependent on how the test taker has answered the previous question. Typically, a wrong answer causes the software to administer a question at a difficulty level below the previous question, and a right answer causes the software to administer a question at a difficulty level higher than the previous question.

CBTI

Computer-Based Test Interpretation. For the purposes of this report, CBTI refers only to the computer-generated interpretation of a test taker's response set. The test administration itself would be in paper-and-pencil format, with the test taker answers entered into an automated system, such as by optical scanner.

CGA

Colour Graphics Adapter. This was the first video system for the personal computer that provided colour text and graphics, neither of which are now adequate for current computer software.

CPT

Computerised Psychological Testing. For the purposes of this report, CPT refers only to the testing situation where both test administration and score interpretation are computerised. Where score interpretation is the only automated feature, the term CBTI has been used.

EGA

Enhanced Graphics Adapter. Replaced the CGA system, providing more colours and better text capabilities.

VDU

Visual Display Unit. Also called a monitor, display, or screen. Typically, this acronym is used for monitors connected with dumb terminals and not for monitors connected to personal computers and so forth.

VGA

Video Graphics Array. Replaced the EGA system, with increased colour graphics, resolution, and text capabilities. VGA, or SVGA (SuperVGA) is the standard system on home personal computers currently produced.

Appendix B. Consent Form & Questionnaires Used in Study 1.

This appendix contains:

The information sheet and consent form

Questionnaire 1.1 (Participant Characteristics Questionnaire)

Questionnaire 1.2 (Anxiety Questionnaire)

Questionnaire 1.3 (General Questionnaire)

Analysis of ACER-BL With Massey University Undergraduates

Study 1: Information Sheet and Consent Form

- Thank you for your interest in my research. This study is designed to examine the use of the aptitude test "ACER-BL" with Massey University undergraduates, and the main focus of this research is to examine people's format preferences in the presentation and administration of this aptitude test.

- This study has two main sections: presentation of the aptitude test; and presentation of a questionnaire relating to this aptitude test. Participation in this study will take up around an hour of your time, all of which will be used within the one testing session.

- While each section of this study must be matched to your student ID, this is only for data control, ie to match up correctly each section of the test you complete, and this information is completely confidential. Only the results of statistical analyses will be presented in my thesis, and any published articles arising from this thesis.

- As this is a pilot study, your score on the ACER-BL will not be provided to you. Your individual results, or an interpretation of your results, cannot be provided as all results will remain in raw score form. A summarised copy of the study results will be posted on the Ground Floor noticeboard in the Psychology Department.

- I have read the information sheet for this study and have had the details of the study explained to me. My questions about this study have been answered to my satisfaction, and I understand that I may ask further questions at any time.

- I also understand that I am free to withdraw from this study at any time, or decline to answer any particular questions in this study. I agree to provide information to the researcher on the understanding that it is completely confidential.

- I wish to participate in this study under the conditions set out on the Information Sheet above.

Signed: _____

Name: _____

Date: _____

Questionnaire for ACER-BL Study 1. (Questionnaire 1.1)

Student ID No._____

Age: _____ _____ Today's Date: _____
(*Yrs*) (*Mths*)

Gender: Male Female (please circle one)

1. Which ethnic group do you feel **most** affiliated with?
   (A) NZ European
   (B) NZ Maori
   (C) NZ Samoan
   (D) NZ Tongan

   (E) Other (please identify)_____

   A B C D E (please circle one)

2. Which **general** family income do you feel **best fits** that of your home (prior to attending university)?
   (A) Under $20,000 per year
   (B) $20,000 to $40,000 per year
   (C) $40,000 and above per year

   A B C (please circle one)

3. What type of **eyesight** do you have?
   (A) Uncorrected vision
   (B) Short-sighted, wore glasses/contacts during test
   (C) Short-sighted, did not wear glasses/contacts during test
   (D) Long-sighted, wore glasses/contacts during test
   (E) Long-sighted, did not wear glasses/contacts during test

   A B C D E (please circle one)

4. What is your **highest** educational attainment, or **level of practical knowledge**, in each of the following subject areas?

(A) Mathematics _____

(B) Statistics _____

(C) English _____

(D) Programming _____

5. What year of university are you **currently** attending?

(A) First-year undergraduate
(B) Second-year undergraduate
(C) Third-year undergraduate

A B C (please circle one)

6. What is your **major**?

_____

7. What is your approximate level of **typing ability**?

(A) Passed typing exams (eg Pitmans)
(B) Touch-typist, speed unknown but higher than 40 words per minute
(C) Touch-typist, speed under 40 words per minute
(D) Know where the keys are, but use two fingers to type
(E) "Hunt-and-peck" typist with some keyboard/typewriter familiarity
(F) Little or no prior use of keyboards or typewriters

A B C D E F (please circle one)

8. Have you ever used a computer "**numeric pad**"?

(A) Yes, I use one at home/work
(B) Yes, but only occasional use
(C) Yes, maybe once or twice
(D) No, never used before

A B C D (please circle one)

9. Have you ever used a computer "**mouse**"?

(A) Yes, I use one at home/work
(B) Yes, but only occasional use
(C) Yes, maybe once or twice
(D) No, never used before

A B C D (please circle one)

10. How **familiar** are you with a **computer**?

(A) Have one at home/work:.....use it at least 4 times a week
(B) Have one at home/work: use up to 4 times a week
(C) A friend has one:.....use it at least 4 times a week
(D) A friend has one: use up to 4 times a week
(E) Use one at least 3 times a month
(F) Use one under 3 times a month
(G) Have used a computer consistently before, but not now
(H) Have used a computer slightly, but not now
(I) Have never used a computer

A B C D E F G H I

(please circle one)

Questionnaire for ACER-BL Study 1. (Questionnaire 1.2)

Student ID No._____

Age: _____ _____ Today's Date: _____
(*Yrs*) (*Mths*)

While you are answering this questionnaire, I would like you to think carefully about how you were feeling throughout the ACER-BL test administration. Look at the answers below each question, and decide which option **most accurately** describes how you felt at that particular time.

1. How anxious were you **before** you sat this test?

(1) not at all anxious
(2) slightly anxious.....barely noticeable and performance not affected
(3) slightly anxious quite noticeable but performance not affected
(4) moderately anxious.....performance not affected
(5) moderately anxious performance may have been affected
(6) very anxious.....performance may have been affected
(7) very anxious performance was affected

1	2	3	4	5	6	7

(please circle one)

Why?

_____

_____

2. How anxious were you **after** you completed the practice questions, but **before** you sat this test?

(1) not at all anxious
(2) slightly anxious
(3) moderately anxious
(4) very anxious

1	2	3	4

(please circle one)

Why?

_____

3. How anxious were you **while** you sat this test?

(1) not at all anxious
(2) slightly anxious.....barely noticeable and performance not affected
(3) slightly anxious quite noticeable but performance not affected
(4) moderately anxious.....performance not affected
(5) moderately anxious performance may have been affected
(6) very anxious.....performance may have been affected
(7) very anxious performance was affected

1	2	3	4	5	6	7

(please circle one)

Why?_____

_____

_____

4. Which **part** of the test were you most anxious about?

(1) not at all anxious
(2) slightly anxious.....barely noticeable and performance not affected
(3) slightly anxious quite noticeable but performance not affected
(4) moderately anxious.....performance not affected
(5) moderately anxious performance may have been affected
(6) very anxious.....performance may have been affected
(7) very anxious performance was affected

A. The first 10 questions of the ACER-BL

1	2	3	4	5	6	7

(please circle one)

B. The next 10 questions of the ACER-BL

1	2	3	4	5	6	7

(please circle one)

C. The last 10 questions of the ACER-BL

1	2	3	4	5	6	7

(please circle one)

5. Which **type** of question were you most anxious about?

(1) not at all anxious
(2) slightly anxious.....barely noticeable and performance not affected
(3) slightly anxious quite noticeable but performance not affected
(4) moderately anxious.....performance not affected
(5) moderately anxious performance may have been affected
(6) very anxious.....performance may have been affected
(7) very anxious performance was affected

A. The analogies, eg dolphin is to submarine as sparrow is to...

1	2	3	4	5	6	7

(please circle one)

B. The synonyms, eg find the word that means most nearly the same as peculiar...

1	2	3	4	5	6	7

(please circle one)

C. The antonyms, eg which two of the following words are opposite in meaning...

1	2	3	4	5	6	7

(please circle one)

D. The premises/conclusions, eg a dog that has bitten a child should not be poisoned because the child will then be poisoned. This statement is illogical because...

1	2	3	4	5	6	7

(please circle one)

6. Which type of ACER-BL **format** have you just completed?

(A) Computer-administration
(B) Paper-and-pencil administration

A B (please circle one)

7. Did ACER-BL administration format (paper or computer) **influence** your test anxiety?

(1) Yes.....adversely and strongly
(2) Yes adversely and moderately
(3) Yes.....adversely and slightly
(4) No influence
(5) Yes.....positively and slightly
(6) Yes positively and moderately
(7) Yes.....positively and strongly.

1	2	3	4	5	6	7

(please circle one)

8. What **other factors** do you think influenced your test anxiety on the ACER-BL?

(A) The ACER-BL time limit.....Yes / No
(B) Difficulty in reading ACER-BL items Yes / No
(C) Ease in reading ACER-BL items.....Yes / No
(D) Difficulty in understanding ACER-BL layout Yes / No
(E) Ease in understanding ACER-BL layout.....Yes / No
(F) Entering incorrect response due to unfamiliarity with input device.....Yes / No
(G) Entering incorrect, but within range, response due to unfamiliarity with input device.....Yes / No

9. What **changes** do you think could be made to the presentation of the ACER-BL in order to **reduce** test-taker anxiety? (eg page layout, screen layout)

_____

_____

10. Have you ever completed a questionnaire with a **similar format** to the ACER-BL, eg PAT test, SAT test, personality test?

Yes / No

11. Have you ever completed a **computer-presented** questionnaire with a similar format to the ACER-BL?

Yes / No

If yes, what was the **name** or **type** of test?

_____

Questionnaire for ACER-BL Study 1. (Questionnaire 1.3)

Student ID No._____

Age: _____ _____ Today's Date: _____
(*Yrs*) (*Mths*)

While you are answering this questionnaire, I would like you to think carefully about how you were feeling throughout the ACER-BL test administration. Look at the answers below each question, and decide which option **most accurately** describes how you felt at that particular time.

1. How anxious were you **before** you sat this test?

(1) not at all anxious
(2) slightly anxious.....barely noticeable and performance not affected
(3) slightly anxious quite noticeable but performance not affected
(4) moderately anxious.....performance not affected
(5) moderately anxious performance may have been affected
(6) very anxious.....performance may have been affected
(7) very anxious performance was affected

1	2	3	4	5	6	7

(please circle one)

Why?

_____

_____

2. How anxious were you **after** you completed the practice questions, but **before** you sat this test?

(1) not at all anxious
(2) slightly anxious
(3) moderately anxious
(4) very anxious

1	2	3	4

(please circle one)

Why?

_____

3. How anxious were you **while** you sat this test?

(1) not at all anxious
(2) slightly anxious.....barely noticeable and performance not affected
(3) slightly anxious quite noticeable but performance not affected
(4) moderately anxious.....performance not affected
(5) moderately anxious performance may have been affected
(6) very anxious.....performance may have been affected
(7) very anxious performance was affected

1	2	3	4	5	6	7

(please circle one)

Why?_____

_____

_____

4. Which **part** of the test were you most anxious about?

(1) not at all anxious
(2) slightly anxious.....barely noticeable and performance not affected
(3) slightly anxious quite noticeable but performance not affected
(4) moderately anxious.....performance not affected
(5) moderately anxious performance may have been affected
(6) very anxious.....performance may have been affected
(7) very anxious performance was affected

A. The first 10 questions of the ACER-BL

1	2	3	4	5	6	7

(please circle one)

B. The next 10 questions of the ACER-BL

1	2	3	4	5	6	7

(please circle one)

C. The last 10 questions of the ACER-BL

1	2	3	4	5	6	7

(please circle one)

5. Which **type** of question were you most anxious about?

(1) not at all anxious
(2) slightly anxious.....barely noticeable and performance not affected
(3) slightly anxious quite noticeable but performance not affected
(4) moderately anxious.....performance not affected
(5) moderately anxious performance may have been affected
(6) very anxious.....performance may have been affected
(7) very anxious performance was affected

A. The analogies, eg dolphin is to submarine as sparrow is to...

1	2	3	4	5	6	7

(please circle one)

B. The synonyms, eg find the word that means most nearly the same as peculiar...

1	2	3	4	5	6	7

(please circle one)

C. The antonyms, eg which two of the following words are opposite in meaning...

1	2	3	4	5	6	7

(please circle one)

D. The premises/conclusions, eg a dog that has bitten a child should not be poisoned because the child will then be poisoned. This statement is illogical because...

1	2	3	4	5	6	7

(please circle one)

6. Which type of ACER-BL **format** have you just completed?

(A) Computer-administration
(B) Paper-and-pencil administration

A B (please circle one)

7. Did ACER-BL administration format (paper or computer) **influence** your test anxiety?

(1) Yes.....adversely and strongly
(2) Yes adversely and moderately
(3) Yes.....adversely and slightly
(4) No influence
(5) Yes.....positively and slightly
(6) Yes positively and moderately
(7) Yes.....positively and strongly.

1	2	3	4	5	6	7

(please circle one)

8. What **other factors** do you think influenced your test anxiety on the ACER-BL?

(A) The ACER-BL time limit.....Yes / No
(B) Difficulty in reading ACER-BL items Yes / No
(C) Ease in reading ACER-BL items.....Yes / No
(D) Difficulty in understanding ACER-BL layout Yes / No
(E) Ease in understanding ACER-BL layout.....Yes / No
(F) Entering incorrect response due to unfamiliarity with input device.....Yes / No
(G) Entering incorrect, but within range, response due to unfamiliarity with input device.....Yes / No

9. What **changes** do you think could be made to the presentation of the ACER-BL in order to **reduce** test-taker anxiety? (eg page layout, screen layout)

_____

_____

10. Have you ever completed a questionnaire with a **similar format** to the ACER-BL, eg PAT test, SAT test, personality test?

Yes / No

11. Have you ever completed a **computer-presented** questionnaire with a similar format to the ACER-BL?

Yes / No

If yes, what was the **name** or **type** of test?

_____

12. What was the **order** of administration formats of the ACER-BL?

(A) Paper-and-pencil / Paper-and-pencil
(B) Paper-and-pencil / Computer
(C) Computer / Paper-and-pencil
(D) Computer / Computer

A B C D (please circle one)

13. Did you feel **most anxious** on the first or second ACER-BL test administration?

First / Second / No difference (please circle one)

14. Did you feel **least anxious** on the first or second ACER-BL test administration?

First / Second / No difference (please circle one)

15. Did you perceive any **eyestrain** during ACER-BL test administration?

Yes / No (please circle one)

16. During which ACER-BL test administration did you perceive the **most eyestrain**?

First / Second / No difference (please circle one)

Appendix C. Participant Characteristic Interactions With Mean Test Score (Study 1): Graphs of Insignificant Interactions

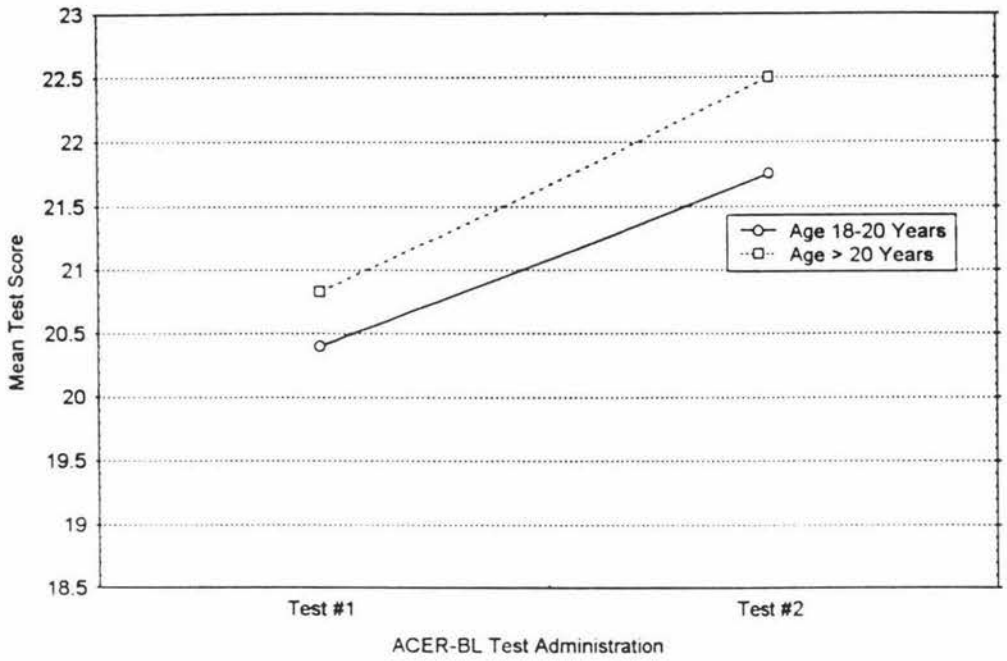

Figure C1. Mean test score for each ACER-BL administration, by participant age.

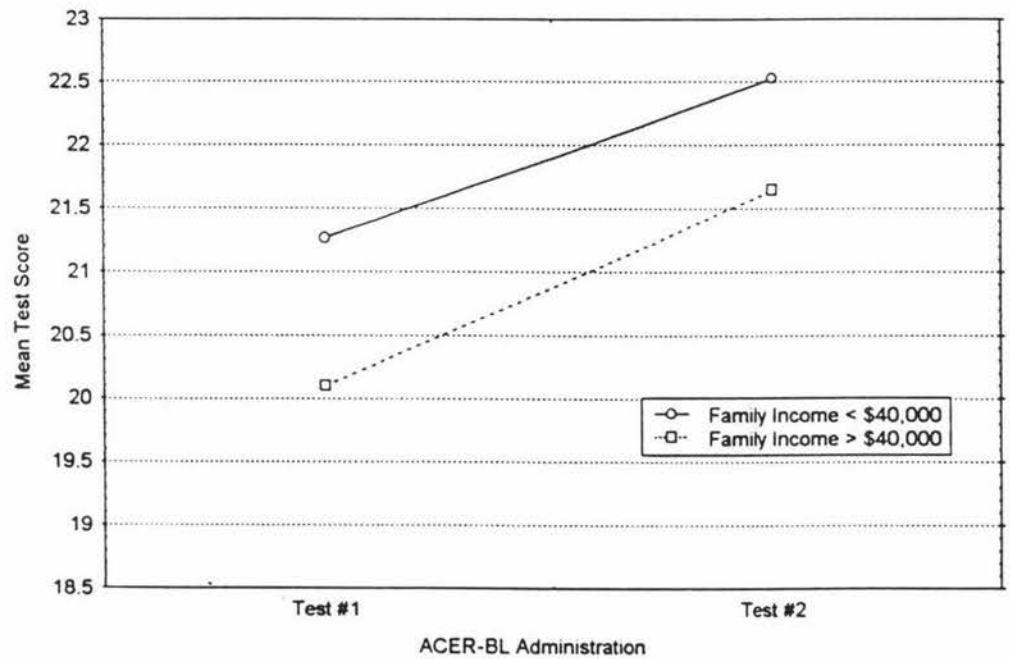

Figure C2. Mean test score for each ACER-BL administration, by annual family income ($NZD).

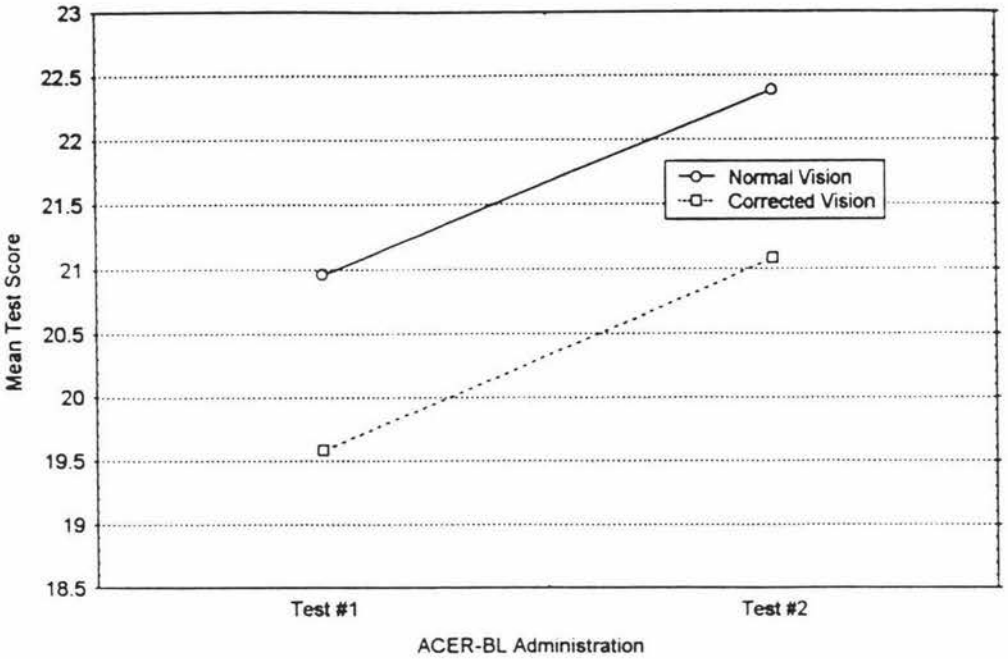

Figure C3. Mean test score for each ACER-BL administration, by participant vision.

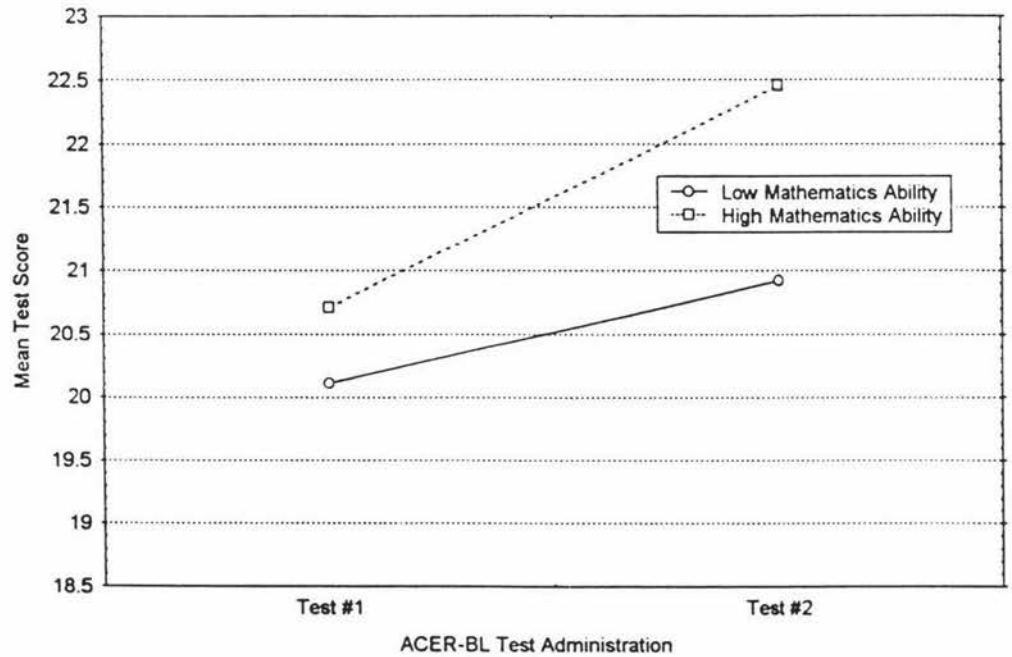

Figure C4. Mean test score for each ACER-BL administration, by participant mathematics ability.

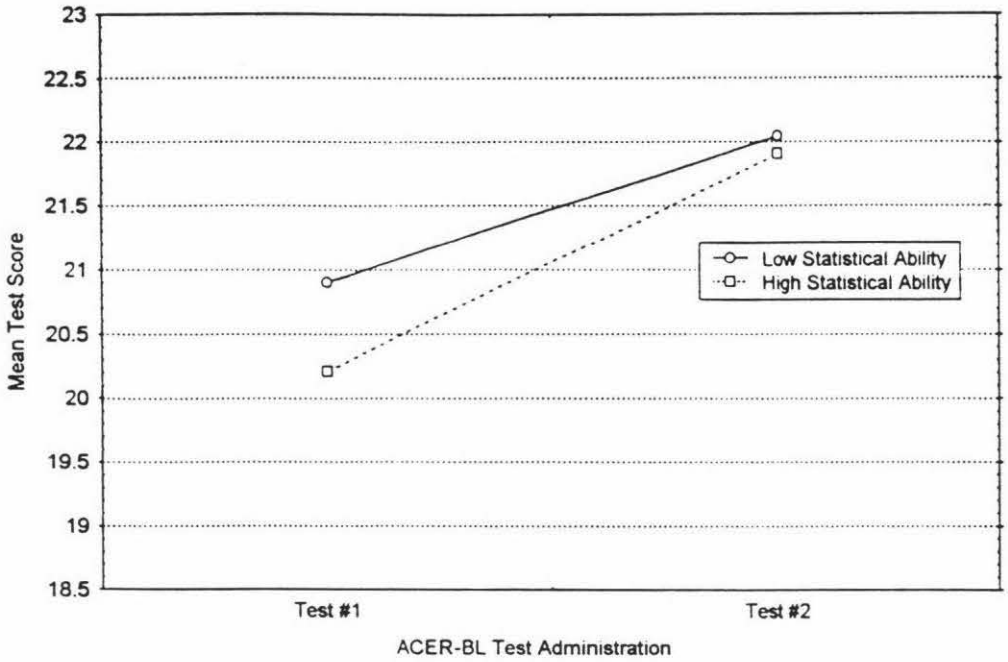

Figure C5. Mean test score for each ACER-BL administration, by participant statistical ability.

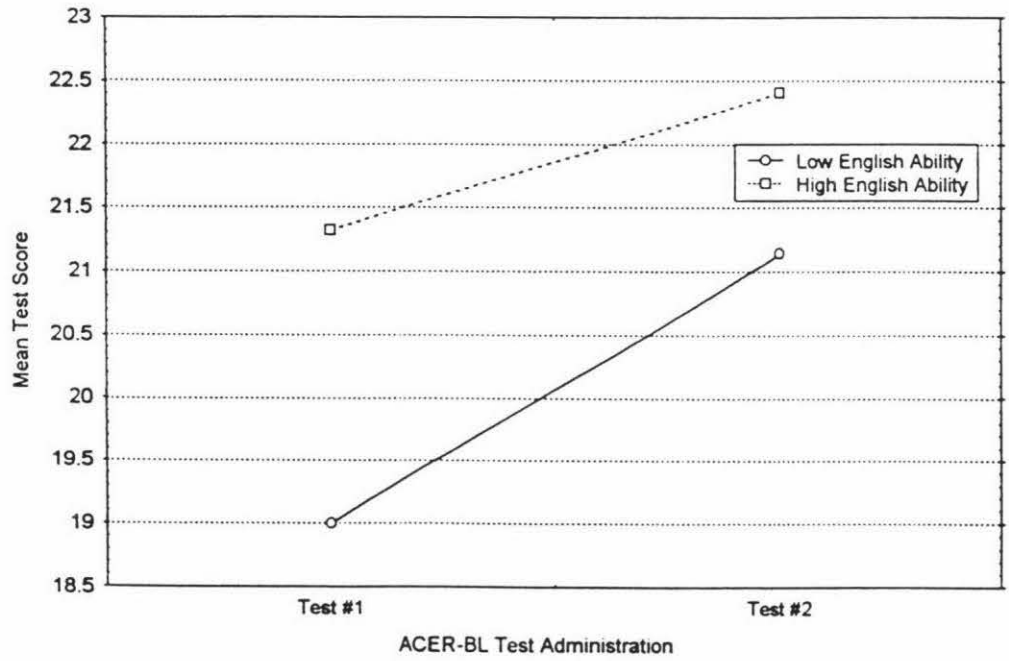

Figure C6. Mean test score for each ACER-BL administration, by participant English ability.

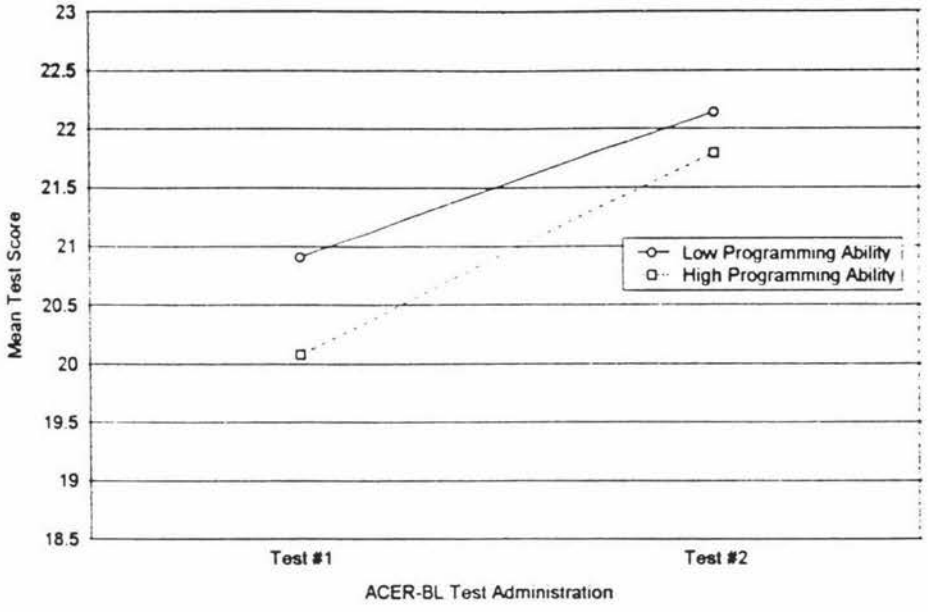

Figure C7. Mean test score for each ACER-BL administration, by participant programming ability.

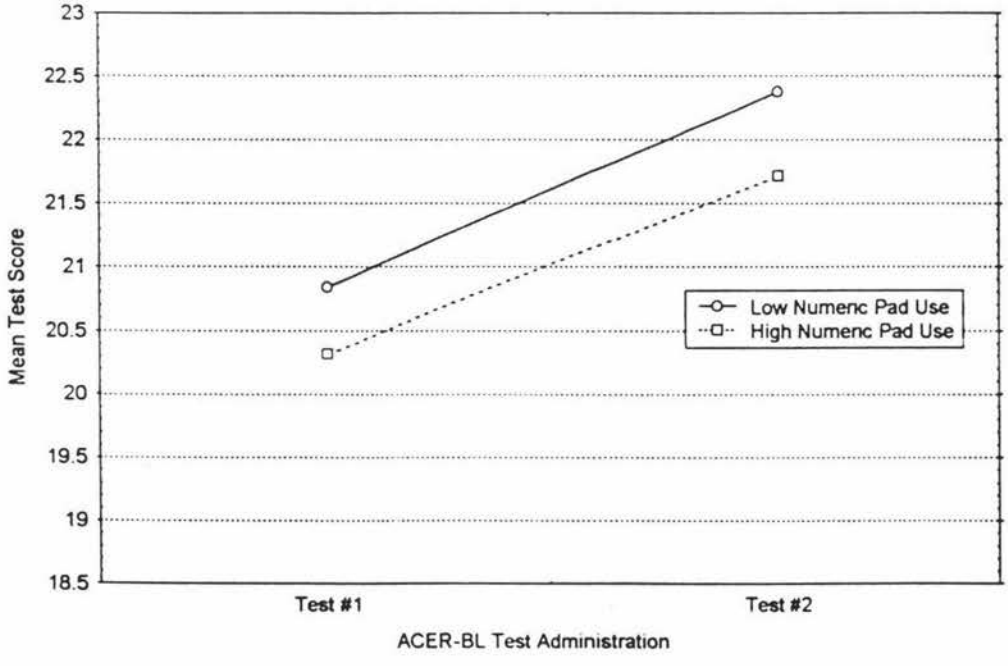

Figure C8. Mean test score for each ACER-BL administration, by participant numeric pad use.

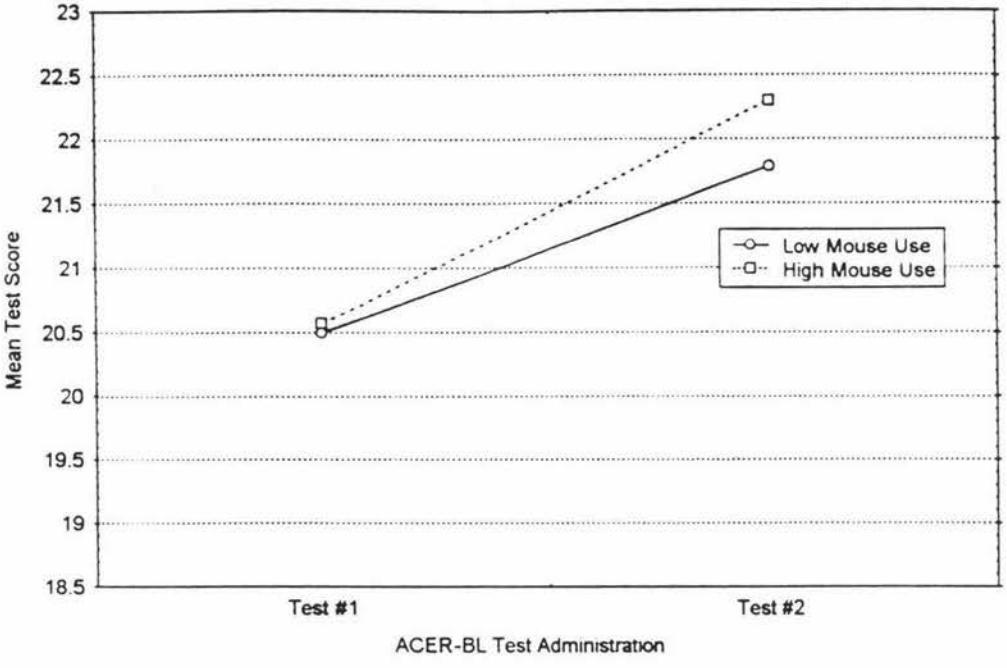

Figure C9. Mean test score for each ACER-BL administration, by participant mouse use.

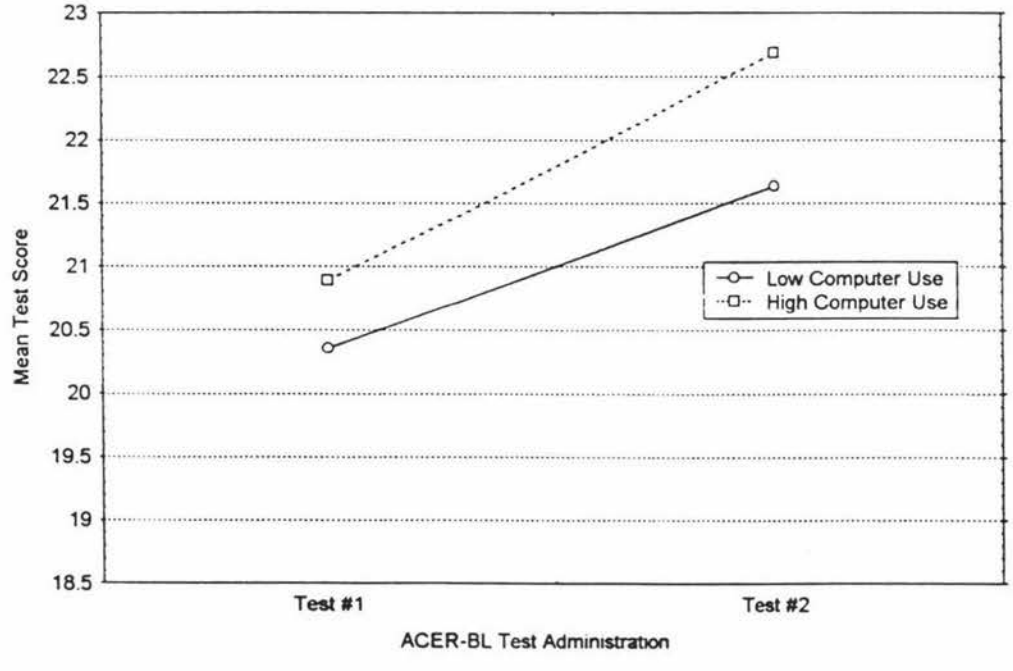

Figure C10. Mean test score for each ACER-BL administration, by participant computer use.

Appendix D. Consent Form & Questionnaires Used in Study 2.

This appendix contains:

The information sheet and consent form

Questionnaire 2.1 (Participant Characteristics Questionnaire)

Questionnaire 2.2 (First Anxiety Questionnaire)

Questionnaire 2.3 (Second Anxiety Questionnaire)

Questionnaire 2.4 (General Questionnaire)

Analysis of ACER-BL With Massey University Undergraduates

Study 2: Information Sheet and Consent Form

- Thank you for your interest in my research. This study is designed to examine the use of the aptitude test "ACER-BL" with Massey University undergraduates, and the main focus of this research is to examine people's input device preferences in the presentation and administration of this aptitude test.

- This study has two main sections: presentation of the aptitude test; and presentation of a questionnaire relating to this aptitude test. Participation in this study will take up around an hour of your time, all of which will be used within the one testing session.

- While each section of this study must be matched to your student ID, this is only for data control, ie to match up correctly each section of the test you complete, and this information is completely confidential. Only the results of statistical analyses will be presented in my thesis, and any published articles arising from this thesis.

- As this is a pilot study, your score on the ACER-BL will not be provided to you. Your individual results, or an interpretation of your results, cannot be provided as all results will remain in raw score form. A summarised copy of the study results will be posted on the Ground Floor noticeboard in the Psychology Department.

- I have read the information sheet for this study and have had the details of the study explained to me. My questions about this study have been answered to my satisfaction, and I understand that I may ask further questions at any time.

- I also understand that I am free to withdraw from this study at any time, or decline to answer any particular questions in this study. I agree to provide information to the researcher on the understanding that it is completely confidential.

- I wish to participate in this study under the conditions set out on the Information Sheet above.

Signed: _____

Name: _____

Date: _____

Questionnaire for ACER-BL Study 2. (Questionnaire 2.1)

Student ID No._____

Age: _____ _____ Today's Date: _____
(*Yrs*) (*Mths*)

Gender: Male Female (please circle one)

1. Which ethnic group do you feel **most** affiliated with?
   (A) NZ European
   (B) NZ Maori
   (C) NZ Samoan
   (D) NZ Tongan

   (E) Other (please identify)_____

   A B C D E (please circle one)

2. Which **general** family income do you feel **best fits** that of your home (prior to attending university)?
   (A) Under $20,000 per year
   (B) $20,000 to $40,000 per year
   (C) $40,000 and above per year

   A B C (please circle one)

3. What type of **eyesight** do you have?
   (A) Uncorrected vision
   (B) Short-sighted, wore glasses/contacts during test
   (C) Short-sighted, did not wear glasses/contacts during test
   (D) Long-sighted, wore glasses/contacts during test
   (E) Long-sighted, did not wear glasses/contacts during test

   A B C D E (please circle one)

4. What is your **highest** educational attainment, or **level of practical knowledge**, in each of the following subject areas?

(A) Mathematics _____

(B) Statistics _____

(C) English _____

(D) Programming _____

5. What year of university are you **currently** attending?

(A) First-year undergraduate
(B) Second-year undergraduate
(C) Third-year undergraduate

A B C (please circle one)

6. What is your **major**?

_____

7. What is your approximate level of **typing ability**?

(A) Passed typing exams (eg Pitmans)
(B) Touch-typist, speed unknown but higher than 40 words per minute
(C) Touch-typist, speed under 40 words per minute
(D) Know where the keys are, but use two fingers to type
(E) "Hunt-and-peck" typist with some keyboard/typewriter familiarity
(F) Little or no prior use of keyboards or typewriters

A B C D E F (please circle one)

8. Have you ever used a computer "**numeric pad**"?

(A) Yes, I use one at home/work
(B) Yes, but only occasional use
(C) Yes, maybe once or twice
(D) No, never used before

A B C D (please circle one)

9. Have you ever used a computer "**mouse**"?

(A) Yes, I use one at home/work
(B) Yes, but only occasional use
(C) Yes, maybe once or twice
(D) No, never used before

A B C D (please circle one)

10. How **familiar** are you with a **computer**?

(A) Have one at home/work:.....use it at least 4 times a week
(B) Have one at home/work: use up to 4 times a week
(C) A friend has one:.....use it at least 4 times a week
(D) A friend has one: use up to 4 times a week
(E) Use one at least 3 times a month
(F) Use one under 3 times a month
(G) Have used a computer consistently before, but not now
(H) Have used a computer slightly, but not now
(I) Have never used a computer

A B C D E F G H I

(please circle one)

Questionnaire for ACER-BL Study 2. (Questionnaire 2.2)

Student ID No. _____

Age: _____ _____ Today's Date: _____
(*Yrs*) (*Mths*)

While you are answering this questionnaire, I would like you to think carefully about how you were feeling throughout the ACER-BL test administration. Look at the answers below each question, and decide which option **most accurately** describes how you felt at that particular time.

1. How anxious were you **before** you sat this test?

(1) not at all anxious
(2) slightly anxious.....barely noticeable and performance not affected
(3) slightly anxious quite noticeable but performance not affected
(4) moderately anxious.....performance not affected
(5) moderately anxious performance may have been affected
(6) very anxious.....performance may have been affected
(7) very anxious performance was affected

1	2	3	4	5	6	7

(please circle one)

Why?

_____

_____

2. How anxious were you **after** you completed the practice questions, but **before** you sat this test?

(1) not at all anxious
(2) slightly anxious
(3) moderately anxious
(4) very anxious

1	2	3	4

(please circle one)

Why?

_____

3. How anxious were you **while** you sat this test?

(1) not at all anxious
(2) slightly anxious.....barely noticeable and performance not affected
(3) slightly anxious quite noticeable but performance not affected
(4) moderately anxious.....performance not affected
(5) moderately anxious performance may have been affected
(6) very anxious.....performance may have been affected
(7) very anxious performance was affected

1	2	3	4	5	6	7

(please circle one)

Why?_____

_____

_____

4. What was your level of anxiety for the following **parts** of the test?

(1) not at all anxious
(2) slightly anxious.....barely noticeable and performance not affected
(3) slightly anxious quite noticeable but performance not affected
(4) moderately anxious.....performance not affected
(5) moderately anxious performance may have been affected
(6) very anxious.....performance may have been affected
(7) very anxious performance was affected

A. The first 10 questions of the ACER-BL

1	2	3	4	5	6	7

(please circle one)

B. The next 10 questions of the ACER-BL

1	2	3	4	5	6	7

(please circle one)

C. The last 10 questions of the ACER-BL

1	2	3	4	5	6	7

(please circle one)

5. What was your level of anxiety for the following **types** of question?

(1) not at all anxious
(2) slightly anxious.....barely noticeable and performance not affected
(3) slightly anxious quite noticeable but performance not affected
(4) moderately anxious.....performance not affected
(5) moderately anxious performance may have been affected
(6) very anxious.....performance may have been affected
(7) very anxious performance was affected

A. The analogies, eg dolphin is to submarine as sparrow is to...

1	2	3	4	5	6	7

(please circle one)

B. The synonyms, eg find the word that means most nearly the same as peculiar...

1	2	3	4	5	6	7

(please circle one)

C. The antonyms, eg which two of the following words are opposite in meaning...

1	2	3	4	5	6	7

(please circle one)

D. The premises/conclusions, eg a dog that has bitten a child should not be poisoned because the child will then be poisoned. This statement is illogical because...

1	2	3	4	5	6	7

(please circle one)

6. Which type of ACER-BL input **format** have you just completed?

(A) Keyboard input
(B) Numeric pad input
(C) Mouse input

A B C (please circle one)

7. Did ACER-BL input format (keyboard, numeric pad, or mouse) **influence** your test anxiety?

(1) Yes.....adversely and strongly
(2) Yes adversely and moderately
(3) Yes.....adversely and slightly
(4) No influence
(5) Yes.....positively and slightly
(6) Yes positively and moderately
(7) Yes.....positively and strongly.

1	2	3	4	5	6	7

(please circle one)

8. What **other factors** do you think influenced your test anxiety on the ACER-BL?

(A) The ACER-BL time limit.....Yes / No
(B) Difficulty in reading ACER-BL items Yes / No
(C) Ease in reading ACER-BL items.....Yes / No
(D) Difficulty in understanding ACER-BL layout Yes / No
(E) Ease in understanding ACER-BL layout.....Yes / No
(F) Entering incorrect response due to unfamiliarity with input device .....Yes / No
(G) Entering incorrect, but within range, response due to unfamiliarity with input device.....Yes / No

9. What **changes** do you think could be made to the presentation of the ACER-BL in order to **reduce** test-taker anxiety? (eg page layout, screen layout)

_____

_____

10. Have you ever completed a questionnaire with a **similar format** to the ACER-BL, eg PAT test, SAT test, personality test?

Yes / No

11. Have you ever completed a **computer-presented** questionnaire with a similar format to the ACER-BL?

Yes / No

If yes, what was the **name** or **type** of test?

_____

Questionnaire for ACER-BL Study 2. (Questionnaire 2.3)

Student ID No._____

Age: _____ _____ Today's Date: _____
(*Yrs*) (*Mths*)

While you are answering this questionnaire, I would like you to think carefully about how you were feeling throughout the ACER-BL test administration. Look at the answers below each question, and decide which option **most accurately** describes how you felt at that particular time.

1. How anxious were you **before** you sat this test?

(1) not at all anxious
(2) slightly anxious.....barely noticeable and performance not affected
(3) slightly anxious quite noticeable but performance not affected
(4) moderately anxious.....performance not affected
(5) moderately anxious performance may have been affected
(6) very anxious.....performance may have been affected
(7) very anxious performance was affected

1	2	3	4	5	6	7

(please circle one)

Why?

_____

_____

2. How anxious were you **after** you completed the practice questions, but **before** you sat this test?

(1) not at all anxious
(2) slightly anxious
(3) moderately anxious
(4) very anxious

1	2	3	4

(please circle one)

Why?

_____

3. How anxious were you **while** you sat this test?

(1) not at all anxious
(2) slightly anxious.....barely noticeable and performance not affected
(3) slightly anxious quite noticeable but performance not affected
(4) moderately anxious.....performance not affected
(5) moderately anxious performance may have been affected
(6) very anxious.....performance may have been affected
(7) very anxious performance was affected

1	2	3	4	5	6	7

(please circle one)

Why?_____

_____

_____

4. What was your level of anxiety for the following **parts** of the test?

(1) not at all anxious
(2) slightly anxious.....barely noticeable and performance not affected
(3) slightly anxious quite noticeable but performance not affected
(4) moderately anxious.....performance not affected
(5) moderately anxious performance may have been affected
(6) very anxious.....performance may have been affected
(7) very anxious performance was affected

A. The first 10 questions of the ACER-BL

1	2	3	4	5	6	7

(please circle one)

B. The next 10 questions of the ACER-BL

1	2	3	4	5	6	7

(please circle one)

C. The last 10 questions of the ACER-BL

1	2	3	4	5	6	7

(please circle one)

5. What was your level of anxiety for the following **types** of question?

(1) not at all anxious
(2) slightly anxious.....barely noticeable and performance not affected
(3) slightly anxious quite noticeable but performance not affected
(4) moderately anxious.....performance not affected
(5) moderately anxious performance may have been affected
(6) very anxious.....performance may have been affected
(7) very anxious performance was affected

A. The analogies, eg dolphin is to submarine as sparrow is to...

1	2	3	4	5	6	7

(please circle one)

B. The synonyms, eg find the word that means most nearly the same as peculiar...

1	2	3	4	5	6	7

(please circle one)

C. The antonyms, eg which two of the following words are opposite in meaning...

1	2	3	4	5	6	7

(please circle one)

D. The premises/conclusions, eg a dog that has bitten a child should not be poisoned because the child will then be poisoned. This statement is illogical because...

1	2	3	4	5	6	7

(please circle one)

6. Which type of ACER-BL input **format** have you just completed?

(A) Keyboard input
(B) Numeric pad input
(C) Mouse input

A B C (please circle one)

7. Did ACER-BL input format (keyboard, numeric pad, or mouse) **influence** your test anxiety?

(1) Yes.....adversely and strongly
(2) Yes adversely and moderately
(3) Yes.....adversely and slightly
(4) No influence
(5) Yes.....positively and slightly
(6) Yes positively and moderately
(7) Yes.....positively and strongly.

1	2	3	4	5	6	7

(please circle one)

8. What **other factors** do you think influenced your test anxiety on the ACER-BL?

(A) The ACER-BL time limit.....Yes / No
(B) Difficulty in reading ACER-BL items Yes / No
(C) Ease in reading ACER-BL items.....Yes / No
(D) Difficulty in understanding ACER-BL layout Yes / No
(E) Ease in understanding ACER-BL layout.....Yes / No
(F) Entering incorrect response due to unfamiliarity with input device .....Yes / No
(G) Entering incorrect, but within range, response due to unfamiliarity with input device.....Yes / No

9. What **changes** do you think could be made to the presentation of the ACER-BL in order to **reduce** test-taker anxiety? (eg page layout, screen layout)

_____

_____

10. Have you ever completed a questionnaire with a **similar format** to the ACER-BL, eg PAT test, SAT test, personality test?

Yes / No

11. Have you ever completed a **computer-presented** questionnaire with a similar format to the ACER-BL?

Yes / No

If yes, what was the **name** or **type** of test?

_____

Questionnaire for ACER-BL Study 2. (Questionnaire 2.4)

Student ID No._____

Age: _____ _____ Today's Date: _____
(*Yrs*) (*Mths*)

While you are answering this questionnaire, I would like you to think carefully about how you were feeling throughout the ACER-BL test administration. Look at the answers below each question, and decide which option **most accurately** describes how you felt at that particular time.

1. How anxious were you **before** you sat this test?

(1) not at all anxious
(2) slightly anxious.....barely noticeable and performance not affected
(3) slightly anxious quite noticeable but performance not affected
(4) moderately anxious.....performance not affected
(5) moderately anxious performance may have been affected
(6) very anxious.....performance may have been affected
(7) very anxious performance was affected

1	2	3	4	5	6	7

(please circle one)

Why?

_____

_____

2. How anxious were you **after** you completed the practice questions, but **before** you sat this test?

(1) not at all anxious
(2) slightly anxious
(3) moderately anxious
(4) very anxious

1	2	3	4

(please circle one)

Why?

_____

3. How anxious were you **while** you sat this test?

(1) not at all anxious
(2) slightly anxious.....barely noticeable and performance not affected
(3) slightly anxious quite noticeable but performance not affected
(4) moderately anxious.....performance not affected
(5) moderately anxious performance may have been affected
(6) very anxious.....performance may have been affected
(7) very anxious performance was affected

1	2	3	4	5	6	7

(please circle one)

Why?_____

_____

_____

4. What was your level of anxiety for the following **parts** of the test?

(1) not at all anxious
(2) slightly anxious.....barely noticeable and performance not affected
(3) slightly anxious quite noticeable but performance not affected
(4) moderately anxious.....performance not affected
(5) moderately anxious performance may have been affected
(6) very anxious.....performance may have been affected
(7) very anxious performance was affected

A. The first 10 questions of the ACER-BL

1	2	3	4	5	6	7

(please circle one)

B. The next 10 questions of the ACER-BL

1	2	3	4	5	6	7

(please circle one)

C. The last 10 questions of the ACER-BL

1	2	3	4	5	6	7

(please circle one)

5. What was your level of anxiety for the following **types** of question?

(1) not at all anxious
(2) slightly anxious.....barely noticeable and performance not affected
(3) slightly anxious quite noticeable but performance not affected
(4) moderately anxious.....performance not affected
(5) moderately anxious performance may have been affected
(6) very anxious.....performance may have been affected
(7) very anxious performance was affected

A. The analogies, eg dolphin is to submarine as sparrow is to...

1	2	3	4	5	6	7

(please circle one)

B. The synonyms, eg find the word that means most nearly the same as peculiar...

1	2	3	4	5	6	7

(please circle one)

C. The antonyms, eg which two of the following words are opposite in meaning...

1	2	3	4	5	6	7

(please circle one)

D. The premises/conclusions, eg a dog that has bitten a child should not be poisoned because the child will then be poisoned. This statement is illogical because...

1	2	3	4	5	6	7

(please circle one)

6. Which type of ACER-BL input **format** have you just completed?

(A) Keyboard input
(B) Numeric pad input
(C) Mouse input

A B C (please circle one)

7. Did ACER-BL input format (keyboard, numeric pad, or mouse) **influence** your test anxiety?

(1) Yes.....adversely and strongly
(2) Yes adversely and moderately
(3) Yes.....adversely and slightly
(4) No influence
(5) Yes.....positively and slightly
(6) Yes positively and moderately
(7) Yes.....positively and strongly.

1	2	3	4	5	6	7

(please circle one)

8. What **other factors** do you think influenced your test anxiety on the ACER-BL?

(A) The ACER-BL time limit.....Yes / No
(B) Difficulty in reading ACER-BL items Yes / No
(C) Ease in reading ACER-BL items.....Yes / No
(D) Difficulty in understanding ACER-BL layout Yes / No
(E) Ease in understanding ACER-BL layout.....Yes / No
(F) Entering incorrect response due to unfamiliarity with input device .....Yes / No
(G) Entering incorrect, but within range, response due to unfamiliarity with input device.....Yes / No

9. What **changes** do you think could be made to the presentation of the ACER-BL in order to **reduce** test-taker anxiety? (eg page layout, screen layout)

_____

_____

10. Have you ever completed a questionnaire with a **similar format** to the ACER-BL, eg PAT test, SAT test, personality test?

Yes / No

11. Have you ever completed a **computer-presented** questionnaire with a similar format to the ACER-BL?

Yes / No

If yes, what was the **name** or **type** of test?

_____

12. What was the **order** of **input devices** you used for the ACER-BL test?

(A) Keyboard / numeric pad / mouse
(B) Keyboard / mouse / numeric pad
(C) Numeric pad / mouse / keyboard
(D) Numeric pad / keyboard / mouse
(E) Mouse / keyboard / numeric pad
(F) Mouse / numeric pad / keyboard

A B C D E F (please circle one)

13. Which ACER-BL administration did you feel **most anxious** on?

(A) First
(B) Second
(C) Third
(D) No difference

A B C D (please circle one)

14. Which ACER-BL test administration did you feel **least anxious** on?

(A) First
(B) Second
(C) Third
(D) No difference

A B C D (please circle one)

15. Did you perceive any **eyestrain** during ACER-BL test administration?

Yes / No (please circle one)

16. During which ACER-BL test administration did you feel the **most eyestrain**?

(A) First
(B) Second
(C) Third
(D) No difference

A B C D (please circle one)

17. During which ACER-BL test administration did you feel the **least eyestrain**?

(A) First
(B) Second
(C) Third
(D) No difference

A B C D (please circle one)

Appendix E: Participant Characteristics Interactions With Input Device ACER-BL Test Score

Table E1. Participant characteristics interactions with ACER-BL score for each input device. Results of repeated measures ANOVAs.

Characteristic		Keyboard	Numeric Pad	Mouse	F-value	P (2-tailed)
Age	M	21.26	21.34	21.18	0.32	.73
	SD	3.54	3.56	3.56		
Gender	M	21.35	21.41	21.25	0.66	.52
	SD	3.47	3.53	3.52		
Ethnicity	M	21.29	21.36	21.24	0.16	.85
	SD	3.56	3.58	3.52		
Family Income	M	21.27	21.34	21.18	0.48	.62
	SD	3.54	3.57	3.56		
Eyesight	M	21.27	21.34	21.18	0.10	.91
	SD	3.54	3.57	3.56		
Mathematical ability	M	21.24	21.32	21.15	0.12	.88
	SD	3.55	3.58	3.56		
Statistical ability	M	21.23	21.32	21.15	1.95	.15
	SD	3.55	3.58	3.56		
English ability	M	21.24	21.32	21.15	0.76	.47
	SD	3.55	3.58	3.56		
Programming ability	M	21.27	21.34	21.18	0.38	.68
	SD	3.54	3.57	3.56		
Year at university	M	21.27	21.34	21.18	0.48	.62
	SD	3.54	3.57	3.56		
Typing ability	M	21.27	21.34	21.18	0.85	.43
	SD	3.54	3.57	3.56		
Numeric pad familiarity	M	21.27	21.34	21.18	0.17	.85
	SD	3.54	3.57	3.56		
Mouse familiarity	M	21.27	21.34	21.18	0.49	.61
	SD	3.54	3.57	3.56		
Computer familiarity	M	21.27	21.34	21.18	0.79	.46
	SD	3.54	3.57	3.56		