THE HENMON-NELSON: COMPUTERIZED

by

DWIGHT DAVID HARLEY

B.Ed., University of British Columbia, 1977

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES

(Department of Educational Psychology)

We accept this thesis as conforming to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

June, 1979

© Dwight David Harley, 1979
In presenting this thesis in partial fulfillment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the Head of my Department or by his representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Educational Psychology
The University of British Columbia
2075 Wesbrook Place
Vancouver, Canada
V6T 1W5

June, 1979.
Abstract

Computerized test administration is a recent application of computer technology to the art of testing that is just now coming into vogue. It was the purpose of this study to determine what effect, if any, computerized administration had on the scores obtained from the Henmon-Nelson Test of Mental Ability, level 4. An interactive computer program was designed to administer this test.

Computerization of standardized psychological tests such as the Slosson Intelligence Test, the Minnesota Multiphasic Personality Inventory, and the California Achievement Test has demonstrated the technical feasibility of computerized testing, but literature searches have shown that limited research has been done with these programs.

This study followed the classical, single group, two trial, repeated measure design. Each of the 36 subjects wrote one of the two parallel forms of the Henmon Nelson in the conventional manner and responded to the other form through the computer. The order and medium by which each subject responded to the different forms was randomized to control for such factors as fatigue and pretest sensitization.

Dependent t-test statistics were used to compare (i) the mean written verbal score with the mean computerized verbal score; (ii) the mean written quantitative score with the mean computerized quantitative score; and (iii) the mean written total score with the mean computerized total score. All three
t-test statistics were found to be nonsignificant, indicating that the null hypotheses should not be rejected. These results suggested that computerized administration of the Henmon-Nelson can provide results comparable to those obtained from a conventional administration.

Research Supervisor: Dr. H. Ratzlaff
## TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 1 - The Problem</td>
<td>1</td>
</tr>
<tr>
<td>Chapter 2 - The Literature</td>
<td>9</td>
</tr>
<tr>
<td>Chapter 3 - Methodology</td>
<td>23</td>
</tr>
<tr>
<td>Chapter 4 - Results</td>
<td>34</td>
</tr>
<tr>
<td>Chapter 5 - Discussion</td>
<td>46</td>
</tr>
<tr>
<td>Bibliography</td>
<td>54</td>
</tr>
<tr>
<td>Appendix A - Demographic Questionaire</td>
<td>56</td>
</tr>
<tr>
<td>Appendix B - Listing Of Computer Program</td>
<td>58</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table I - Split-half Reliabilities (Henmon-Nelson) ..... 19
Table II - Coefficients Of Equivalence (Henmon-Nelson) .. 20
Table III - Standard Errors Of Measurement (Henmon-Nelson)

Table IV - Possible Testing Assignments ................. 28
Table V - Frequency Of Gender ................................ 34
Table VI - Frequency Of Age .................................. 34
Table VII - Frequency Of Major ............................... 35
Table VIII - Frequency Of Year Of Enrollment ............. 35
Table IX - Frequency Of Prior Terminal Use ............... 35
Table X - Means, Standard Deviations And Standard Errors

Table XI - Proportional Means ................................ 37
Table XII - Post Hoc Means And (Standard Deviations) ..... 40
Table XIII - Post Hoc Summary .............................. 41
LIST OF FIGURES

Figure 1 - Two-trial Repeated Measure Design .................. 26
Acknowledgement

I wish to extend my appreciation to Garry Forwood for his technical expertise in the design and implementation of the computer program used for the administration of the Henmon Nelson.

I would like to express my gratitude to my research supervisor Dr. H. Ratzlaff, and committee members Dr. O. Oldridge and Dr. M. Westrom, for their help and counsel.

I wish to thank my parents David and June Harley, for all their help, support, and encouragement throughout the course of my studies.

Finally to my wife Margaret, my sincere thanks for her understanding and encouragement which were necessary in completing this study.
CHAPTER 1

The Problem

Computer technology was first introduced to the testing scene during the early 1950's. Computers were used primarily to replace clerks, and their scoring machines, who could mark and record only a few hundred test forms in a one hour period. The computers that were available during this period produced a name list (identification list) with corresponding scores and a few summary statistics such as the mean, mode, and standard deviation.

As technology advanced, so did the role of computers in testing. Computers now produce not only a list of names and scores but comprehensive statistical distributions and analyses during the time that the answer sheets are scanned and scored. With the use of optical scanners for sensing the response sheets, and computers for scoring and processing, the time required for test analysis has been decreased by well over 100 times.

However, in the testing environment computers can perform tasks other than those related to data analysis and retrieval from data supplied to them by answer sheets. More recently computers are being used in the preparation of the test itself. Test preparation can be supported by computers in a variety of ways ranging from item banking and item generation to item selection and test printing. Considering the information
processing capabilities of present day computers, CATC\(^1\) is a most natural development in the art of testing. (Lippey, 1974)

To date computers have been used in test preparation, scoring, and analysis. Each of these tasks requires clerical operations and straightforward decisions. These tasks are algorithmic, and hence machine executable. Another task which is algorithmic in nature as well as a logical extension of computer technology into the realm of testing is computerized-test-administration.

In order to reduce the differential effects of the testing environment, examiner's manuals go to great lengths in an attempt to achieve uniformity and control over the procedures to be followed in the administration of tests. A better way to ensure uniformity is to program these procedures and have the computer follow this process in the administration of the test. Computerized test administration greatly reduces the effect of the environment on the testing situation. It is with this concept of computerized test administration that this study dealt.

1. The Problem

Computerized test administration is a recent application of computer technology to the area of testing that is just now coming into vogue. It has been utilized, to a large extent, in

\(^1\) Computer Assisted Test Construction
conjunction with computerized adaptive testing (Ferguson, 1969). Computerized adaptive testing (CAT) is a form of testing during which the testee responds to a branching or programmed test interactively through the use of a conversational terminal. Research based on the theoretical and empirical aspects of CAT has so far been inconclusive, but psychometric research has proved promising and has encouraged further work in this area.

However, most if not all research done with CAT, as well as with other applications of computerized test administration, has apparently glossed over one not-so-minor detail. That is, does the testee, using a computer terminal to respond to a test, achieve the same score as he would responding to the same test in the conventional manner? This was the general problem to which the study addressed itself.

A problem of this dimension is clearly unanswerable in the context of just one study. Considering all the types of tests presently in use as well as the number of conventional response forms, the researcher is compelled to reduce the problem to a manageable size by selecting one test and examining the effect that computerized administration has on that particular instrument.

In keeping with this line of thought, this study attempted to determine what effect, if any, computerized administration had on the scores obtained from the Henmon-Helson Test of Mental Ability, level 4 (Nelson, Lamke, and Keslo, 1961).
2. Definition Of Terms

The following is a collection of key terms accompanied by definitions that indicate the precise sense in which these terms are used in this study.

a) **Computer administration**: presenting test stimuli and accepting responses to these stimuli from testees via conversational terminals.

b) **Conventional administration**: presenting test stimuli and accepting responses to these stimuli from testees in the medium (pencil and paper, oral, etc.) by which the test was originally intended to be administered.

c) **Conversational terminal**: a typewriter like device, such as the Ann Arbor terminals which were used in this study, with which the user can communicate interactively, in terms of real time, with a time sharing computer system. In the case of the Ann Arbor terminals, information was displayed to the user on a cathode-ray-tube (CRT) screen.

d) **Page mode**: a terminal mode by which entire "screens-full" of information are displayed on the terminals CRT screen. These "pages" are changed in their entirety, paralleling the action of turning a page in a book. Pages may be redisplayed as many times as desired in the exact format as they were originally.

e) **Scores**:

   i) **quantitative**: the number of correct responses given on the quantitative subtest of the Henmon-Nelson Test of Mental Ability, level 4.

   ii) **verbal**: the number of correct responses given on the verbal subtest of the Henmon-Nelson Test of Mental Ability,
level 4.

iii) total: the sum of the quantitative and verbal scores on the Henmon-Nelson Test of Mental Ability, level 4.

f) Written form: the Henmon-Nelson Test of Mental Ability, level 4, administered and taken in the conventional, paper and pencil manner.

3. Research Questions

Given that the problem of the study was to determine whether or not computerized administration of the Henmon-Nelson Test of Mental Ability, level 4, has any effect on the obtained scores, three research questions emerged.

1. Is there a significant difference between the average raw quantitative score obtained from a computerized administration of the Henmon-Nelson Test of Mental Ability, level 4 and the average raw quantitative score obtained from a conventional administration of the test?

2. Is there a significant difference between the average raw verbal score obtained from a computerized administration of the Henmon-Nelson Test of Mental Ability, level 4 and the average raw verbal score obtained from a conventional administration of the test?

3. Is there a significant difference between the
average raw total score obtained from a computerized administration of the Henmon-Nelson Test of Mental Ability, level 4 and the average raw total score obtained from a conventional administration of the test?

Post hoc analyses were carried out with respect to other variables of interest, such as sex and age. Analyses of this nature took the form of repeated measure ANOVAs with one grouping factor.

A five percent significance level was used due to the exploratory nature of this study. It was intended to show that comparable results can be obtained through computer administration.

4. Delimitation of the Study

The study was carried out using student volunteers enrolled in first year university through first year graduate school at the University of British Columbia.

The study employed only Ann Arbor CRT terminals which are supported by a PDP-11/40 mini-computer attached to an Amdahl 470 V/6, Model II computer. The use of other types of terminals available at the University of British Columbia's computer installation would have necessitated complex I/O (input/output procedures) format changes in the computer program which administered the test.

In the context of the global problem concerning tests as a
whole, it is only possible to generalize the findings of the study to other tests in so far as they are similar to the Henmon-Nelson Test of Mental Ability, level 4.

When comparing tests to the Henmon-Nelson for the purposes of extending any generalizations of the findings, it is important to note that the Henmon-Nelson consists totally of multiple choice formatted responses to verbal and quantitative stimuli. Extending any generalizations to include true and false responses should not present a problem as true and false is a modified version of the multiple choice format. However, matching response formats are entirely different and should become the subject of other research.

5. Justification

Justification for this study was best seen in the light of the general problem. Solution of the immediate problem was only of minor practical importance in itself. However, as a basis from which to extend future research into the interface between testees and computer associated devices used for testing, this study was of significance.

When it is demonstrated that comparable results can be obtained from computer administered tests and their corresponding written forms, the entire testing environment can be revolutionized into a more efficient and practical process. As discussed previously, computer administered testing allows for the standardization of the administrative procedures. Having testees respond directly to a computer eliminates
clerical errors in transposing answer sheets into a machine readable form. Also using computers to score the tests further reduces clerical errors and achieves greater objectivity in scoring. The problem of test security is reduced since there will not be test booklets and answer sheets to lose or misplace. However this problem can not be entirely eliminated since it now becomes confounded with the new problem of computer security.

Computer administered testing eliminates the examiner effect as well as problems involved with test materials, printing, storage, distribution, and answer sheet processing. Computer administered testing should greatly increase the efficiency of the testing process.

Logistical advantages are not the sole offerings of computerized test administration. Computers used in this fashion enable us to obtain information that would be impossible or impractical to collect using the conventional modes of response (e.g. item response latency; the amount of time in which it takes the testee to answer each item). Other examples include the number of times pages are viewed, frequency of changed responses per item, a list of previous responses, and so on. These kinds of information will allow for further empirical research to be done in the area of response style.

Clearly there is a need for automation of the testing process. The following chapter examines the work done in the area of computerized testing and provides a critique of the Henmon-Nelson Test of Mental Ability, level 4.
CHAPTER 2

The Literature

In the past few years, computer programs have been developed which allow for the administration of standardized psychological tests. The purpose of this study was to determine whether performance on the Henmon-Nelson Test of Mental Ability, level 4, was affected by this new mode of presentation. In order to provide a foundation for this study, the literature review will cover the following areas: (1) automated testing; (2) computerized testing; and (3) Henmon-Nelson Test of Mental Ability, level 4.

1. Automated Testing

Automated testing is the forerunner of computerized test administration. It enjoyed the same advantages over conventional testing that computerized administration now enjoys. These advantages include greater standardization of test administration procedures, elimination of clerical errors encountered in the recording and scoring of responses, as well as the capability of gathering data regarding response styles that would otherwise be impractical to collect.

One of the earliest demonstrations of the usefulness of a technological approach to test administration was made by Elwood (1969) who developed an automated testing booth programmed to administer the Wechsler Adult Intelligence Scale (WAIS). The device administered all performance and verbal subtests of the WAIS and recorded each response made to the individual test.
items. Two subtests were automatically scored and response style data collected.

Automated testing is a mechanical approach to test administration. Elwood's booth consisted of a portable sound proof room containing a response panel and two cases of twelve drawers. The drawers contained the materials for the various subtests of the WAIS. Each drawer was spring loaded, solenoid operated, and remotely controlled by an eight-track paper tape reader and a BSR electronic digital logic system located outside the booth. The BRS system recorded the amount of time the subject spent on each item. Subjects were taken into the booth by a research assistant who explained the operations of the booth. Subjects were then given practice at pressing the "READY" button which opened a demonstration drawer containing materials which gave the subject experience at working with the test pieces. When the drawer was replaced the research assistant would leave the booth and the subject would press the "READY" button to begin the test. Research assistants scored the subjects' performance by inspecting the contents of the drawers and noting the time print-out received from the teletype which was attached to the BRS logic system. Scoring standards described in the WAIS manual were followed.

Investigations into the test-retest reliability of Elwood's automated WAIS were carried out by Griffith and Elwood (cited in Hedl et al., 1969). Verbal subscore reliability was estimated at $r=0.99$ while the reliability of the performance scale was $r=0.97$. Reliability of the full scale was approximately $r=0.98$. 
A comparison of the automated WAIS to the traditional WAIS was made by Orr (cited in Hitti et al., 1969). He found reliabilities of $r=0.94$, $r=0.82$, and $r=0.93$ for the verbal, performance, and full scale scores respectively.

Such was the extent of Elwood's automated WAIS in the early 1970's. Exceptional results were obtained in the related studies, but problems which reduced the devices efficiency existed. Only half of the testing process was automated. Psychologists would still have to score the test which was a costly procedure both in terms of time and labour. Clearly the next goal for researchers in automated testing should be the complete and total automation of the entire testing process. The time was now right for the union of computer technology and the procedures of test administration.

2. Computerized Testing

One of the first attempts at completely automating a standardized psychological test was the computerization of the Slosson Intelligence Test (SIT). Hedl, O'Neil, and Hansen (1971) designed a computer program to administer the SIT to subjects via CRT-graphic type terminals. The program recorded all student responses and response latencies.

In taking the test the subject first read the item which was displayed on the CRT terminals and then responded by typing on the terminal's keyboard. The computer then evaluated the response as totally correct, totally incorrect, or partially correct. If the response was partially correct the subject was
then prompted by the computer to elaborate. The subject elaborated in the same way by typing his response into the computer. The procedure followed by the computer in the administration of the SIT parallels that of the procedure suggested by the authors of the SIT.

All scoring was done automatically by the computer. Verbal responses to the test were scored through the use of algorithms. Each response was scanned for combinations of keywords which were matched to lists stored in the machine. If certain combinations of words appeared in the response, it was scored as correct while other combinations of words constituted an incorrect response. Quantitative scores were matched against answer sets which were also stored on disk in the computer.

In the testing of the scoring mechanisms of the program, 25 college students wrote the computer based version of the SIT. The computer calculated a score and produced an exact copy of the dialogue carried on between it and the subject. This dialogue was then given to psychologists to evaluate and score. The list of 25 pairs of scores (computer score and psychologist score) was then correlated and a Pearson r of 0.83 was obtained. This would seem to indicate that the scoring procedures of the SIT program were reasonably reliable.

The study went on to determine whether computer administration of the SIT provided results equivalent to results from conventional administration. An attempt was also made to compare the results of the computer-based SIT to the WAIS. Using a 3x3 latin square (counterbalanced) design (n=48)
subjects were tested with the WAIS, the computer version of the SIT, and the conventional SIT. There was a one week interval between each of the three testings. Random assignment was used to control for examiner effects. The Pearson product-moment correlation coefficient between the scores of the computer-based SIT and the scores of the conventional SIT was 0.75, significant at the .01 level. Non-significant Pearson correlations were found between the computerized SIT and the WAIS (r=0.54) as well as between the conventional SIT and the conventional WAIS (r=0.52).

It should be noted that Pearson's correlation between the automated WAIS and the conventional WAIS (r=0.93) is much greater than the correlation between the computerized SIT and the conventional SIT (r=0.75). This is in part due to the fact that both forms of the WAIS were hand scored while one form of the SIT was scored by hand and the other form scored by computer. By comparison, the Pearson r between hand scoring of the computer based SIT and computer scoring of the computer based SIT was 0.83. It should also be noted that the test-retest reliability coefficient of the WAIS was 0.96.

Automated testing has not been limited solely to individualized tests. Group tests have also been computerized. As of 1973 Hoye and Wang (1973) report that computerized versions of the California Achievement Test, the Minnesota Multiphasic Personality Inventory (MMPI), and the Kuder Vocational Interests Inventory have been developed. Lekan (1970) indicates that these computerized tests were available
prior to 1970. However literature searches have shown that limited research has been done with these programs. One paper (Lushene, O'Neil, and Dunn 1972) did state that correlations between the scores of the computerized MMPI and the scores of the conventional MMPI were as high as those obtained from test-retest or parallel forms. Empirical data confirming this was not reported.

More recently (Hitti, Riffer, and Stuckless, 1972) the Ravens Progressive Matricies Test (RPMT), a test of non-verbal general aptitude, was computerized and tested on a sample of 76 deaf, post secondary students. One of the objectives of the study was to determine if performance was affected by mode of test presentation. Since there is a sex bias on the RPMT a stratified random process was used to assign the students to the two groups. One group was tested under computerized conditions while the second group was tested in the usual fashion. The analysis showed that there was no significant difference on performance with respect to the mode of presentation. However a significant ($\alpha=.05$) sex difference was found, but this was consistent with the data already available on the RPMT. A significant interaction effect was not found, even though visual inspection of the means would suggest one.

Thus, studies have demonstrated the feasibility of using computers to administer tests. Results so far have been encouraging but it still remains to be confirmed whether computer administered tests provide comparable results to those obtained under conventional testing.
3. Henmon Nelson Test Of Mental Ability

The Henmon-Nelson Test of Mental Ability is a timed, group administered test which is designed to "measure those aspects of mental ability which are important for success in academic work and in similar endeavors outside the classroom."(Nelson et al., 1961)

Four levels of the test have been developed to measure the range of ability found in grades K through first year graduate school. The interest of the present study is limited to the uppermost level which is appropriate for grades 13 to graduate school.

The test package consists of two test forms, A and B, along with separate answer sheets. A manual appropriate for both forms is also supplied. Three kinds of answer forms can be used with Henmon-Nelson: IBM 805 and IBM 1230 machine or hand scoring answer sheets and a self-scoring answer sheet which eliminates the need for a scoring key.

Administration of the test is simple. Special training is not required, however familiarization with the examiner's manual is necessary to obtain more reliable measures.

The authors feel that their test, as well as all other group mental ability tests, should be categorized as Level B as defined by the American Psychological Association's Code of Standards for Test Distribution. This would require the administrators of such tests to be well versed in educational measurement, statistics, guidance, as well as human learning and
development. However, the authors do concede that reasonable results can be obtained by a competent instructor who carefully follows the instructions provided in the examiner's manual.

Raw score is defined as the total number of correct responses with the underlying scale continuous and interval in nature. Raw scores for the quantitative and verbal subsets are determined with the total score seen as the sum of the two subtest scores. A table is provided to convert each of the three raw scores to percentile ranks. These norms were gathered on college freshmen who were tested early in the school year. Norms for other subgroups of the target population have not been provided. The authors attempt to justify this omission by claiming that local norms are more useful than national norms since the caliber of student attracted to specific institutions varies. This may be true, but it renders the test less than useful, unless the user is prepared to go to the trouble of establishing his own norms.

Care should be taken when interpreting percentile rank scores. The user should be aware that percentile ranks are not equal. They inflate raw score differences near the center of the distribution and reduce differences toward the extremes. It must be noted that percentile ranks should be used only as an ordinal scale rather than as an interval scale. The manual includes one brief paragraph regarding the interpretation of the percentile tables, but cautions regarding over-interpretation of these scores are not made. Inclusion of equivalent stanine scores would help alleviate the problems encountered when
working with percentile ranks. They would also help avoid a single score from being mis-interpreted.

The examiner's manual is well-designed with particular attention given to the clarity of instructions as well as to the possible uses of the Henmon-Nelson tests. This should make the test very useful to those unfamiliar with standardized testing, once reliable norms have been obtained. The accompanying test booklets had a somewhat crowded format.

The test was constructed from a pool of 400 items which had been extensively pilot tested. Most of the items were taken from the Iowa Test of Mental Ability and the previous edition of the Henmon-Nelson. This pool was again tested and the results analyzed with respect to difficulty index, variance, and contribution to overall reliability. By analyzing these statistics and the content of each item, two parallel forms were constructed.

A variety of different styles of answer sheets were used during the pilot testing. Each type of response form was uniformly distributed throughout the population. The manual does not provide any statistics dealing with the effect of the different answer forms on the item or test results.

Sex bias of the resulting forms was investigated. Samples of 100 men and 100 women were randomly drawn from the norming population. At the ten percent alpha level, the means and standard deviations of groups across forms did not significantly differ.
The normative sample consisted of 2004 American college freshmen. 1002 subjects wrote each form of the Henmon-Nelson within the first two months of their first semester of college attendance. The sample was stratified by location, size, and type of institution such that it reflected the general college population at that time (Fall, 1958). As previously noted, the test was not normed for other subsets of the target population as the authors felt local norms would be more useful than national norms.

Data regarding the reading level of the college version of the Henmon-Nelson is not provided. A subjective examination of the test leads one to believe that it is heavily dependent upon reading skills. The manual provides correlational data between the Henmon-Nelson and the Nelson-Denny Reading Test. The correlation between total raw scores is approximately 0.78 (Pearson r). This empirically indicates that the Henmon-Nelson is dependent upon reading skills.

The test in general, is heavily biased toward the verbal factor of mental ability. Approximately 60% of the total number of items belong to the verbal scale whereas only 40% of the items are quantitative in nature. This affects the reliability of the quantitative scale and weights the test in favor of verbal ability. Also the two types of items are not uniformly distributed throughout the test. This obviously will affect the score of the slower subject who does not finish within the allotted time.
The Henmon-Nelson, as noted previously, is a speed test with a forty minute time limit. The examiner's manual does not report any data regarding the effect of time upon an individual's score. However the authors are quick to concede that speed will adversely affect the score of subjects with a slow response style.

Odd-even reliability coefficients are listed for both forms A and B. These values are given in Table I. Q, V, and Total represent the quantitative scales respectively. The manual is unclear as to whether or not these values have been corrected by the Spearman-Brown Prophecy formula. Computations were based upon a random sample of 100 subjects selected from the normative sample.

Table I - Split-Half Reliabilities (Henmon-Nelson) 

\[
\begin{array}{ccc}
N=100 & \text{Form A} & \text{Form B} \\
\hline 
Q & .921 & .889 \\
V & .925 & .934 \\
\text{Total} & .950 & .940 \\
\end{array}
\]

It should be noted that the split half method of estimating reliability leads to inflated values when applied to speed tests.

Alternate forms reliability estimates are given in Table II. This is a more realistic approach to obtaining reliability estimates of speed tests than by using the split-half method (Magnusson, 1967). Test forms A and B were administered to the same subjects with a time interval of 35 days between testings.
Justification is not given for the rather long interval between the successive testings. The retesting should have been carried out within a much shorter time period (e.g., less than 10 days) since longer periods allow for sources of invalidity such as contemporary history, maturation and experimental mortality to weaken the internal validity of the alternate forms study.

Table II - Coefficients of Equivalence (Henmon-Nelson)

<table>
<thead>
<tr>
<th></th>
<th>Form A</th>
<th>Form B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N=52 Mean</td>
<td>Mean</td>
</tr>
<tr>
<td>Q</td>
<td>19.58</td>
<td>23.46</td>
</tr>
<tr>
<td>V</td>
<td>28.50</td>
<td>28.69</td>
</tr>
<tr>
<td>Total</td>
<td>48.08</td>
<td>52.15</td>
</tr>
</tbody>
</table>

From Table II it can be seen that there is a four point raw score increase between form A and B which is most likely due to a practice effect on the quantitative subtest. This effect could have been controlled if a counterbalanced design would have been employed in the testing.

Coefficients of rationale-equivalence (internal consistency) reliability are not provided in the manual. These non-correlational estimates of reliability would have given some indication of the homogeniety of the Henmon-Nelson as a whole.

Standard errors of measurement are given in Table III. The manual provides an excellent section on the interpretation of this statistic.
Table III - Standard Errors of Measurement (Henmon-Nelson)

<table>
<thead>
<tr>
<th>N=100</th>
<th>Form A</th>
<th>Form B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q</td>
<td>1.98</td>
<td>2.40</td>
</tr>
<tr>
<td>V</td>
<td>3.05</td>
<td>2.86</td>
</tr>
<tr>
<td>Total</td>
<td>3.64</td>
<td>3.57</td>
</tr>
</tbody>
</table>

The authors of the Henmon-Nelson have attempted to establish criterion-related validity by correlating it with other measures of mental ability, such as grade point average (GPA), the Nelson-Denny Reading Test and other levels of the Henmon-Nelson tests. Correlations with GPA range from 0.417 (quantitative subscore) to 0.783 (total score). The correlations between other levels of the Henmon-Nelson and its college level range from 0.696 to 0.791. It is interesting to note that in all cases correlations between the quantitative score and the criterion score were lowest while correlations between the total score and the criterion were always the highest. Attempts to establish construct validity were not reported in the manual.

The Henmon-Nelson Test of Mental Ability is practical in format and easy to administer, score, and interpret. The test possesses good reliability and reasonable validity. This makes the test useful to classroom teachers and counselors who make decisions regarding academic considerations for a variety of pupils.

The technical feasibility of computerized testing has been demonstrated, however it remains to be shown that comparable
results can be obtained from computer based testing. The following chapter describes the methodology involved in determining the effect of computerization upon the Henmon-Nelson.
CHAPTER 3

Methodology

The purpose of this study was to assess the effect of computer administration on the scores resulting from a multiple-choice test that was administered interactively via conversational, CRT terminals. Methodologically, the study was straightforward. In brief, each respondent completed two alternate forms of a multiple-choice formatted, standardized test. One testing was done via computer while the other was carried out in the conventional manner. The resulting two sets of scores were analyzed to determine the effect of computerized administration on a standardized test.

1. Measuring Instruments

Henmon-Nelson

The test chosen to be used in the study was the Henmon-Nelson Test of Mental Ability, level 4. It was selected on the basis of three criteria.

First, the test used must have parallel forms with adequate parallel forms reliability. As tabled in a previous chapter, the Henmon-Nelson has an overall reliability of approximately 0.86 (Pearson r). This level was considered acceptable for the purposes of this study.

Secondly, the chosen test must be true/false, multiple-choice, or matching in format. This was required in order to facilitate the use of answer keys in machine scoring. The Henmon-Nelson consists exclusively of multiple-choice type
Finally, the test must contain only verbal or numerical stimuli and responses so as to minimize the complexity of the computer program which was used to administer the test. Graphical stimuli would necessitate the use of terminals with powerful display capabilities. Availability of these types of terminals at the University of British Columbia's computer installation is limited and very expensive.

Validity was not used as a criterion in the selection of the test. Obtained scores were not interpreted. The sole concern of the project was the consistency of measure across mode of presentation.

**Demographic Questionnaire**

A demographic questionnaire was administered to all respondents before the testing procedures began. A copy of this questionnaire can be found in Appendix A. This information was stored in the computer, and used as required for analysis.

**Computer Program**

Vital to this study was a computer program that was developed to administer the Henmon-Nelson Test of Mental Ability. It is a program written in PL/1 designed to be used on the University of British Columbia's Amdahl 470 V/6, Model II computer which is under control of a Michigan Terminal System (MTS) operating system. The program is intended to be accessed through the use of the Ann Arbor conversational terminals. A complete listing of the program can be found in Appendix B.

When the subject first sat down at the terminal the subject
identification number was entered. The program then displayed a series of simple directions regarding the use of the terminal. The subject was instructed by the computer on how to enter responses, proceed to the next page, return to a previous page, and how to change answers. The program then displayed the instructions for taking the test in much the same manner as they appear in the examiner's manual. The three sample items that were used with the written form were presented to give the subject an opportunity to "try out" the terminal. The computer then gives the subject the option of reviewing the instructions or beginning the test. When the respondent decided he was ready to begin the test and indicated this to the computer, the program selected the appropriate form for administration on the basis of the subject's identification number.

In designing the program, every attempt was made to duplicate the procedures used when taking the test in the conventional manner. Items were presented by the page (terminals are set in page mode) in a crowded format, paralleling that of the test booklet. Items could be answered in any order, responses could be changed any number of times and pages could be 'turned' at will. Emphasis was placed upon keeping the terminals' operating instructions to a minimum.

The program has an automatic feature which kept track of elapsed time from the moment the first page was presented until the 40 minute time limit had expired. Time remaining was displayed in the upper right hand corner of the terminal's screen and changed each time a page was flipped.
The program has several features useful in data collection. A record of total time spent on each page was kept along with number of times each page was displayed. Responses for all items were stored and the number of response changes per item was recorded. The program scores the test, records the three scores, and stores all collected data in a file for further analysis.

Design

The study followed the classical, single group, two trial, repeated measure design. It was a one-way design, intended only to assess differences between the two trials.

<table>
<thead>
<tr>
<th></th>
<th>$A_1$</th>
<th></th>
<th>$A_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>$Y_{11}$</td>
<td>$Y_{11}$</td>
<td></td>
</tr>
<tr>
<td>$S_2$</td>
<td>$Y_{21}$</td>
<td>$Y_{21}$</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>$S_n$</td>
<td>$Y_{n1}$</td>
<td>$Y_{n1}$</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1 - Two-Trial Repeated Measure Design

The two observations, $A_1$ and $A_2$, are the scores obtained from testing through the computer and testing in the conventional fashion.

Most of the current literature that deals with research design (Campbell and Stanley, 1966; Kerlinger, 1973) list history, maturation, and statistical regression as possible threats to the internal validity of this design. However, in this study these threats are controlled. Both history and maturation are largely controlled by the extremely short time
interval between testings (10 minutes). Statistical regression was not a concern since only one group was used and subjects were not selected on the basis of extreme scores on previous testings.

The order and medium by which each subject responded to the different forms was randomized. This controlled factors such as fatigue and pretest sensitization. Randomization of order also controlled the practice effect that was found in the original norming procedures of the test. There was an equal number of subjects who were first observed under the $A_1$ condition to those who were observed first under $A_2$. Randomization also helped in the control of history and maturation for exactly the same reason. The procedure employed for the randomization is discussed in detail later in this chapter.

This choice of design allowed subjects to answer both forms of the Henmon-Nelson using each mode of response in random combinations, thus allowing for isolation of the effect of the different response modes without interference from extraneous variables. This increased the power of the study and made the chosen design a most efficient vehicle for the solution of the problem.

3. Population

The Henmon-Nelson, level 4 is appropriate for students enrolled in first year undergraduate work through first year graduate school. The study used volunteers from the University of British Columbia's general student population who fell into
the above category. Undergraduate classes in the Faculty of Education were searched for volunteers. A maximum of 90 minutes per subject was allowed to complete the testing. A sample size of 36 subjects was used.

4. Data Collection

Each subject was given a copy of the demographic questionnaire to complete when he first came to the lab. When the questionnaire was completed he gave the form to the experimenter who then randomly assigned an identification number to the subject. Each of the identification numbers had previously been randomly assigned to a testing process. Each assignment gave the order in which the test forms were administered and which response mode was used for each of the testings. There were four possible assignments.

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A computer B written</td>
</tr>
<tr>
<td>2</td>
<td>B written A computer</td>
</tr>
<tr>
<td>3</td>
<td>B computer A written</td>
</tr>
<tr>
<td>4</td>
<td>A written B computer</td>
</tr>
</tbody>
</table>

Table IV indicates the order as well as the combinations of mediums and test forms in which the tests were administered. The experimenter directed the subject to either a terminal or the examination room depending upon the assigned process.

Terminals

The experimenter signed the subject on at the terminal and
then invigilated the terminal session. Invigilation required that the assistant be present in the terminal room in the event of hardware/software malfunction during the testing. Only three terminals were in use at any one time, due to the facilities available. Instructions and directions were given by the computer. The invigilator was available to clarify instructions. In the event that the invigilator was required to clarify instructions, he made careful note of the nature of the problem and subject and submitted this data for subjective analysis. Once the testing part of the study had begun, the directions given by the computer were not modified.

The computer first provided an introduction to its operations and then presented three sample items for the subject to try. These items were taken from the Henmon-Nelson's administration manual. The subject was given the opportunity to review the instructions and sample questions once again if he desired. If he wanted to continue on to the test, the first page of the test was then presented and the program's 40 minute time limit was begun. At the end of 40 minutes the program stopped the administration of the test, stored the necessary data, presented a concluding message to the subject thanking him for his cooperation and then "signed off." The subject then reported to the research assistant for further instructions.

**Written Form**

The experimenter assigned the subject to one of three chairs at a table in the examination room and administered the appropriate form of the Henmon-Nelson following the prescribed procedure outlined in the examiner's manual. The subject was
supplied with a test booklet, self scoring answer sheet, and a pencil. Each subject was allotted the prescribed 40 minute time limit to complete the test.

At the end of 40 minutes each paper was collected. This data was keypunched on computer cards in preparation for machine scoring and data storage. In between the writing of the two test forms, subjects were given a 10 minute rest period, during which they were not allowed to converse with other subjects; coffee was available.

Testing was carried out in the Educational Research Service Center (ERSC) located in the basement of the Scarfe building at the University of British Columbia. Times for testing were arranged at the convenience of the subjects. Testing was carried out over an 8 day period.

Data Collected

Data was collected on each subject from three different sources. Data from the first source, the demographic questionnaire, was coded, keypunched, and read into the computer for machine storage and use in later analysis.

All the data that was collected during the computer administration of the test was automatically stored by the program at the end of the 40 minutes. The types of data collected during this procedure included individual item responses, the total time spent on each page, the number of times each page was displayed, and the number of response changes per question. The scores on the verbal and quantitative subtests as well as the total test score were stored.
Responses from the written form were key-punched in preparation for machine scoring. Since item data was collected and stored on disk, machine scoring of the conventional testing was done in the interests of time and efficiency. The only types of data that were feasible to collect from the written form were response data and scores on subtests and totals. This demonstrated one of the advantages of computer administration over the conventional method of test administration.

The final data file contained three records per subject, each headed by subject identification. The first record contained the demographic data, the second record the data collected during computer administration, while the third record consisted of data collected during the conventional administration. All data and program files were backed up on magnetic tape in order to protect against a system malfunction.

5 Statistics

With the design following a two trial, repeated measure paradigm, the most efficient means of analysis was the paired (dependent) t-test (Winer, 1971). Since a covariance term is found in the denominator of the t expression, it is sometimes referred to as the correlated t.\footnote{Recall that covariance is proportional to the correlation coefficient} Repeated measures is actually a matching process where every subject is matched with himself;
hence dependency across measures.

This dependency however, is useful. It helps in the control of some of the extraneous variables which can influence the dependent variable. In other words, the dependent samples reduce the subject to subject variability (residual variance).

The usual assumptions of normality of population and homoscedasticity were made.

The three hypotheses, stated in the null form were as follows.

1. There is not a statistically significant difference between the average raw quantitative score obtained from a computerized administration of the Henmon-Nelson Test of Mental Ability, level 4 and the average raw quantitative score obtained from a conventional administration of the test, at the $\alpha=.05$ level of significance.

2. There is not a statistically significant difference between the average raw verbal score obtained from a computerized administration of the Henmon-Nelson Test of Mental Ability, level 4 and the average raw verbal score obtained from a conventional administration of the test, at the $\alpha=.05$ level of significance.

3. There is not a statistically significant
difference between the average raw total score obtained from a computerized administration of the Henmon-Nelson Test of Mental Ability, level 4 and the average raw total score obtained from a conventional administration of the test, at the $\alpha=.05$ level of significance.

The analysis consisted of three non-directional t-tests; one for the quantitative score, one for the verbal score and one for the total score. The statistical hypotheses for each of the three tests were:

$$H_0 : u_1 = u_2$$
$$H_1 : u_1 \neq u_2$$

Scores on each subtest as well as the total score were analyzed by means of SPSS (Nie, Hull, Jenkins, Steinbrenner, and Bent, 1975), using the T-Test/Pairs procedure card. Missing data was not encountered in the analysis.

Pearson product-moment correlation coefficients were also computed to give an indication of the magnitude of the relationship between the scores obtained from the computer administration of the Henmon-Nelson and the scores obtained in the traditional manner. SPSS was used to determine this coefficient as well as testing its significance.
CHAPTER 4

Results

The analysis of the data collected in the study was divided into four categories: 1) descriptive analysis of the sample; 2) analysis directed towards answering the research questions; 3) post hoc analysis; and 4) subjective analysis. Each of the above divisions are dealt with in turn.

1. Sample

A sample of 36 subjects was used in the study. The subjects were equally divided among the four testing procedures as listed in Table IV. Table V through Table IX give frequency breakdowns with respect to demographic variables.

Table V - Frequency Of Gender

(i) Gender

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>19</td>
</tr>
<tr>
<td>Female</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td>36</td>
</tr>
</tbody>
</table>

Table VI - Frequency Of Age

(ii) Age

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Less Than 18</td>
<td>1</td>
</tr>
<tr>
<td>18-23</td>
<td>22</td>
</tr>
<tr>
<td>24-29</td>
<td>11</td>
</tr>
<tr>
<td>30+</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>36</td>
</tr>
</tbody>
</table>
### Table VII - Frequency Of Major

**(iii) Major**

<table>
<thead>
<tr>
<th>Major</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fine Arts</td>
<td>8</td>
</tr>
<tr>
<td>Humanities</td>
<td>3</td>
</tr>
<tr>
<td>Social Sciences</td>
<td>6</td>
</tr>
<tr>
<td>Science</td>
<td>15</td>
</tr>
<tr>
<td>Phys Educ</td>
<td>4</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>36</strong></td>
</tr>
</tbody>
</table>

### Table VIII - Frequency Of Year Of Enrollment

**(iv) Year of enrollment**

<table>
<thead>
<tr>
<th>Year of enrollment</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>Graduate</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>36</strong></td>
</tr>
</tbody>
</table>

### Table IX - Frequency Of Prior Terminal Use

**(v) Previous use of computer terminals**

<table>
<thead>
<tr>
<th>Use</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>14</td>
</tr>
<tr>
<td>No</td>
<td>22</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>36</strong></td>
</tr>
</tbody>
</table>

Five of the 14 subjects were familiar with computer terminals as a result of their previous employment as bank
tellers. Five of the remaining nine had used terminals in high school while the last four subjects gained their experience at the University of British Columbia.

Of the 36 subjects, 28 were enrolled in the Faculty of Education (18 elementary; 10 secondary) while the remaining 8 were from the Faculties of Arts and Science.

2. Data Analysis

Table X provides a list of means, standard deviations, and standard errors in raw score metric for each of the six dependent variables: computerized verbal score (CV), written verbal score (WV), computerized quantitative score (CQ), written quantitative score (WQ), computerized total score (CT), and written total score (WT).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV</td>
<td>42.056</td>
<td>9.940</td>
<td>1.657</td>
</tr>
<tr>
<td>WV</td>
<td>41.028</td>
<td>11.770</td>
<td>1.962</td>
</tr>
<tr>
<td>CQ</td>
<td>27.445</td>
<td>7.534</td>
<td>1.256</td>
</tr>
<tr>
<td>WQ</td>
<td>26.194</td>
<td>8.632</td>
<td>1.439</td>
</tr>
<tr>
<td>CT</td>
<td>69.500</td>
<td>15.890</td>
<td>2.648</td>
</tr>
<tr>
<td>WT</td>
<td>67.222</td>
<td>18.509</td>
<td>3.085</td>
</tr>
</tbody>
</table>

Meaningful comparisons of means and standard deviations can only be made between the two verbal scales, the two quantitative scales, and the two full scales due to the raw score metric.
Table XI provides the proportional means to facilitate inter-scale comparison.

<table>
<thead>
<tr>
<th></th>
<th>Computerized</th>
<th>Written</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>0.701</td>
<td>0.684</td>
</tr>
<tr>
<td>Q</td>
<td>0.686</td>
<td>0.655</td>
</tr>
<tr>
<td>Total</td>
<td>0.695</td>
<td>0.672</td>
</tr>
</tbody>
</table>

Examination of the means in Tables X and XI show that the computerized scores are consistently higher than the written scores across the three scales. This could suggest the presence of a nonsignificant novelty effect. Table X also shows a smaller variance of scores across each scale for the computerized forms.

**Hypothesis 1**

In considering the first null hypothesis (page 32) regarding differences between the average, raw, computerized verbal score and the average, raw, written verbal score, a dependent t-statistic was calculated \( t=0.71, \text{df}=35, p=0.485 \) which indicated that there was no significant difference between the average, raw, computerized verbal score and the average, raw, written verbal score and that the null hypothesis \( H_0: \mu_1=\mu_2 \) should not be rejected. A Pearson product moment correlation coefficient of 0.688 was also calculated and found to be significantly different from zero \( (p<0.001) \). This suggested a significant correlation between the raw computerized verbal scores and the raw written verbal scores.
Hypothesis 2

In considering the second null hypothesis (page 32) regarding differences between the average, raw, computerized quantitative score and the average, raw, written quantitative score, a dependent t-statistic was calculated \( (t=0.96, \ df=35, \ p=0.342) \) which indicated that there was no significant difference between the average, raw, computerized quantitative score and the average, raw, written quantitative score and that the null hypothesis \( (H_0: u_1 = u_2) \) should not be rejected. A Pearson product moment correlation coefficient of 0.544 was also calculated and found to be significantly different from zero \( (p<0.001) \). This suggested a significant correlation between the raw computerized quantitative scores and the raw written quantitative scores.

Hypothesis 3

In considering the third null hypothesis (page 32) regarding differences between the average, raw, computerized total score and the average, raw, written total score, a dependent t-statistic was calculated \( (t=0.88, \ df=35, \ p=0.384) \). This result indicated that there was no significant difference between the average, raw, computerized total score and the average, raw, written total score and that the null hypothesis \( (H_0: u_1 = u_2) \) should not be rejected. A Pearson product moment correlation coefficient of 0.604 was calculated and found to be significantly different from zero \( (p<0.001) \). This suggested a significant correlation between the raw computerized total scores and the raw written total scores.
In all three cases the null hypothesis was supported as the available literature had suggested. Thus it appears reasonable to conclude that computerized administration of the Henmon-Nelson Test of Mental Ability, level 4, can produce results equivalent to those obtained from conventional administrations. These results also suggest that computerized administration of tests similar in nature and format to the Henmon-Nelson, does not affect the resulting scores.

3. Post Hoc Analysis

Post hoc analyses took the form of a one trial factor repeated measures ANOVAs with each of the demographic variables acting in turn as grouping factors. Table XII contains the means and standard deviations for each of the variables used in the analysis. The variable "year" records the subject's present enrollment status at the University of British Columbia while "comp" signals previous computer terminal experience. "Major" indicates the subject's academic major or major area of study.

A summary of the post hoc findings with respect to the various grouping variables is presented in Table XIII. F ratios regarding the within factor (mode of presentation) and interactions are not provided since they were all found to be nonsignificant. Marginal means for the within factor can be found in Table X. In Table XIII an F ratio followed by a single asterisk (*) indicates 5% significance while a double asterisk (**) indicates a 1% significance level.
### Table XII - Post Hoc Means And (Standard Deviations)

<table>
<thead>
<tr>
<th></th>
<th>CV</th>
<th>WV</th>
<th>CQ</th>
<th>WQ</th>
<th>CT</th>
<th>WT</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENDER</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>44.32</td>
<td>43.68</td>
<td>29.58</td>
<td>28.05</td>
<td>73.89</td>
<td>71.74</td>
</tr>
<tr>
<td></td>
<td>(7.39)</td>
<td>(10.75)</td>
<td>(7.54)</td>
<td>(9.20)</td>
<td>(13.64)</td>
<td>(18.28)</td>
</tr>
<tr>
<td>Female</td>
<td>39.52</td>
<td>38.06</td>
<td>25.06</td>
<td>24.12</td>
<td>64.59</td>
<td>62.18</td>
</tr>
<tr>
<td></td>
<td>(11.91)</td>
<td>(12.46)</td>
<td>(6.98)</td>
<td>(7.69)</td>
<td>(17.16)</td>
<td>(17.96)</td>
</tr>
<tr>
<td>AGE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-23</td>
<td>38.83</td>
<td>37.87</td>
<td>25.39</td>
<td>24.52</td>
<td>64.22</td>
<td>62.39</td>
</tr>
<tr>
<td></td>
<td>(9.67)</td>
<td>(12.37)</td>
<td>(7.38)</td>
<td>(8.69)</td>
<td>(15.45)</td>
<td>(18.96)</td>
</tr>
<tr>
<td>24+</td>
<td>47.77</td>
<td>46.62</td>
<td>31.08</td>
<td>29.15</td>
<td>78.85</td>
<td>75.77</td>
</tr>
<tr>
<td></td>
<td>(7.84)</td>
<td>(8.39)</td>
<td>(6.59)</td>
<td>(7.99)</td>
<td>(12.27)</td>
<td>(14.71)</td>
</tr>
<tr>
<td>MAJOR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arts</td>
<td>42.50</td>
<td>44.63</td>
<td>22.75</td>
<td>20.63</td>
<td>65.25</td>
<td>65.25</td>
</tr>
<tr>
<td></td>
<td>(9.27)</td>
<td>(9.67)</td>
<td>(8.79)</td>
<td>(8.48)</td>
<td>(16.85)</td>
<td>(16.59)</td>
</tr>
<tr>
<td>Humants</td>
<td>43.33</td>
<td>34.67</td>
<td>31.33</td>
<td>20.67</td>
<td>74.67</td>
<td>55.33</td>
</tr>
<tr>
<td></td>
<td>(11.59)</td>
<td>(13.05)</td>
<td>(3.78)</td>
<td>(9.07)</td>
<td>(15.37)</td>
<td>(22.12)</td>
</tr>
<tr>
<td>Soc Sc</td>
<td>40.17</td>
<td>40.33</td>
<td>26.83</td>
<td>29.00</td>
<td>67.00</td>
<td>69.33</td>
</tr>
<tr>
<td></td>
<td>(10.59)</td>
<td>(10.50)</td>
<td>(8.23)</td>
<td>(7.29)</td>
<td>(16.57)</td>
<td>(16.52)</td>
</tr>
<tr>
<td>Science</td>
<td>42.60</td>
<td>40.80</td>
<td>30.20</td>
<td>29.67</td>
<td>72.80</td>
<td>70.74</td>
</tr>
<tr>
<td>Phys Ed</td>
<td>41.00</td>
<td>40.50</td>
<td>24.50</td>
<td>24.25</td>
<td>65.50</td>
<td>64.75</td>
</tr>
<tr>
<td></td>
<td>(5.66)</td>
<td>(7.33)</td>
<td>(7.33)</td>
<td>(4.86)</td>
<td>(12.77)</td>
<td>(10.34)</td>
</tr>
<tr>
<td>YEAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 - 2</td>
<td>41.38</td>
<td>36.88</td>
<td>28.75</td>
<td>26.13</td>
<td>70.13</td>
<td>63.00</td>
</tr>
<tr>
<td></td>
<td>(11.89)</td>
<td>(17.23)</td>
<td>(7.36)</td>
<td>(9.57)</td>
<td>(18.75)</td>
<td>(25.81)</td>
</tr>
<tr>
<td>3 - 4</td>
<td>40.00</td>
<td>38.83</td>
<td>23.58</td>
<td>20.58</td>
<td>63.58</td>
<td>59.42</td>
</tr>
<tr>
<td></td>
<td>(8.59)</td>
<td>(10.56)</td>
<td>(6.79)</td>
<td>(6.65)</td>
<td>(12.89)</td>
<td>(14.87)</td>
</tr>
<tr>
<td>5+</td>
<td>43.94</td>
<td>44.75</td>
<td>29.69</td>
<td>30.43</td>
<td>73.63</td>
<td>75.19</td>
</tr>
<tr>
<td></td>
<td>(10.15)</td>
<td>(8.68)</td>
<td>(7.43)</td>
<td>(7.36)</td>
<td>(16.03)</td>
<td>(14.15)</td>
</tr>
<tr>
<td>COMP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>47.86</td>
<td>48.14</td>
<td>32.57</td>
<td>33.35</td>
<td>80.43</td>
<td>81.50</td>
</tr>
<tr>
<td></td>
<td>(6.71)</td>
<td>(4.95)</td>
<td>(4.94)</td>
<td>(4.85)</td>
<td>(10.55)</td>
<td>(8.45)</td>
</tr>
<tr>
<td>No</td>
<td>38.36</td>
<td>36.50</td>
<td>24.18</td>
<td>21.64</td>
<td>62.55</td>
<td>58.14</td>
</tr>
<tr>
<td></td>
<td>(10.00)</td>
<td>(12.67)</td>
<td>(7.13)</td>
<td>(7.33)</td>
<td>(14.88)</td>
<td>(17.45)</td>
</tr>
</tbody>
</table>


Table XIII - Post Hoc Summary

<table>
<thead>
<tr>
<th></th>
<th>Verbal</th>
<th>Quantitative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>F=2.550</td>
<td>F=3.392</td>
<td>F=3.612</td>
</tr>
<tr>
<td></td>
<td>(df=34, p=.120)</td>
<td>(df=34, p=.074)</td>
<td>(df=34, p=.066)</td>
</tr>
<tr>
<td>Major</td>
<td>F=0.140</td>
<td>F=1.206</td>
<td>F=0.172</td>
</tr>
<tr>
<td></td>
<td>(df=33, p=.392)</td>
<td>(df=33, p=.328)</td>
<td>(df=33, p=.131)</td>
</tr>
<tr>
<td>Year</td>
<td>F=0.965</td>
<td>F=4.555 *</td>
<td>F=2.166</td>
</tr>
<tr>
<td></td>
<td>(df=33, p=.392)</td>
<td>(df=33, p=.018)</td>
<td>(df=33, p=.131)</td>
</tr>
<tr>
<td>Age</td>
<td>F=7.792 **</td>
<td>F=4.860 *</td>
<td>F=8.284 **</td>
</tr>
<tr>
<td></td>
<td>(df=34, p=.009)</td>
<td>(df=34, p=.034)</td>
<td>(df=34, p=.007)</td>
</tr>
<tr>
<td>Comp</td>
<td>F=12.845 **</td>
<td>F=32.586 **</td>
<td>F=26.468 **</td>
</tr>
<tr>
<td></td>
<td>(df=34, p=.001)</td>
<td>(df=34, p=.001)</td>
<td>(df=34, p=.001)</td>
</tr>
</tbody>
</table>

From Table 4.9 it is evident that there are nonsignificant gender and major-area-of-study effects. A gender effect was not expected since the Henmon-Nelson is reported not to have a sex bias. However, a significant "major" effect was anticipated due to the highly verbal nature of the test, but none in fact materialized.

The variable "year" had a significant F ratio when analyzed with the quantitative scores as the criterion measure. Analysis using Scheffe's multiple contrast method was carried out and the average raw quantitative score of subjects in their third and fourth year at the University of British Columbia (group 2) was found to be significantly lower than that of the subjects in their fifth and graduate years (group 3). This is most likely due to type I error and artifacts of sampling. The remaining two pairwise comparisons were both nonsignificant.
The age factor was blocked into two levels; 23 years and younger or 24 years and older. This was done to obtain sufficient cell sizes to facilitate the analysis of variance procedures. Significance was found on all three criterion variables and examination of the cell means indicated the average raw scores on all scales for the younger subjects were substantially lower than those of the older group.

The final variable which was used in the analysis, previous-computer-use, provided interesting results. Significant F ratios (p < .001) were obtained for each of the three scales. Subjects who had previously used computer terminals consistently scored higher on each scale than did those subjects without any previous computer exposure. Interactions were nonsignificant as was the within factor (mode of presentation) effect. This could imply that the subjects who have used computers are more "intelligent", as defined and measured by the Henmon-Nelson Test of Mental Ability, level 4, than those who have not. It could also suggest that computer science develops skills which sensitize subjects to the Henmon-Nelson. These are two areas for future studies to explore.

4. Subjective Analysis

At the end of the first day of testing a minor problem with the computer administration program was accidentally discovered during data-file back up procedures. The problem was traced to a routine that checked for duplicate subject identification numbers. The routine took the first three digits of the
identification number and checked the corresponding element in an array consisting exclusively of zeros and ones. If the element was a "0", it was then changed to a "1" and the subject continued with the testing program. However, if the element was non-zero (i.e. "1"), a duplicate identification number had been entered and the subject was asked to reenter his number. If the same condition occurred, the subject was instructed by the computer to see the experimenter. The problem was that, under freak conditions, portions of the checking array was blanked out, thus not allowing a subject whose number corresponded with a blanked out element to be tested. Fortunately, the problem was discovered before any complications arose.

The CAI\(^1\) portion of the program that instructed the subjects in the procedures required in the operation of the terminal worked remarkably well. The program determined, that subjects took approximately 5 to 6 minutes to complete this segment which included the practice test items. Instructions were apparently clear enough as problems regarding terminal procedures did not arise. These instructions however, should have made clear to subjects that the program would automatically terminate at the end of the 40 minutes of testing, as one or two subjects became "befuddled" when the program terminated.

One subject who was assigned to group 3 (B computer and A written) completed the computerized B form and began to take

\(^{1}\) Computer Assisted Instruction
the written A form. After a very short time the subject became frustrated and left the examination room. Through discussion with the experimenter, it was discovered that the frustration was due to English being the subject's second language causing great difficulty in completing the highly verbal Henmon-Nelson. Data collected on this subject was treated as incomplete and excluded from the analysis. It is interesting to note that the subject became frustrated while taking the written form and not the computerized form.

In general, the attitudes of the subjects towards computerized testing, as judged through informal interviews, were favourable. Most subjects referred to the computerized version as "fun". Every subject preferred the computerized version to the conventional mode of administration. Some said it was "better", in terms of speed and ease of response to the verbal items, but slower for quantitative items as they had to be copied onto paper before they could be solved. One subject found it difficult to look at a CRT screen for 45 minutes while another subject found the reading distance (i.e., distance from eye to focal plane) to be too great.

Computerized administration of the Henmon-Nelson was relatively inexpensive. The cost of administration was approximately $4.27 (computer dollars) per person. This includes the $2.25 fee for connect time and a nominal charge

---

1 based upon a rate of $3.00 per hour at approximately 45 minutes per person
for disk storage.

This concludes the analysis phase of this study. The following chapter will present a brief summary, a discussion of the study's limitations, and suggestions for future research.
CHAPTER 5

Discussion

1. Summary

The general problem to which this study addressed itself, was to determine what effect computerized administration had on the scores obtained from a standardized test. Clearly, a problem of such size would be very difficult if not impossible to answer adequately from the results of just one study. It was necessary to reduce the scope of the problem to a manageable size. Thus it became the purpose of this study to determine what effect, if any, computerized administration had on the scores obtained from the Henmon-Nelson Test of Mental Ability, level 4. From this problem the study's null hypotheses emerged.

1. There is not a statistically significant difference between the average raw quantitative score obtained from a computerized administration of the Henmon-Nelson Test of Mental Ability, level 4 and the average raw quantitative score obtained from a conventional administration of the test, at the $\alpha = .05$ level of significance.

2. There is not a statistically significant difference between the average raw verbal score obtained from a computerized administration of the Henmon-Nelson Test of Mental Ability, level 4 and the average raw verbal score obtained from a conventional administration of the test, at the $\alpha = .05$ level of
significance.

3. There is not a statistically significant difference between the average raw total score obtained from a computerized administration of the Henmon-Nelson Test of Mental Ability, level 4 and the average raw total score obtained from a conventional administration of the test, at the \( \alpha = 0.05 \) level of significance.

Dependent t-test statistics and Pearson product moment correlation coefficients were calculated with respect to each of the three hypotheses and in each case the null hypothesis was not rejected. The Pearson r coefficients indicated significant correlations between the computerized and written scores. These results supported the three null hypotheses and suggested that computerized administration of the Henmon-Nelson can provide results comparable to those obtained from a conventional administration.

Post hoc analyses were carried out to assess the effect of the different demographic variables. Nonsignificant F ratios were found for gender and major-area-of-study while significance resulted for age and previous-computer-use. Subjects older than 23 years scored considerably higher than those 23 and younger. Those students with previous computer experience scored higher than those without. A "year-of-enrollment" effect was found, but it was attributed to sampling error.
2. Weaknesses To Be Considered

The structure of the proposed study was considered sound. Extraneous variables that could influence the dependent measures were controlled both by design and the use of randomization. The statistical model on which the analysis was based was straightforward and for the purposes of this study, elegant. However, as in most if not all research, designs do fall short of the ideal and this study was no exception. One or two minor shortcomings existed but these did not radically affect the findings.

The most serious problem was one relating to the nature of the sample. Since the sample consisted totally of volunteers, a bias was introduced. This bias was a result of the sample not being strictly representative of the target population. The subjects were volunteers and possibly had ulterior motives for participating in the study. These motives could have influenced the observed measures under the experimental conditions.

Members of the target population who did not volunteer also bias the study. For example, it is unlikely that "computer-phobics" volunteered as subjects for a study which used computer-subject interaction as a means of collecting data. People belonging to such subpopulations should be included in the testing since generalizations will be made about the population to which they belong. A randomly selected sample is ideal for almost any study but the possibility for that here, simply did not exist. Clearly, inclusion (in the proper proportion) of people who do not volunteer would eliminate the
bias that resulted from using only volunteers. Using volunteers as subjects in studies of this nature did present problems with respect to external validity, but they probably were not serious enough to drastically affect the results of the study.

Other weaknesses in the study were related to the computer program that was used to administer the test. In presenting the test it was intended to display each page in a format paralleling that of the test booklet. Unfortunately the 37 line CRT display does not permit exact duplication. Hence, the items were not presented in the same page groupings as they appeared in the test booklet. Since different item groupings tend to cause score fluctuations, it is probable that this source of uncontrollable variance caused a slight increase in the error variance. This increase was not considered large enough to affect the final results of the study.

A minor point but still worth noting, is that all alphabetic characters were displayed exclusively in upper case. It was not practical in terms of computer time and real time to have the display utilizing both upper and lower cases. Considering the age of the subjects, confusion arising from this, was not anticipated and did not occur.

Another computer related weakness is that a warning was not given to the user when an illegal character (ie. not a "1" or "2" or "3" or "4" or "5") was given as a response. Technically, it was a simple addition to the program, but problems regarding the 37 line CRT display limitation of the Ann Arbor terminals made the change impractical. The program ignored all illegal
characters in the input.

In reviewing his answers, the subject proceeded page by page. He had to go to either the page immediately preceding or immediately ahead of the present page. He did not have the freedom to turn directly to a given page. This technically, was again easy to change but it was also necessary to keep the terminal's operating procedures down to a bare minimum. As a result the change was regarded as an unnecessary luxury.

3. Future Research

With computerized test administration being a relatively recent development in the art of testing, many new areas exist for future related research. To facilitate discussion these areas have been divided into three groups. Clearly these groups are not mutually exclusive; they serve only as a means of identifying the focus of each group of suggested studies.

The first group of studies includes all those which are directly related to the computerization of the Henmon-Nelson. These would include reliability studies in which subjects took both computerized versions with parallel forms reliability coefficients calculated and compared to that of the manual (i.e. 0.88).

Computerized test-retest reliability studies should also be done in order to fully appreciate the effect of computerized administration on the Henmon-Nelson.

Further studies should be carried out to determine if the
significant $F$ ratio corresponding to year of enrollment across the quantitative scale, is in fact attributable to sampling error. A large sample drawn at random should be used in a study of this nature. A repeated measure design with a single, six level grouping factor would be the best model to follow in the analysis.

The significance of "previous-computer-usage" suggests two series of studies to be undertaken. First, one series should explore the possibility that people who use computers do score significantly higher on the Henmon-Nelson's three scales than those who have not used a computer. Secondly, studies should be done to determine if the skills tested by the Henmon-Nelson are those which are developed in the study of computer science. And thirdly, could the Henmon-Nelson Test of Mental Ability be used as an aptitude test for computer science? This third group of studies presupposes positive results from the second series.

Studies could be done to assess the effects of implementing the technical programming refinements listed previously in this chapter. These studies would attempt to determine if the results of this paper are still valid. Changes in the presentation format of the test could also be made and their effects analyzed. For example, items could be presented in groups of five per screen-page. This would drastically alter the appearance of the test and perhaps result in the scores being radically different.

The second group of studies addresses itself to the more general topic of computerized testing. This study limited
itself to the administration of one particular test on one particular type of terminal. Studies should be carried out that examine the different ways the different tests behave on different terminal types. Perhaps different results would have been obtained if, for example, the Henmon-Nelson was administered on a Vu-Com terminal or an IBM 3270 type terminal. Again, repeated measure designs with grouping factors would be the most efficient means of analysis.

Another series of studies would examine the different methods of presenting the terminal's operating instructions. The instructions in this study were presented via a CAI type routine at the beginning of the terminal session. This worked remarkably well as there were no problems or queries arising from this procedure. However, there are other ways in which these instructions could have been presented and it is important to examine the efficiencies of these alternate methods in order to improve computerized testing.

The final group of studies would concern itself with the psychometric properties of the different kinds of information which now is seen as practical to collect using computerized testing. These studies will answer such questions as: Can response latency be related to score, reliability, or even measurement error? What is the relationship between the number of times the response to an item was changed and the difficulty index and item variance? The answers to these questions and others will open new doors in the areas of test score theory and response style analysis.
With computerized testing being technically feasible, it is hoped that future studies will demonstrate the merits of this new field and thus prove it to be a useful and revolutionary development in the art of testing.
Bibliography


Appendix A

The following questionnaire was used to collect general information on each of the subjects participating in the study. It was distributed to each subject prior to the beginning of his part in the testing.
Subject Number........
Time..................
Date..................

1. SEX:  ○ Male   ○ Female

2. AGE:  ○ 18     ○ 18-23   ○ 24-29  ○ 30+

3. MAJOR: ○ Fine Arts
           ○ Humanities
           ○ Social Sciences
           ○ Science
           ○ Physical Education or Recreation
           ○ Other  (specify..................)

4. Are you enrolled in the Faculty of Education?
   ○ No   ○ Elementary   ○ Secondary

5. Which year are you enrolled in?
   ○ 1   ○ 2   ○ 3   ○ 4   ○ 5   ○ Grad

6. Have you ever used a computer terminal before?
   ○ Yes   ○ No

   If yes, where?  ○ High School
                   ○ University
                   ○ Work   (specify...............)
                   ○ Other  (specify...............)


Appendix B

The following pages represent a source listing of the computer program which was used to administer the Henmon-Nelson Test Of Mental Ability, level 4. It is written in PL/1 and designed to be used on the University of British Columbia's Amdahl 460 V/6, Model II computer. The program is designed to use only the Ann Arbor CRT terminals at the University of British Columbia which are supported by a PDP-11/40 mini computer attached to the Amdahl 470.
MAIN: PROCEDURE OPTIONS (MAIN);
/* THIS PROGRAM PRESENTS AN INTELLIGENCE TEST TO A STUDENT BASED UPON WHICH OF FOUR */
/* GROUPS HE/SHE IS IN. */
/* GROUP 1 TAKES TEST A BY COMPUTER, THEN TAKES TEST B MANUALLY. */
/* GROUP 2 TAKES TEST B MANUALLY, THEN TAKES TEST A BY COMPUTER. */
/* GROUP 3 TAKES TEST B BY COMPUTER, THEN TAKES TEST A MANUALLY. */
/* GROUP 4 TAKES TEST A MANUALLY, THEN TAKES TEST B BY COMPUTER. */
/* EACH STUDENT IS GIVEN A UNIQUE I.D. BY THE EXPERIMENTER. THIS I.D. IS */
/* RECORDED IN THE STUDENT FILE. IF THE EXPERIMENTER ATTEMPTS TO SPECIFY AN */
/* I.D. WHICH HAS ALREADY BEEN ALLOCATED, THE PROGRAM PROMPTS HIM FOR A NEW ONE*/
/* AS THE STUDENT TAKES THE TEST, HIS RESPONSES AND SOME FREQUENCY AND TIME */
/* DATA ARE RECORDED. WHEN THE TEST IS COMPLETED OR TIME IS UP, THE TEST IS */
/* SCORED AND ALL COLLECTED DATA IS STORED IN THE STUDENT FILE AT MTS LINE */
/* I.D. * 100. */
/* RESPONSES TO THE MANUAL TEST ARE KEY-PUNCHED, THEN READ BY A SECOND PROGRAM */
/* WHICH EDITS THEM FOR ERRORS, SCORES THEM AND STORES THEM IN THE STUDENT FILE*/
/* AT MTS LINE I.D. * 100 + 1. */

/* STUDENT FILE RECORDS: */
DCL 1 HEADER_RECRC,
  2 I_D_FLAG (300) CHAR (11);
DCL 1 TESTER_RECRC,
  5 I_D
  5 GROLP,
  5 ANSWER_DATA,
    10 ANSWERS
    10 ANSWER_FREQ (100)
  5 SCORE_DATA,
    10 NUMERIC_SCORE
    10 VERBAL_SCORE
    10 TOTAL_SCORE
  5 PAGE_FREQ (15)
  5 TIME_ON_PAGE (15)
  5 TOTAL_TIME
/* TEST FILE RECORDS: */
DCL 1 TEST_INFO,
  5 CORRECT_ANSWER (100) CHAR (11),
  5 QUESTION_TYPE (100) CHAR (11),
  5 LAST_PAGE
DCL 1 PAGE_INFO,
  5 LINES_PER_PAGE
  5 C_PER_PAGE
  5 FIRST_QUEST
DCL TESTLINE CHAR (100) VARYING;
/* FILE DECLARATIONS: */
CCL TESTFILE
FILE DIRECT KEYED ENVIRONMENT (U(350) INDEXED);
DCL STUDFILE
FILE DIRECT KEYED ENVIRONMENT (U(350) INDEXED);

/** MISCELLANEOUS DECLARATIONS: */

CCL PAGE_HDR,
  5 HDR_FILL_1 CHAR (6) INIT (' PAGE '),
  5 PAGE_NBR PIC '99',
  5 HDR_FILL_2 CHAR (44) INIT (''),
  5 HDR_FILL_3 CHAR (22) INIT ('TIME REMAINING: ABOUT '),
  5 TIME_REM PIC '99',
  5 HDR_FILL_4 CHAR (4) INIT (' Min');
CCL PAGE_HDR_STRING CHAR (80) DEFINED PAGE_HDR;
CCL ELAPSED ENTRY RETURNS (BIN FLCAT),
SIGNOFF ENTRY;

DCL ANSWER_SCALE,
  5 ANS_SCALE_FILL CHAR (18) INIT ('OUESTION NUMBERS:'),
  5 ANS_SCALE_QUEST (14) PIC 'ZZZZ';
DCL ANSWER_SCALE_STRING CHAR (74) DEFINED ANSWER_SCALE;

CCL ANSWER (110)
  ANSWER_STRING1 CHAR (11) DEFINED ANSWERS,
  ANSWER_CHAR1 (37) CHAR (1) DEFINED ANSWER_STRING1,
  ANSWER_STRING2 CHAR (11) DEFINED ANSWER_STRING2,
  ANSWER_CHAR2 (37) CHAR (1) DEFINED ANSWER_STRING2,
  REVIEW_REPLY CHAR (10),
  CHANGE_PAGE_CHAR CHAR (1),
  KEY CHAR (4) BASED (PTR),
  STUDENT_KEY CHAR (4) BASED (STUDENT_PTR);

DCL FAGENO
  QUEST_NBR BIN FIXED (15) INIT (1),
  SLCT_NBR BIN FIXED (15);
DCL LINE
  STUDENT_LINE BIN FIXED (31),
DCL STARTTIME
  PAGE_TIME BIN FLOAT (21),
  PAGE_TIME BIN FLCAT (21);

/* SET UP FILE KEYS FOR READS AND WRITES: */
  PTR = ACCR (LINE); /* TESTFILE KEY */
  STUDENT_PTR = ACCR (STUDENT_LINE); /* STUDFILE KEY */
/* SET UP TERMINAL FOR OUTPUT */
  CALL CNTL ('*SINK*','HISTORY=0');
  CALL CNTL ('*SINK*','ROLL*');
  CALL CNTL ('*SINK*','TAB=OFF', '5,9,13,17,21,25,29,33,37,41');
/* ATTACH AND OPEN OUTPUT FILE: */
  CALL INSTRUCTIONS;
  CALL CNTL ('*SINK*','LINE*');
  CALL GET_TEST;
STARTTIME = ELAPSE;<
/* PRESENT TEST TO SUBJECT: */
DO WHILE (ELAPSED - STARTTIME < 2400 & PAGENO <= LAST_PAGE):
  CALL DISPLAY_PAGE(PAGENO);
  CALL GET_ANSWERS(PAGENO);
  IF ELAPSED - STARTTIME < 2400 THEN CALL SELECT_PAGE;
END; /* DO */
/* CALCULATE AND SAVE TEST DATA AND QUIT: */
CALL CALC_SCCRES;
/* SAVE STUDENT DATA: */
REWRITE FILE (STUDFILE) FROM (HEADER_RECORD) KEY (STUDENT_KEY);
STUDENT_LINE = I_D * 100000;
WRITE FILE (STUDFILE) FROM (TESTEE_RECORD) KEYFRCM (STUDENT_KEY);
CALL TERM INATE;

INSTRUCTIONS:  PRCC:
DCL INSTFILE FILE DIRECT KEYED ENVIRONMENT(U(350) INDEXED);
DCL ANSFLAG BIT(1);
ICFLAG BIT(1);
GOFLAG BIT(1);
CCL REPLY (?) CHAR (1);
ANSWERS CHAR (3) INIT ('152'),
NBR_WRONG FIXED BIN (15);
CALL CNTL ('*SIK*', 'LINE*');
CALL ATTACH ('STUDFILE=STUDFILE*');
OPEN FILE (STUDFILE) UPDATE;
STUDENT_LINE = 1000; /* MTS LINE */
READ FILE (STUDFILE) INTO (HEADER_RECORD) KEY (STUDENT_KEY);
CALL ATTACH ('INSTRFILE=TEST-INST*');
OPEN FILE (INSTFILE) INPUT;
PAGENO=1;
DO I = 1 TO 3;
  CALL DISPLAY_INSTR(I);
END;
ICFLAG = 'C'B;
DO WHILE (ICFLAG = 'O'B):
  CALL DISPLAY_INSTR(GOICB); /* PAGE 4 */
  IF ANSWER_CHAR2(1) >= '0' & ANSWER_CHAR2(1) <= '9' &
     ANSWER_CHAR2(2) >= '0' & ANSWER_CHAR2(2) <= '9' &
     ANSWER_CHAR2(3) >= '0' & ANSWER_CHAR2(3) <= '9'
  THEN
    I_D = SUBSTR (ANSWER_STRING2, 1, 3);
    GROUP = ANSWER_CHAR2(4);
    IF I_D > 0 & I_D < 3 & GROUP > '0' & GROUP < '5'
    THEN
      IF I_D_FLAG(I_D) = 'C'
      THEN
        GC: I_D_FLAG(I_D) = '1';
        ICFLAG = '1'B;
      ELSE
        IF ICFLAG = 'C'B
      END:
END: /* DO */
THEN

END;

IF ICFLAG = '0'B

END;

CALL DISPLAY_INSTR(1101B); /* PAGE 7 */
CCFLAG = '0'B;
DO WHILE (CCFLAG = '0'B);

CALL DISPLAY_INSTR(11CCB); /* PAGE 8 */
CALL CNTL('*SINK*',*TAB1=ON');
CALL DISPLAY_INSTR(1001B); /* PAGE 9 */
CALL CNTL('*SINK*',*TAB1=OFF');

DO I = 1 TO 3;

REPLY(I) = ANSWER_CHAR2(4*I-3);
IF REPLY(I) < '1' OR REPLY(I) > '5'
THEN REPLY(I) = '-';

END;

ANSFLAG = 'C'B;

CC WHILE (ANSFLAG = '0'B);

CALL DISPLAY_INSTR(11101B); /* PAGE 10 */

DISPLAY('ENTER ANY CHANGES: ')
REPLY(ANSWER_STRING2);
CALL CNTL('*SINK*',*TAB1=OFF');
NBR_WRONG = 3;
DO I = 1 TO 3;

IF ANSWER_CHAR2(4*I-3) = '-' THEN

REPLY(I) = ANSWER_CHAR2(4*I-3);
IF REPLY(I) < '1' OR REPLY(I) > '5' THEN REPLY(I) = '-';

END;

ELSE

END; /* 24 BLANK LINES */
CALL DISPLAY_INSTR(1100B); /* PAGE 12 */
IF ANSWER_CHAR2(1) = 'N'
THEN ANSFAG = '1'B;
END;

ANSW_RGBHAR2(1) = 'L';
WHILE (ANSWER_CHAR2(1) = 'T' & ANSWER_CHAR2(1) = 'R')
CALL DISPLAY_INSTR(1101B); /* PAGE 13 */
IF ANSWER_CHAR2(1) = 'T' THEN GOFLAG = '1'B;
END;

WHILE (ANSWER_CHAR2(1) = 'N')
CALL DISPLAY_INSTR(1102B); /* PAGE 14 */
END;

DISPLAY(‘- ’); DISPLAY(‘- ’); /* 6 BLANK LINES */
CLOSE FILE (INSTFILE);
CALL CNTL (*SINK**, *CALL*);

DISPLAY_INSTR: PROC (P):
DCL P FIXED BIN (15);
DCL PAGE_REC,
  5 NUMLINES PIC '99',
  5 TITLE_FILLER CHAR(8);
LINE = P*1000000;
READ FILE (INSTFILE) INTO (PAGE_REC) KEY (KEY);
DO LINE = P*1000 + NUMLINES TO (P*1000 + NUMLINES) BY 1000;
  READ FILE (INSTFILE) INTO (TESTLINE) KEY (KEY);
  DISPLAY (TESTLINE);
END;
IF P <= 11 THEN
DO: LINE = (P*100 + NUMLINES)*1000;
  READ FILE (INSTFILE) INTO (TESTLINE) KEY (KEY);
  DISPLAY (TESTLINE) REPLY (ANSWER_STRING2);
END;
END DISPLAY_INSTR;
END INSTRUCTIONS;

GET_TEST: PROC ;
DCL TEST_NAME CHAR (20);
DCL TESTFILE=TEST.A;
DCL TESTFILE=TEST.B;
READ FILE (TESTFILE) INTO (TEST_INFO) KEY (KEY);
END GET_TEST;

DISPLAY_PAGAE: PROC (P):
DCL P BIN FIXED (15);
PAGE_FREQ(P) = PAGE_FREQ(P) + 1;
PAGE_TIME = ELAPSED;
PAGE_NBR = P;
TIME_REM = TRUNC((24*60 - (ELAPSED - STARTTIME))/60);
DISPLAY(PAGE_HDR_STRING);
/* READ PAGE HEADER FROM TESTFILE: */
LINE = PAGENO * 10000;
/* MTS LINE = PAGECRC * 100 */
READ FILE (TESTFILE) INTO (PAGE_INFO) KEY (KEY);
/* PRINT THE QUESTIONS: */
DC LINE = PAGENO * 10000 + 1000 TO PAGENO * 10000 + LINES_PER_PAGE * 1000 BY 1000;
READ FILE (TESTFILE) INTO (TESTLINE) KEY (KEY);
DISPLAY(TESTLINE);
/* PRINT THE ANSWER SCALE: */
CALL DISPLAY_ANSWER_SCALE;
CALL GET_PREV_ANSWERS;
IF PAGE_FREQ(F) = 1 THEN
  DISPLAY('ANSWER STRING1:');
ELSE
  DISPLAY('PREVIOUS ANSWERS:  ');  
END;
END DISPLAY_PAGE;

DISPLAY_ANSWER_SCALE: PROC;
/* FILL IN AND PRINT THE ANSWER SCALE: */
ANS_SCALE_CUEST = 0;
J = FIRST_CUEST;
DO I = 1 TO C_PER_PAGE;
  ANS_SCALE_CUEST(I) = J;
  J = J + 1;
END:
DISPLAY(ANSWER_SCALE_STRING);
END DISPLAY_ANSWER_SCALE;

GET_PREV_ANSWERS: PROC;
ANSWER_STRING1 = ' ';
DO I = 1 TO C_PER_PAGE;
  IF ANSWER(FIRST_QUEST+I-1) = '*' THEN
    ANSWER_CHAR1((I-1)*4+1) = '*';
  ELSE
    ANSWER_CHAR1((I-1)*4+1) = ANSWER(FIRST_QUEST+I-1);
END;
END GET_PREV_ANSWERS;

GET_ANSWERS: PROC (F);
DCL P VIA FIXED (15);
CALL CTRL (**SINK**, **TABI=ON**);
IF PAGE_FREQ(F) = 1 THEN
  DISPLAY('ENTER YOUR ANSWERS: ')
  REPLY (ANSWER_STRING2);
ELSE
  DISPLAY('ENTER CORRECTIONS: ')
  REPLY (ANSWER_STRING2);
  CALL CTRL (**SINK**, **TABI=CFF**);
  PAGE_TIME = ELAPSED - PAGE_TIME;
  TIME_ON_PAGE(P) = TIME_CN_PAGE(P) + PAGE_TIME;
END;
DC I = 1 TO Q_PER_PAGE;
SLCT_NBR = 4*I - 3;
QUEST_NBR = FIRST_QUEST + I - 1;
IF ANSWER_CHAR2(SLCT_NBR) = ' ' THEN
  DC: ANSWER_FREQ(QUEST_NBR) = ANSWER_FREQ(QUEST_NBR) + 1;
  IF ANSWER_CHAR2(SLCT_NBR) < '1' OR ANSWER_CHAR2(SLCT_NBR) > '5'
  THEN ANSWER_CHAR2(SLCT_NBR) = ' '; 
  ANSWER(QUEST_NBR) = ANSWER_CHAR2(SLCT_NBR);
END;
END GET_ANSWERS;

SELECT_PAGE: PRCC;
IF PAGENO < LAST_PAGE THEN
  DO:
    IF PAGENC = 1 THEN
      CISPLAY('TYPE "N" TO SEE THE NEXT PAGE: ') REPLY(CHANGE_PAGE_CHAR);
    ELSE
      CISPLAY('TYPE "N" TO SEE THE NEXT PAGE, OR "P" TO SEE THE PREVIOUS PAGE AGAIN: ') REPLY(CHANGE_PAGE_CHAR);
      IF CHANGE_PAGE_CHAR = 'N' THEN
        PAGENO = PAGENO + 1;
      ELSE
        IF CHANGE_PAGE_CHAR = 'P' THEN
          PAGENC = PAGENC - 1;
        ELSE
          IF ELAPSED - STARTTIME < 2400 /* 40 MINUTES */ THEN
            CISPLAY('YOU ARE FINISHED THE TEST. DO YOU WANT TO REVIEW? (YES,NO) ') REPLY(REVIEW_REPLY);
            IF SUBSTR(REVIEW_REPLY,1,1) = 'N' THEN
              PAGENC = PAGENC + 1;
            ELSE
              PAGENC = 1;
            END;
          ELSE
            CISPLAY('YOU ARE FINISHED THE TEST.');
            PAGENO = PAGENC + 1;
          END;
        END;
      END;
    ELSE
      IF ELAPSED - STARTTIME < 2400 /* 40 MINUTES */ THEN
        DC: CISPLAY('YOU ARE FINISHED THE TEST. DO YOU WANT TO REVIEW? (YES,NO) ') REPLY(REVIEW_REPLY);
        IF SUBSTR(REVIEW_REPLY,1,1) = 'N' THEN
          PAGENC = PAGENC + 1;
        ELSE
          PAGENC = 1;
        END;
      ELSE
        CISPLAY('YOU ARE FINISHED THE TEST.');
        PAGENO = PAGENC + 1;
      END;
    END;
  END:
ELSE
  DO: CISPLAY('YOU ARE FINISHED THE TEST.');
  PAGENO = PAGENC + 1;
  END:
END SELECT_PAGE;

CALC_SCCPES: PRCC:
TOTAL_TIME = ELAPSED - STARTTIME;
DC I = 1 TO 100;
IF ANSWER(I) = CORRECT_ANSWER(I) THEN
  DC: TOTAL_SCCRE = TOTAL_SCCRE + 1;
IF QUESTION_TYPE(I) = 'V'
THEN
    VEREAL_SCORE = VEREAL_SCORE + 1;
ELSE
    NUMERIC_SCORE = NUMERIC_SCORE + 1;
END;
END;
END CALC_SCCRES;

TERMINATE: FRCC:
CLOSE FILE (TESTFILE);
FILE (STUCFILE):
DISPLAY (*-THANKYOU FOR PARTICIPATING IN THIS PART OF THE EXPERIMENT.*):
DISPLAY (*OPLEASE SEE THE ATTENDANT FOR FURTHER INSTRUCTIONS.*):
CALL SIGNOFF:
END TERMINATE:
END MAIN:

$SIG