Comparison of the Standard and Computerized Versions of the College Level Examination Program General Examination in English Composition

Alan L. Muhlestein

Utah State University

Follow this and additional works at: https://digitalcommons.usu.edu/etd

Part of the Psychology Commons

Recommended Citation
Muhlestein, Alan L., "Comparison of the Standard and Computerized Versions of the College Level Examination Program General Examination in English Composition" (1981). All Graduate Theses and Dissertations. 5986.
https://digitalcommons.usu.edu/etd/5986
COMPARISON OF THE STANDARD AND COMPUTERIZED VERSIONS OF THE
COLLEGE LEVEL EXAMINATION PROGRAM GENERAL
EXAMINATION IN ENGLISH COMPOSITION

by

Alan L. Muhlestein

A dissertation submitted in partial fulfillment
of the requirements for the degree
of
DOCTOR OF PHILOSOPHY
in
Psychology

Approved:

UTAH STATE UNIVERSITY
Logan, Utah
1991
ACKNOWLEDGEMENTS

I wish to express my gratitude to Dr. Keith T. Checketts for including me in this research and for the selfless giving of his time, expertise, and encouragement not only during the course of this study but throughout the many years of my graduate training.

I also extend my appreciation to Dr. Glenn Maw and Mrs. Gloria Earl of the testing center at Utah State University for their support and assistance during the data collection phase of this study. The efforts of the remainder of my committee, Drs. Michael Bertoch, Joan Kleinke, David Stein, and Ronald Thorkildsen are greatly appreciated. I wish also to acknowledge the personnel at Educational Testing Services for making the computer-administered and paper examinations available for this research.

I am forever grateful to my family members whose warm support and genuine interest have helped me more than they realize. Finally for my wife, Rosanne, who provided unmeasurable assistance, patience, and support, my most sincere "thank you."

Alan Muhlestein
TABLE OF CONTENTS

ACKNOWLEDGEMENTS ................................................... ii
LIST OF TABLES .......................................................... v
LIST OF FIGURES ....................................................... vi
ABSTRACT ...................................................................... vii

Chapter

I. INTRODUCTION ......................................................... 1

II. REVIEW OF THE LITERATURE ........................................ 5

Advantages of Computerized Testing ............................... 5
Problems with Computer Testing .................................... 6
Examples of Computerized Testing .................................. 8
Findings in the Current Literature ................................. 12
CLEP Examinations and the Computer ............................ 14

III. PROCEDURES .......................................................... 17

Pilot Study .................................................................... 17
Population and Sample .................................................. 19
Instruments ................................................................... 21

Paper-and-pencil CLEP Test ......................................... 21
Computer-Administered CLEP Test ................................. 24

Hardware ................................................................... 24
Software ................................................................... 25

Design ......................................................................... 27
Analysis .................................................................... 28

IV. RESULTS AND DISCUSSION ........................................... 30

Equivalency of Scores Obtained from the
Computer-administered and Paper and
Pencil Forms ................................................................. 30
Test Reliability ............................................................. 33

Equivalence Reliability ................................................. 33
Internal Consistency Reliability ....................................... 34

Discussion ................................................................... 36
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Examination Groups</td>
<td>21</td>
</tr>
<tr>
<td>2</td>
<td>Two-Way Analysis of Variance of CLEP General English Scaled Scores</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>with Examination Group and Order of Administration as Main Effects</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Group Scaled Score Means and Standard Deviations</td>
<td>32</td>
</tr>
<tr>
<td>4</td>
<td>Pearson Product-Moment Correlations Between Computer and Paper-and-</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>pencil Forms for the Entire Sample and for the Computer-first and</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Paper-and-Pencil-First Groups</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Kuder-Richardson Formula 20 Internal Consistency Reliability</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Coefficients</td>
<td></td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>1</td>
<td>English CLEP scaled scores for paper-and-pencil and computer-administered groups and first and second testing sessions</td>
<td>33</td>
</tr>
</tbody>
</table>
Comparison of the Standard and Computerized Versions of the College Level Examination Program General Examination in English Composition

by

Alan L. Muhlestein, Doctor of Philosophy
Utah State University, 1991

Major Professor: Dr. Keith T. Checketts
Department: Psychology

The purpose of this study was to test whether the computer-administered College Level Examination Program (CLEP) General Examination in English Composition produced scores equivalent to those obtained from the traditional paper-and-pencil version. The CLEP examination and its adaptation for computer administration and the results of a pilot study are presented.

The subjects in this study were volunteers who took the CLEP English Composition Examination in order to earn college credit and were randomly assigned to either the computer-first or paper-and-pencil-first groups. Each subject took both forms of the examination with approximately one half of the subjects taking each version first.

Analysis of variance failed to detect a significant effect for test form or an interaction of test form and
order of administration. Equivalence reliability coefficients and internal consistency coefficients also indicated that the computer administration did not significantly alter the results of the examination. In general, the results of this study support the hypothesis that the computer-administered version of the CLEP General Examination in English Composition produces results equivalent to those obtained from the traditional paper-and-pencil version.
CHAPTER I
INTRODUCTION

Computers have been increasingly used to assist in the administration, scoring and interpretation of psychological assessments over the last couple of decades. Computers have been employed with a variety of evaluation instruments including projective and personality assessments, intellectual measures, interest inventories, and achievement tests. With the advent of relatively inexpensive personal computers, "stand alone" computer-assisted testing stations are becoming affordable and plentiful. Software packages to support such computer-assisted testing are also increasing in supply as well as demand.

Several advantages of computerized testing have been reported in the literature. Bartram and Bayliss (1984) suggest that automated testing allows for more efficient use of skilled testing personnel, can minimize the effects that the examiner has on the examinee, eliminates scoring errors, and generates interpretive reports rapidly. Brown (1984) adds that computerized testing also provides consistent, standardized administration which improves reliability.

Even with the advantages of computerized testing and the increased use of automated versions of tests, some questions remain. One question which should be addressed before undertaking serious implementation of computerized
testing is the acceptability of the computer by the consumer. Do examinees actually prefer paper and pencil to computerized administration of the same assessment instrument? Some authors report that this is not the case (see Bartram & Bayliss, 1984; Burke & Normand, 1987; Eaves, 1986; Lukin, Dowd, Plake, & Kraft, 1985; and Moreland, 1985). Hofer and Green (1985) further suggest that there was a lack of convincing evidence that any demographic group is disadvantaged by computerized testing when compared to traditional administration. Eaves (1986) also found with a sample of undergraduate students, faced with computerized testing, that experience with computers offered little, if any, advantage over computer-naive counterparts if the software was "user friendly." User friendly, in part, means that the program provides understandable instructions to the examinee and responses require a minimum of possible data entry options, that is, only a few required keystrokes versus typing lengthy answers.

A more important question involving computer-assisted testing is whether the equivalence of the paper-and-pencil and computerized versions of the test can be demonstrated. Bartram and Bayliss (1984), Burke and Normand (1987), Jackson (1985), Moreland (1985), and Wilson, Genco and Yager (1985), each provide discussions of the need for proof of the equivalence of the two administration methods.
The American Psychological Association (1986) directly states the following with respect to test form equivalence:

When interpreting scores from the computerized versions of conventional tests, the equivalence of scores from computerized versions should be established and documented before using norms or cutting scores obtained from conventional tests. Scores from conventional and computer-administrations may be considered equivalent when (a) the rank orders of scores of individuals tested in alternative modes closely approximate each other, and (b) the means, dispersions, and shapes of the score distributions are approximately the same, or have been made approximately the same by rescaling the scores from the computer mode. (Guideline 16, p. 14)

It is generally agreed that, before an assessment which is developed from an existing paper-and-pencil version is offered for computer administration, the equivalence of the two forms needs to be adequately demonstrated.

Computer-administered forms of the College Level Examination Program (CLEP) tests to earn college credit have recently been developed. Paper-and-pencil versions of the CLEP examinations have been widely employed by colleges and universities for approximately the last 20 years. The computerized versions are alternative forms of the currently used paper-and-pencil versions and are designed to run on readily available microcomputers. As mentioned above, it is necessary to empirically determine the equivalence of the paper-and-pencil and computer-administered versions before results of computerized administration can be used with existing norms or cutoff scores.
The purpose of this study was to investigate the equivalence of results produced by the computerized and paper-and-pencil versions of the CLEP General Examination in English Composition. The research question was, Do the scores obtained by individuals taking the CLEP Examination for English credit differ as a result of the method of administration (i.e., microcomputer versus paper-and-pencil) or as a result of order of administration (i.e., computer or paper-and-pencil form first)? The specific null hypotheses were the following:

1. There is no difference in the scores obtained on the CLEP General Examination in English Composition when administered in paper-and-pencil and computerized formats.

2. There is no difference in the scores obtained on the CLEP General Examination in English Composition due to the order of administration of the computerized and paper-and-pencil versions.
CHAPTER II
REVIEW OF THE LITERATURE

This chapter is primarily devoted to summarizing the relevant literature concerning the use of computers in testing. The important advantages and disadvantages of computer-administered testing are discussed. Examples of the various types of tests that have been adapted for computer administration are also provided and the findings of the current literature in this area are presented. This chapter concludes with a discussion of CLEP examinations and their adaptation to computer administration.

Advantages of Computerized Testing

Computers have been used in some aspect of test administration, scoring, or interpretation for well over twenty years. Smith (1963) suggests that computers can contribute to the areas of examination, testing, and measuring characteristics of individuals. In this early discussion of testing by computer, the author calls for extensive research to promote the usage of computers to facilitate the goals of psychological testing.

Since that time, much has been published concerning the use of computers in testing. Reviews by Bartram and Bayliss (1984), Brown (1984), Burke and Normand (1987), and Mazzeo and Harvey (1988) each address the past developments and current status of computer-assisted testing. Each
discussion contains examples of benefits for computerized testing which include: cost effectiveness through the easy repetition of the computerized version of a test and the increase in efficiency of skilled personnel who employ computer-assisted testing; reduction or elimination of scoring errors; consistent, standardized administrations; increased reliability and speed of production of test reports; and the flexibility of computers to be programmed to administer, score, and interpret a variety of instruments. In addition, these authors also suggest that computer-administered testing is generally favored by, or at least acceptable to, examinees.

Problems with Computer Testing

In contrast to the advantages of computer-assisted testing, some problems and disadvantages are also mentioned. Bartram and Bayliss (1984) point out that paper-and-pencil tests lend themselves to single administration to large groups of people. Similar group administration of computer-administered tests could be prohibitively costly and computer administration is probably best suited to individual tests. The authors suggest that this problem will decrease in the future as offices will likely contain numerous microcomputers linked together in a network which can facilitate group administration of computerized tests. In addition, they
suggest that group tests may be outdated by instruments specifically developed for computer administration.

Burke and Normand (1987) purport that problems with computer-assisted testing can arise from incomplete preparation for the human element. The authors suggest that computer-administered tests should contain clear test instructions, provide time for practice items, have a response key which allows an item to be skipped, and permit the examinee to return to review previously answered or skipped questions.

A larger problem with computer-administered versions of standardized paper-and-pencil tests is confidently employing existing norms or cutoff scores. Burke and Normand (1987) point out that existing norms can be used with the computerized version only if the equivalency of the two versions has been demonstrated. Moreover, they suggest that demonstrating equivalence is more straightforward with ability tests than with personality measures. Further, they state: "More specifically, a high degree of equivalence between different modes of presentation is expected for power (versus speed) tests that are fixed in length, have little if any change in format, and require some form of multiple-choice response" (p. 47).
Examples of Computerized Testing

Computer-administered versions have been developed and evaluated for a number of traditional paper-and-pencil tests. Hoffman and Lundberg (1976) administered computer monitored and paper-and-pencil examinations with true-false, multiple-choice, and matching formats to pharmacy students. The two modes of administration produced equivalent scores in the multiple-choice and true-false formats but significantly different scores were obtained in the matching tests. The authors concluded that recall type items such as multiple choice and true-false questions are better suited to computer administration than are matching questions and other items of a problem solving nature.

Rock and Nolen (1982) report on a pilot study of computerized and standard versions of the Raven Coloured Progressive Matrices Test. Fifteen subjects, ages 7 through 14 years, were administered the computerized version of the test. The mean results of this administration were compared to the mean score of the normative sample and no significant difference was detected. The authors suggest that, in light of their findings, investigation into the application of computerized testing should continue.

Greaud and Green (1986) investigated the equivalence of conventionally-administered and computer-presented speeded clerical tests. They reported that the computer-
administered tests were at least as reliable as the traditional tests and that examinees were faster when responding on the computer. Correlations between computerized and conventional versions were reportedly between .56 and .68 when the items were identical in the paper-and-pencil and computer-administered forms. When the items were presented differently in the two administration modes (i.e., individually on the computer and in blocks of seven items on the paper-and-pencil form) correlations were between .28 and .61. The authors suggest that the difference in item presentation changed the nature of the testing task and was responsible for the lower correlations observed. Mean score differences were detected between the computerized and traditional modes of administration and the authors suggest that either new norms be established or a score transformation be completed to achieve equivalence of scores. In the latter case, traditional norms could be used with scores from computer administration.

The comparability of standard and computer-administered Vocational Preference Inventory was reported by Hodgkinson (1986). Ninety-nine subjects completed both versions of the inventory in a counterbalanced design with one week between administration. The author compared alternate forms reliability coefficients (computer versus paper-and-pencil) with correlations reported in the manual for the original version. The reliability coefficients
compared favorably and the author concludes that the computerized and traditional versions are psychometrically parallel. Use of the automated version of the Vocational Preference Inventory was recommended in place of the standard form where appropriate.

Levy and Barowsky (1986) compared computerized and paper-and-pencil administration of the Harris adaptation of the Goodenough Draw-A-Man Test with 40 computer-naive adolescents. Analysis of the results failed to detect any significant correlation between the two forms of the test. According to the authors, if computerized versions of this test are to be used, the applicability of the old norms must be established or a restandardization of the new method needs to be completed. Levy and Barowsky point out that the Draw-A-Man Test requires a chained, or sequential response for which computerized administration may be poorly suited. They further suggest that other assessments (e.g., achievement tests) which offer discrete responses may be more appropriate for computer administration.

Numerous accounts of the use of computerized versions of personality tests are present in the literature. Two will be discussed here as examples of computer adaptation of paper-and-pencil personality tests. Lushene, O'Neil and Dunn (1974) present their research on the equivalent validity of a computerized version of the Minnesota Multiphasic Personality Inventory (MMPI). Subjects were 63
female college students who completed both forms of the MMPI in a counterbalanced order with a one week separation between administrations. After completing the second administration, subjects indicated which version of the MMPI they preferred. A larger, but not significant, number of subjects stated a preference for the computerized version. Test-retest coefficients were calculated between the two forms and were all reportedly statistically significant at the .01 level and ranged from .45 to .85 for the whole sample. Overall, the authors conclude that the computerized version is as valid as the booklet form.

White, Clements, and Fowler (1985) also reported on a comparison of the standard and computerized administration of the MMPI. Subjects completed the MMPI twice, with an average of one week between administrations. Subject groups completed either the booklet form twice, computerized version twice, or each version once. A preference for the computerized version was expressed by a significant majority of those completing both forms. The computer-administered version required less time to complete than the booklet form. The authors also report that the two formats proved to be equivalent on test-retest correlations, mean scale scores and stability of high-point codes.

Mazzeo and Harvey (1988) reviewed nine studies of computer-administered personality tests including the two
mentioned above. Overall, they found experimental results of the equivalence of the conventional and computerized forms to be mixed suggesting that each test which is adapted to computer administration should be evaluated for equivalence with the traditional form.

Findings in the Current Literature

There are several themes which run through the literature concerning computer-assisted testing. One is the relative advantages of computerized test administration over traditional, paper-and-pencil versions. Some authors remark on the potential benefits that automation can bring to testing (see Bartram and Bayliss, 1984; Brown, 1984; and Burke and Normand, 1987). Cost effectiveness, increased efficiency of skilled testing personnel, consistent, standardized administrations which improves test reliability; reduction or elimination of scoring errors; increased reliability and speed of production of test reports; and the flexibility of computers to be programmed to administer, score, and interpret a variety of instruments are some of the advantages reported by the authors.

Acceptability of computerized tests to examinees has also been a point of discussion in the literature. The concern exists that if an individual being tested is uneasy or anxious about using computers test performance may be
influenced. Various authors (Burke & Normand, 1987; Lukin et al, 1985; Lushene et al, 1974; Moreland, 1985; and White et al, 1985) suggested that examinees react favorably to computer-assisted test administration.

The most common thread running through the literature concerned using cutoff scores or norms based on traditional versions with the automated forms. General agreement has existed that if normative data from a traditional version is to be used with the computerized test, equivalence of the two forms needs to be demonstrated (see American Psychological Association, 1986; Bartram & Bayliss, 1984; Burke & Normand, 1987; Hofer & Green, 1985; and Wilson et al, 1985). Six articles were reviewed here in which the equivalence of automated and traditional forms was investigated (Greaud & Green, 1986; Hodgkinson, 1986; Hoffman & Lundberg, 1976; Levy & Barowsky, 1986; Rock & Nolen, 1982; and White et al, 1985). With the exception of Rock and Nolen (1982), each report suggested that the two versions of the instrument evaluated were found to be equivalent. Rock and Nolen suggested that power (versus speeded) tests with discrete response choices may be most adaptable to computer administration. The computerized CLEP examinations fit into that category.
The CLEP examinations are sponsored by the College Board, scored by Educational Testing Services (ETS), and were designed to cover material taught in introductory-level courses at many universities and colleges. There are 35 CLEP examinations including five general and 30 subject examinations. The general examinations cover the areas of English Composition, Humanities, Mathematics, Natural Sciences, and Social Science and History. All of the examinations were reviewed by faculties at various colleges and universities in order to ensure that they cover the requisite material for the courses represented by the tests. The decision to award credit for CLEP examination is made by the individual college or university, not by the College Board or ETS. Universities or colleges are not required to accept CLEP scores for credit and each institution is free to decide which examinations, if any, will be employed at their facility (College Entrance Examination Board, 1987).

The majority of the CLEP examinations are multiple-choice tests with four or five response options. One point is awarded for each correct answer. Partial credit is deducted for incorrect responses, that is, one-fourth point for a five-choice item and one-third point for a four-choice item. No credit is deducted for unanswered items. The resulting raw score is converted to a scaled score.
between the values of 200 to 800 inclusive. These scaled scores are used to make the results of various forms of CLEP examinations equivalent (College Entrance Examination Board, 1987).

The General Examination in English Composition was designed to assess the skills typically required for successful completion of first-year college composition courses. Two versions of the English Composition test are offered by ETS. One is an all multiple-choice version consisting of 95 five-option items in two sections. The other version is comprised of 45 multiple-choice items and an essay. The all multiple-choice version has been adapted for computer administration (Raffeld, Checketts, Muhlestein & Mazzeo, 1990).

The first section of the all multiple-choice exam contains 55 items intended to focus on logical, structural, and grammatical relationships within sentences. Section II is comprised of 40 items which require the analysis of elements of language within larger prose passages. Many of the items in the second section are based on short prose passages. Forty-five minutes are allotted for completion of each section (College Entrance Examination Board, 1987).

The computer-administered CLEP General Examination in English Composition employs the same question format and multiple-choice responses, a similar number of questions, and the same time limits as the conventional version. Both
forms are scored as described above, and raw scores are converted to scaled scores. The equivalence of the scores obtained from the computerized and paper-and-pencil versions of the CLEP General Examination in English Composition was the subject of this research.
CHAPTER III
PROCEDURES

This chapter describes the procedures employed in this study. Included are descriptions of the population and sample, instrumentation, research design, and data analyses. The chapter begins with a review of the pilot study, which compared the equivalency of the prototype computer-administered form and an existing paper-and-pencil form of the CLEP General English test.

Pilot Study

An earlier version of the computer-administered CLEP General English test was investigated in a pilot study conducted at Utah State University (USU) in 1988. Examination of the results of that study revealed that there was not a statistically significant difference in the scores obtained on the two forms when the paper-and-pencil version was administered first. However, when the computer-administered version was completed first, subjects obtained significantly higher scores on the paper-and-pencil form.

It was assumed that the differences observed in test scores were the result of two separate effects, a practice effect and an effect for mode of administration. The practice effect was defined as an increase in scores on a retest with a similar examination within a relatively short
period of time. The mode of administration effect was defined as scores that are effected by the type of test administered. In this specific case the mode of administration effect that appeared to be present was one in which scores on the computer version of the test were lower than scores on the traditional paper-and-pencil version of the test.

When the computer version was completed first, the two effects combined and resulted in a significantly higher score on the paper-and-pencil version. Conversely, for the paper-and-pencil-first group, the effects essentially canceled each other out which resulted in statistically similar scores for both forms. That is, the expected gain for practice was negated by the lower scores achieved on the computer version.

One possible explanation for the mode of administration effect was that the instructions provided for the computer administered examination were too brief or insufficiently clear to enable examinees to become proficient with the workings of the computer program. Changes designed to remedy this problem have been implemented in the CLEP test to be utilized in this study.
Population and Sample

The CLEP examinations were designed to allow college students an opportunity to earn credit or gain exemption in a particular subject by their performance on a test. Those individuals who obtain at least the minimum cutoff score receive credit for the introductory level course(s) in the area tested. Credit at USU is awarded when a subject obtains a score equal to or higher than the cutoff score set by the Utah System of Higher Education. Credit by CLEP testing is optional and not all students attempt the examinations.

The target population for this study included all those who take the CLEP General English examination for credit or exemption. Since those individuals who took the CLEP General English test at the Testing Center at USU took the tests for credit it is assumed that they were a representative subset of the target population.

Subjects included all those who applied to take the CLEP General English examination at the testing center between July and December 1989 and agreed to complete both the paper-and-pencil and the computer-administered versions of the test.

When a prospective subject applied to take the CLEP General English examination, he or she was informed of the research being conducted and encouraged to complete both forms of the test. As an incentive for participation, the
regular fee for the test was reduced by one half for those who completed both forms, and the subjects were awarded credit based on the higher of the scores from the two tests. This incentive resulted in 115 volunteer subjects from 135 individuals taking the English CLEP test during the period of this study. The most common reason provided by the 20 individuals who declined to participate in the study was that they had insufficient time to complete both test forms. There was no evidence that employing volunteers as subjects biased the subject sample in this study.

Subjects were randomly assigned to groups by order of administration of the two forms of the test, that is, computer-administered version first or paper-and-pencil version first. Fifty-seven subjects completed the paper-and-pencil version first while the other 58 completed the computer-administered version first. The design is shown graphically in Table 1 below. Subjects completed both test versions during the same day and were not informed of the test results until both versions had been completed.
Table 1

**Examination Groups**

<table>
<thead>
<tr>
<th>Groups</th>
<th>First</th>
<th>Second</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer-first (N = 58)</td>
<td>Computer</td>
<td>Paper-pencil</td>
</tr>
<tr>
<td>Paper-pencil First (N = 57)</td>
<td>Paper-pencil</td>
<td>Computer</td>
</tr>
</tbody>
</table>

**Instruments**

**Paper-and-Pencil CLEP Test**

The CLEP General Examination in English Composition is intended to measure the skills required to pass most first year college composition courses. Two versions of the test are available, an all multiple-choice version and a version consisting of multiple-choice items and an essay.

The examination was originally developed in 1963 as part of the College Comprehensive Test and was 60 minutes in length. Substantial modifications were made to the exam in 1978 and, consequently, new reference data were collected. Reference data were collected by administering the examination to students completing first year English courses in the spring of that year. A total of 2,290 students at 37 colleges and universities completed the two
forms of the all multiple-choice examination which are currently in use (College Entrance Examination Board, 1984).

The 1978 version of the test was developed under the direction of a committee of six college professors from different institutions. The committee's objective was to develop a test with high content validity. The committee purports that the test has good content validity for satisfying the requirement of English Composition at most colleges. It is recommended, however, that content validity be further addressed by personnel at the college or university at which the test will be employed (College Entrance Examination Board, 1984).

The reliability of the 1978 multiple-choice versions of the General Examination in English Composition was estimated with the Kuder-Richardson Formula 20 (KR-20). A reliability coefficient is an estimate of the proportion of variance in scores which is due to real differences in ability, as compared to chance or differences resulting from factors other than those being tested. A KR-20 coefficient of .90 or higher is assumed to indicate satisfactory reliability. Reliability estimates calculated with this formula for the two forms developed in 1978 are both .92 (College Entrance Examination Board, 1984).

The standard error of measurement (SEM) was also computed as a measure of test reliability. The standard
error of measurement is an index of the probable range of an examinee's obtained score and the true score. A true score assumes the test measures with perfect accuracy. The SEM for the two 1978 forms of the all multiple-choice test are 28.3 and 27.6 (College Entrance Examination Board, 1984).

The first section of the exam contains items which are intended to focus on logical, structural, and grammatical relationships within sentences. The second section contains items which are intended to focus on an analysis of the elements of language within larger prose passages. Many of the items in Section II are based on short prose passages. Students are given 45 minutes to complete each section of the test.

The raw score (formula score) is obtained by subtracting a fraction of the number wrong from the number correct on each section of the test. For example, one-fourth of a point is deducted for each incorrect answer on the General Examination in English Composition. Subtracting a portion of the incorrect responses was implemented as a correction for guessing and was intended to result in a total formula score near zero when all items were answered by guessing. Formula scores are then linearly converted to scaled scores from 200 to 800 with a mean of 500 (College Entrance Examination Board, 1984).
The 1978 version of the all multiple-choice test was employed in this study. The test was administered to groups of examinees by proctors who were employed at USU Testing Center. The examinees were seated at individual tables in a testing room which was separated from the rest of the testing center by a wall with an observation window. The examinees were provided with a test booklet, answer sheet, and soft lead pencils. The proctor instructed the group in completing the personal data section of the answer sheet and assisted examinees in this process as required. When each examinee had completed the personal data section, the proctor read the standardized instructions to the group and started the timing for Section I. The proctor observed the testing and kept track of the time. When time elapsed for Section I the proctor started the group on Section II and reset the timer. At the end of the time for Section II the proctor collected all test materials and dismissed the group. Testing center staff were also responsible for scoring the examinations in accordance with the procedures described above.

Computer-Administered CLEP Test

Hardware. The computer-administered version of the CLEP General English examination was administered by microcomputers at stand alone test stations in a second examination room at the testing center. IBM PC AT compatible computers with monochrome video display were
employed at each station. The microcomputers contained hard disk drives on which the CLEP test program and results data files were stored.

Each computer station was equipped with a printer which was used to print a report of the examination results. The score printed on the report was compared with the cutoff score by testing center staff to determine whether credit had been earned.

**Software.** All software employed in this study was developed by and copyrighted to Educational Testing Service (ETS) and remains their property. Software for each testing station included a demographic information screen, examination instructions, and a practice test. The demographic information screen allowed the examinee to enter relevant personal data, such as name, date of birth, address, Social Security number, etc. Before leaving the demographic information section, the examinee was queried by the computer about the correctness of the information entered. Examinees could correct errors before proceeding to the next section of the examination.

The instructions presented by the computer program in the current version have been expanded and improved from those in the earlier form of the computer-administered CLEP General English test. For example, the computer presented guidelines encouraging examinees to study the instructions carefully before beginning the test. The features of the
examination program were explained to the subject by messages on the video display. At any time examinees could preview upcoming questions, or return to and revise questions previously answered. A help window, with a list of features available to the examinee, could be displayed on and removed from the screen at any time. Pressing the "T" key followed by the enter key displayed the time remaining for the current section.

The examination was administered in two 45 minute timed sections. When the last question of a section had been answered and if time remained, the program displayed a screen listing all the question numbers and the recorded answers. Any question could then be reviewed and revised by typing the number of that question. This process could be repeated until the examinee was satisfied with all the responses or the time expired. If an examinee finished before the time elapsed, the section could be exited by typing the "F" key and following the prompts provided by the program.

The practice test was expanded in the current CLEP General English test beyond that provided in the earlier form. Again, examinees were encouraged to spend sufficient time with the practice test to try out and become familiar with the options described above. Subjects were allowed up to 30 minutes for the practice test. For the practice and both sections of general test, time was tracked by the
computer. The more detailed instructions and the enhanced practice test were implemented to increase the equivalency of the computer-administered and paper-and-pencil versions of the CLEP General English examination.

Design

Signs describing the option subjects had for completing both versions of the CLEP General English examination for a reduced fee were placed at the door of the testing center and at the registration desk within that office. Those applying to take the English examination were also asked individually by testing center staff if they were aware of the research being conducted and whether they wished to participate. Additional information about the study was provided at that time as requested by candidates. Potential subjects were encouraged to participate but none were pressured to do so. Participation remained voluntary.

CLEP examination candidates who agreed to complete both the paper-and-pencil and computer-administered versions of the English test were randomly assigned to groups by order of administration. Fifty-eight subjects were administered the computer form first and 57 completed the paper-and-pencil version first. Random assignment to these groups was employed to control for the effects of order of administration on test results.
USU Testing Center staff administered the paper-and-pencil version of the CLEP examination to the subjects as outlined above. For the computer-administered version, subjects were instructed about and administered the examination by the computer program as described above. Testing center staff were responsible for printing and interpreting the score report. Following these procedures facilitated consistent, standardized administration of both versions of the examination as required by ETS policy. Once both versions of the examination were completed, credit award decisions were made on the basis of the higher score. Reports of credit earned were forwarded to the USU Records offices per usual testing center procedures.

Analysis

A two way analysis of variance for repeated measures was used to test the hypothesis of no difference in the scores obtained on the CLEP General Examination in English Composition when administered in paper-and-pencil and computer-administered formats.

Correlations between the scaled scores of the computer-administered and paper-and-pencil formats were calculated. Correlations were completed for the entire sample as well as for the computer-first and paper-and-pencil-first groups.
Personnel at ETS retained proprietary rights over all the CLEP testing materials including the test booklets and answer sheets for the paper-and-pencil form and the computer program and output files for the computer-administered form. The paper-and-pencil answer sheets and response data files for the computer-administered version contained item level data from which additional analyses could be performed. From the item level data KR-20 internal consistency reliabilities were computed for each of the four examination conditions, computer-first and second and paper-pencil first and second. The results of these analyses were provided by ETS.
CHAPTER IV
RESULTS AND DISCUSSION

The equivalency of the results obtained from the computer-administered and paper-and-pencil forms of the CLEP General Examination in English Composition was investigated and the results of that investigation are presented in this chapter. The research data were analyzed to detect differences in mean scores due to group assignment, order of test administration, and the interaction of the group assignment and order of administration. Equivalence reliability and internal consistency reliability were also computed. The chapter concludes with a discussion of the findings of this study.

Equivalency of Scores Obtained from the Computer Administered and Paper-and-Pencil Forms

To test the hypothesis of no difference in the scores obtained on the CLEP General Examination in English Composition when administered by computer or by traditional paper-and-pencil methods, a 2 x 2 analysis of variance with repeated measures was employed. The outcome measures were the subjects' scaled scores on the two forms of the test. The two factors (or main effects) in this analysis were experimental group (computer-first or paper-and-pencil-first) and order of administration (computer or paper-and-pencil-administered-first). The analysis of variance
summary table (Table 2) indicates no significant differences between groups ($F = 0.30, p = 0.587$) or for the interaction of group by order ($F = 0.63, p = 0.427$). A significant effect was found for order of administration ($F = 12.65, p = 0.001$). The absence of a significant interaction effect suggests that the mode of administration (i.e., computer or paper-and-pencil) did not differentially affect test scores. Group means and standard deviations may be found in Table 3.

Table 2

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Signif. of F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examination Group (between-subjects) Effects:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Cells Group</td>
<td>113</td>
<td>11107.23</td>
<td>3290.72</td>
<td>.30</td>
</tr>
<tr>
<td>Order of Administration (within-subjects) Effects:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Cells Order</td>
<td>113</td>
<td>887.69</td>
<td>11225.36</td>
<td>12.65</td>
</tr>
<tr>
<td>Group by Order</td>
<td>1</td>
<td>563.52</td>
<td>.63</td>
<td>.427</td>
</tr>
</tbody>
</table>

Examination of Table 3 reveals that for both the computer-first and paper-and-pencil-first groups the mean
Table 3

**Group Scaled Score Means and Standard Deviations**

<table>
<thead>
<tr>
<th>Examination Groups</th>
<th>First Test Taken</th>
<th>Second Test Taken</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Computer Form)</td>
<td>(Paper Form)</td>
<td></td>
</tr>
<tr>
<td>Computer-first</td>
<td>Mean = 483.2</td>
<td>Mean = 500.3</td>
<td>Mean = 481.8</td>
</tr>
<tr>
<td>(n = 58)</td>
<td>SD = 81.5</td>
<td>SD = 76.9</td>
<td>SD = 79.4</td>
</tr>
<tr>
<td>(Paper Form)</td>
<td>(Computer Form)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paper-and-pencil</td>
<td>Mean = 478.8</td>
<td>Mean = 489.6</td>
<td>Mean = 484.2</td>
</tr>
<tr>
<td>First (n = 57)</td>
<td>SD = 75.5</td>
<td>SD = 75.6</td>
<td>SD = 75.4</td>
</tr>
<tr>
<td>Total</td>
<td>Mean = 481.0</td>
<td>Mean = 495.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD = 78.3</td>
<td>SD = 76.1</td>
<td></td>
</tr>
</tbody>
</table>

Score for the second test taken is higher than the mean score for the first test taken. Further, the difference between the means for both groups is similar, approximately 17 points for the computer-first group and nearly 11 points for the paper-and-pencil-first group. A graphic representation of the group means can be seen in Figure 1. This pattern is essentially what would be expected from the practice effect without a significant effect for mode of administration.
Figure 1. English CLEP scaled scores for paper and pencil and computer administered groups and first and second testing sessions.

Test Reliability

Equivalence Reliability

To assess the equivalence reliability of the computer-administered CLEP General English test with the paper-and-pencil form, Pearson product-moment correlations were calculated between the two forms for the entire sample and for each examination group (computer-first and paper-and-pencil-first) separately. As can be seen from Table 4 the correlation between first and second test scores for both groups combined was $r = 0.837$. The within groups
correlation for the computer-first group was $r = 0.856$ and for the paper-and-pencil-first group was $r = 0.849$.

Table 4


<table>
<thead>
<tr>
<th>Group</th>
<th>Correlation</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entire Sample</td>
<td>.837</td>
<td>115</td>
</tr>
<tr>
<td>Computer-first</td>
<td>.856</td>
<td>58</td>
</tr>
<tr>
<td>Paper-First</td>
<td>.849</td>
<td>57</td>
</tr>
</tbody>
</table>

These correlations suggest that the rank order of the scores on the two forms of the test for each of the groups is similar. That is, examinees' scores tended to maintain a similar place within their group ranking for both test forms. From these results, it appears that the mode of administration did not significantly effect the reliability of the test.

**Internal Consistency Reliability**

As reported in the previous chapter, the reliability of the 1978 multiple-choice version of the General Examination in English Composition was estimated with the Kuder-Richardson Formula 20 (KR-20). Using the data collected in this study, personnel at ETS computed KR-20
internal consistency reliability coefficients for the paper-and-pencil and computer versions of the test. The KR-20 reliability estimates range from 0.878 to 0.911 and are presented in Table 5 below.

Table 5
Kuder-Richardson Formula 20 Internal Consistency Reliability Coefficients.

<table>
<thead>
<tr>
<th>Examination Groups</th>
<th>First Test Taken</th>
<th>Second Test Taken</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer-first (n = 58)</td>
<td>.886</td>
<td>.911</td>
</tr>
<tr>
<td>Paper-and-pencil First (n = 57)</td>
<td>.878</td>
<td>.883</td>
</tr>
</tbody>
</table>

From Table 5 it can be seen that the KR-20 coefficients are lower than those obtained for the 1978 version except for the second administration of the computer-first group. The other three coefficients are approximately .88 which is near the level of .90 which was expected for the 1978 sample (College Entrance Examination Board, 1984). It should be pointed out that the lowest of the four KR-20 coefficients was obtained on the paper-and-pencil form administered first, which is the administration
that most closely emulates the 1978 standardization procedure.
Discussion

According to the American Psychological Association (1986), scores obtained from a computer-administered version of an existing paper-and-pencil test can be considered equivalent if the means, dispersions, and shapes of the score distributions are approximately the same and the rank order of scores from the two versions approximate each other. The results presented above suggest that these conditions have been met for the CLEP General Examination in English Composition.

Analysis of variance failed to detect any significant difference in mean scores obtained from the computer-administered and paper-and-pencil forms of the test. Likewise, no significant difference was found for the interaction of examination group affiliation with the order of test administration. The significant difference detected for order of test administration, as presented graphically in Figure 1, is as would be expected from a practice effect. That is, a gain in test scores is anticipated upon retest with a similar examination within a short period of time. Examination of Figure 1 reveals that the gain for each examination group was similar which provides further support that the two forms produce equivalent results.

Correlations of the scores obtained from the two tests suggest that the rank order of the scores from the two test
versions do approximate each other. From Table 4 it can be seen that the correlation coefficients are nearly the same for the entire sample and both of the examination groups. Further, the correlations are high enough in value to meet the second test of equivalence described above.

Finally, internal consistency reliability coefficients were computed for the scores obtained from both groups and both examinations. The KR-20 internal consistency reliability coefficients obtained in this study are close to those which were obtained in the 1978 standardization. In addition, the KR-20 reliability coefficients for the first administration for both examination forms are nearly identical, providing additional evidence for the equivalency of the computer-administered and paper-and-pencil forms of the CLEP General Examination in English Composition.
CHAPTER V
SUMMARY AND CONCLUSIONS

A computer-administered version of the CLEP General Examination in English Composition has been developed and previously investigated in a pilot study to determine the equivalence of the results produced by it and the traditional paper-and-pencil form. The results of the pilot study revealed that the mean test scaled score was lower for the computer administration than for the paper-and-pencil administration. The magnitude of the mean score difference was such that the computer-administered version could not be directly substituted for the paper-and-pencil test using the existing cutoff scores. Modifications designed to reduce this difference in scores were implemented in the computerized version of the test. These modifications included enhanced instructions to familiarize examinees with the workings of the computer program, a provision for marking items for later review, and an expanded practice test which included items of each type included in the examination. It was the equivalence of the scores obtained from the revised computer-administered and traditional paper-and-pencil versions of the CLEP General Examination in English Composition which was the subject of this research.

The results of this study suggest that the modifications made to the computer-administered English
Composition Test have eliminated the mode of administration effect on the average scaled scores for the two groups. The mean difference between paper-and-pencil and computer-administered scores was only about 2.5 points. Further, the computer-first and paper-and-pencil-first groups obtained higher average scores on second administration which is what might be expected from the practice effect.

The pattern of results in the current study is unlike the pattern of results obtained in the pilot study. In the pilot study the mean scores for the paper-and-pencil-first group were essentially the same for the first and second administrations while the mean scores increased significantly for the computer-first group upon retest. It was hypothesized that the practice and mode of administration effects were accumulative for the computer-first group resulting in the large observed gain in mean scores upon retest. It is further hypothesized that the practice and mode of administration effects were approximately equal in magnitude and canceled each other out for the paper-and-pencil-first group resulting in the nearly equivalent mean scores for first and second administrations.

The pattern of results from the current study suggests that the practice effect is present without a noticeable mode of administration effect.
The equivalence reliability of the computer-administered and paper-and-pencil versions of the CLEP General English Test was also investigated. Pearson product-moment correlations were calculated between the two test forms for each examination group and for the entire sample. These reliability coefficients were similar for each group and ranged from .837 for the entire sample to .856 for the computer-first group with the paper-and-pencil-first group at .849. The coefficient values are high enough to suggest that the rank order of scores for the two groups did not differ substantially in this sample. Further, the coefficient values for both experimental groups are similar enough that order of administration seems to have little effect on the rank order of test scores. These results suggest that the computer version produces reliable results when compared with the traditional paper-and-pencil form of the examination.

Personnel at Educational Testing Services computed Kuder-Richardson Formula 20 (KR-20) internal consistency reliability coefficients for the data collected in this study. Their results were similar to those obtained in the 1978 standardization of the paper-and-pencil form of the CLEP General Examination in English Composition. The KR-20 coefficients for the first administration of both forms of the test were nearly identical (.88) and close to the value of .90 which was expected for the 1978
standardization. From these results it is concluded that the computer administration did not alter the internal consistency of the CLEP General English Test.

In general, the results of this study support the hypothesis that the computer-administered version of the CLEP General Examination in English Composition produces results equivalent to the traditional paper-and-pencil version. In view of these findings, the computer version can be confidently utilized as an alternative to the paper-and-pencil version.

Methodological Limitations

Strengths

One strength of this study was the research design employed to test the hypothesis of no difference in the scores obtained from the paper-and-pencil and computer-administered versions of the CLEP General Examination in English Composition. The research design was a two group crossover design in which each subject completed both forms of the test. Approximately half of the subjects took the computer form first with the remainder completing the paper-and-pencil form first. By having each subject complete both test forms a comparison of the effects for mode of administration could be made. That is, any given subject should do about as well on one form as the other if the forms are truly equivalent. The effect for order of
administration of the test forms was controlled by randomly assigning half of the subjects to take each of the tests first and then switching to the other form for the second administration. That way, any effect for taking one or the other form first would be cancelled out.

Another advantage of this design is that it allowed for correlations of test scores between the two forms to be readily computed. Since each subject took both forms it was a simple matter to determine whether the rank order of the scores from the experimental groups differed.

An additional strength of this study was in the subject sample itself. The subjects were individuals applying to, or currently attending Utah State University. Further, each subject requested to take the CLEP General Examination in English Composition for credit at the university. Since the subjects could earn six credits in English Composition by achieving a score at or above the cutoff level, they were representative of those for whom the test was designed, and had adequate motivation to do their best. In other words, it was a real life application of the examination.

The third major strength of the study was actually an interaction of the above two strengths. Since the subjects could earn college English credit they were motivated to do well on the test. The design of the study was such that subjects could earn credit by achieving a score at or above
the cutoff level on either of the two forms. Subjects were not informed of their test scores until they had completed both forms of the test. Consequently, the subjects remained motivated to do well on both forms which added strength to the research design and improved the credibility of the results.

Weaknesses

One weakness in this study was the sample size. Although the subjects were motivated in their completion of the examination and the sample was large enough for statistical analysis, it was not large enough for additional investigations. For example, the sample was too small to partial out subgroups for specific analysis. An investigation of the effect that sex had on test results could be worthwhile. The subjects' previous computer experience is another area in which analysis may be fruitful. Subjects' age, educational level (entering freshmen versus juniors or seniors) and college entrance examination scores (SAT, ACT, etc.) are other factors which could provide useful data. The sample in this study was of insufficient size to partial out such subgroups for meaningful analysis. In one sense, although these analyses may be fruitful, or at least interesting, they are not absolutely necessary. The CLEP examinations are not limited to individuals at one particular level on any of the above dimensions. They are available for any college
student who may benefit from taking the test. Given that fact, a heterogeneous, volunteer sample of actual examinees such as was employed in this study was adequate for the purpose of comparing the results obtained from the two test forms.

Future Research

The results of this study support the application of the computer-administered CLEP General Examination in English Composition as an alternative form to the traditional paper-and-pencil version. Additional research in the future may be useful in determining how individual variables such as previous computer experience, educational level, and so forth, effect the results obtained from the computer-administered CLEP General English Test. Similar studies with this test could also be performed at other universities and colleges to replicate these results with different samples.

The most important observation for future research comes from the difference in the results obtained in this study and the pilot study. The computer-administered form of the CLEP General English Test did not produce results equivalent to those obtained from the paper-and-pencil form. This demonstrates the need to empirically verify that each test adapted for computer administration produces results which are equivalent to the traditional version if
existing normative data and/or cutoff scores are to be employed. Consequently, much research is still to be done in determining which traditional tests can be successfully adapted to computer administration.
REFERENCES


VITA

Alan L. Muhlestein

Candidate for the Degree of

Doctor of Philosophy

Dissertation: Comparison of the Standard and Computerized Versions of the College Level Examination Program General Examination in English Composition

Major Field: Professional-Scientific Psychology

Biographical Information


Education: Bachelor of Science degree with a Psychology Major from Southern Utah State College in Cedar City Utah in 1982; Master of Science in Counseling Psychology, Utah State University, Logan, Utah, 1987; completed the requirements for the degree of Doctor of Philosophy in Combined Professional-Scientific Psychology at Utah State University in 1991.

Internship: American Psychological Association accredited internship at Western State Hospital in Fort Steilacoom, Washington, including rotations in forensic, geriatric, and adult inpatient psychology.