THE DEVELOPMENT AND TESTING OF
A THREE-SECTION CLOZE TEST
OF ENGLISH PROFICIENCY

THESIS

Presented to the Graduate Council of
North Texas State University in Partial
Fulfillment of the Requirements

For the Degree of

MASTER OF ARTS

By

Lauralee Lindholm, B.A.
Denton, Texas
August, 1985

The purpose of this research was to develop and test a three-section cloze test of English proficiency and to norm it for use as a means of level placement. The study sample consisted of ESL students at Brookhaven Community College and the Intensive English Language Institute of North Texas State University, as well as a group of native speakers.

Four types of statistical analysis were used: analysis of variance, Pearson product-moment correlations, a t-Test, and a multiple comparison procedure, the Scheffé test.

The cloze test was sensitive to significant differences between every level at both schools. Subsequently it was normed to a four-level system and score ranges for each level were suggested.
TABLE OF CONTENTS

LIST OF TABLES ........................................ iv

CHAPTER

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

Statement of the Problem ......................... 2
Purpose of the Study .............................. 2
Significance of the Study ......................... 3
Hypotheses ........................................ 3
Basic Assumptions ................................. 4
Limitations ...................................... 4

II.  REVIEW OF RELATED LITERATURE ............... 5

The Unitary Competence-Divisible Competence Controversy ......................... 6
Integrative Testing ............................... 7
Cloze Testing Research ......................... 9
Cloze Test Construction and Administration .................. 10
Reliability and Validity .......................... 12
Cloze Test Variations ............................ 14
Problems and Prospects ........................... 16

III.  PROCEDURES ................................... 19

Population ........................................ 19
Selection of Data Producing Sample ........ 20
The Instrument .................................... 21
Procedure for Collecting Data ................ 22
Design of the Study .............................. 24
Procedures for Analysis of Data ............... 24

IV.  PRESENTATION OF DATA ......................... 26

Hypothesis 1 ....................................... 28
Hypothesis 2 ....................................... 32
Hypothesis 3 ....................................... 33
Hypothesis 4 ....................................... 38
Reliability ........................................ 39
Validity ........................................... 39
Norming the Test .................................. 40
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.</td>
<td>Total Scores by Level.</td>
<td>28</td>
</tr>
<tr>
<td>II.</td>
<td>Analysis of Variance for Total Score</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Denton, Levels 2, 3, 4</td>
<td></td>
</tr>
<tr>
<td>III.</td>
<td>Scheffé Test for Total Score</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Denton, Levels 2, 3, 4</td>
<td></td>
</tr>
<tr>
<td>IV.</td>
<td>Analysis of Variance for Total Score</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Brookhaven, Levels 1, 2, 3, 4</td>
<td></td>
</tr>
<tr>
<td>V.</td>
<td>Kruskal-Wallis One-Way ANOVA for Total Score</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Brookhaven, Levels 1, 2, 3, 4</td>
<td></td>
</tr>
<tr>
<td>VI.</td>
<td>Scheffé Test for Total Score</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Brookhaven, Levels 1, 2, 3, 4</td>
<td></td>
</tr>
<tr>
<td>VII.</td>
<td>Homogeneity of Variance</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Denton and Brookhaven</td>
<td></td>
</tr>
<tr>
<td>VIII.</td>
<td>Pearson Product-Moment Correlations of Total Scores with Level, Denton and</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Brookhaven</td>
<td></td>
</tr>
<tr>
<td>IX.</td>
<td>Section Scores for Denton</td>
<td>34</td>
</tr>
<tr>
<td>X.</td>
<td>Section Scores For Brookhaven</td>
<td>36</td>
</tr>
<tr>
<td>XI.</td>
<td>Reliability Coefficients</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>Brookhaven and Denton</td>
<td></td>
</tr>
<tr>
<td>XII.</td>
<td>T-Test of Total Cloze Test Scores</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Level 5 Denton and Native Speakers</td>
<td></td>
</tr>
<tr>
<td>XIII.</td>
<td>Proposed Cloze Test Score Ranges by Level for a Four-Level Program</td>
<td>41</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION

English as a Second Language (ESL) programs are being developed and expanded all across the U.S., particularly in metropolitan areas. Accompanying this growth in programs have come new developments in textbooks, tests, and teaching techniques. Every facet of ESL has been subject to scrutiny and innovation.

Because placing students in proper levels is central to the ESL teaching process, proficiency testing has been an area of particular interest. A variety of techniques for measuring proficiency have been developed and tested and several key factors in an effective testing procedure have been identified, such as cost, ease of administration, time required, and discriminability at all proficiency levels. However, all too often a test that rates well in one area fails in another.

Brookhaven Community College in Farmers Branch, Texas, has an ESL program in their Continuing Education Division. All courses in this division must pay their own way, and in order to keep student expense down, overhead costs are kept to a minimum. Class time is at a premium since most classes meet only once or twice a week. ESL teachers and administrators have long felt the need for an English proficiency
measure to aid in level placement, but have failed to find one that met budgetary, time, and personnel constraints.

Development of a proficiency test that is inexpensive, short, and easy to administer, would help not only Brookhaven, but other low-budget ESL programs as well. It would provide administrators with a tool for quick and accurate level assignment. It would help teachers by making it possible to group students homogeneously, thus giving a well-defined target level for instruction. Most important, it would benefit students because instruction could be provided at the upper limits of their level of capability. Krashen (1983) identifies a maximally effective language learning situation as one in which students are exposed to language that contains structure just beyond their level of competence. When language is understood, language acquisition takes place and competence is increased.

Statement of the Problem

The problem in this research was to identify a method of language proficiency testing which meets the needs of a non-credit ESL program in the Continuing Education division of an urban community college and then to develop and test an instrument using that method.

Purpose of the Study

The purpose of this research was to establish a range of English proficiency scores for each level of ESL
instruction in a four-level program by means of a testing instrument developed and normed for that purpose.

Significance of the Study

ESL classes which include students with a wide range of abilities are difficult to teach. Inevitably, some of the students are bored while others are frustrated. For optimal learning to take place, students should be grouped on the basis of language ability. Proficiency tests are a great help in evaluating students. In general, the standardized tests already on the market are lengthy and expensive. They do not discriminate very well at the lower levels because they are too difficult. In addition, trained personnel are required for administration and scoring.

Since many ESL programs lack the time, money, and personnel to administer the tests presently available, a simple and economical test is needed. This research focused on developing and testing such a measure of English proficiency.

Hypotheses

For the purposes of this research the following hypotheses were tested in relation to ESL classes in the Continuing Education Division of Brookhaven Community College (hereafter called Brookhaven) and in the Intensive Language Institute of North Texas State (hereafter called Denton).
1. There will be significant differences in total test scores between class levels at both schools.

2. Classes at Denton will be significantly more homogeneous than classes at Brookhaven.

3. There will be significant differences in section test scores between class levels at both schools.

4. There will be a significant correlation between cloze scores and CELT scores.

Basic Assumptions

The following assumptions are implicit in this study.

1. The use of two schools and a variety of class teachers to administer the tests negated the effect of any single test giver.

2. Cloze testing measures global language skills and provides an adequate assessment of overall language facility.

Limitations

This research was limited to two populations: ESL students in the Continuing Education Division of Brookhaven Community College at Farmers Branch, Texas, and students in the Intensive English Language Institute of North Texas State University at Denton, Texas. Students who participated in the research were not randomly selected. They were part of intact classes at both schools.
Educators have made the assumption that degrees of knowing a language exist, especially a foreign language. These degrees of knowing can be referred to as language proficiency levels. Vollmer (1983:29) defines language proficiency as "the generalized ability of a person to make more or less use of a foreign language as an instrument of social interaction in relevant future situations."

In the past, language proficiency was viewed as consisting of a number of distinguishable competencies, and a variety of discrete-point tests were developed in an attempt to measure these aspects of language performance. Oller and Hinofotis (1980) state that the components usually singled out by such tests include listening comprehension, grammatical structure, reading vocabulary, reading comprehension, writing ability, oral fluency, accent, oral grammatical accuracy, oral vocabulary usage, and conversational comprehension.

The trend continued in the direction of identifying and measuring individual, isolated skills for many years. The major proficiency tests on the market today reflect this emphasis: Test of English as a Foreign Language (TOEFL),
Michigan Test of English Language Proficiency (often referred to as the Michigan test), Comprehensive English Language Test (CELT), Center for English as a Second Language (CESL) Placement Battery at Southern Illinois University, etc.

The Unitary Competence-Divisible Competence Controversy

Research in testing in the 1970's indicated that a significant amount of the variance in individual tests could be accounted for by one global factor. Oller and Hinofotis (1980) found that a single factor accounted for no less than sixty-five percent of the total variance in several batteries of proficiency tests. Scholz, et al. (1980) studied twenty-two English language tests and discovered that at least one of the tests in each of the five major categories shared sixty-seven percent or more of its variance with a single global factor. Hisama (1980) showed that a single factor could account for as much as eighty-two percent of the total variance in tests of syntax, listening comprehension, and reading comprehension.

These data all supported what Oller and Hinofotis (1980) call the Unitary Competence Hypothesis. This hypothesis attributes language proficiency to a single global factor which accounts for almost all variations in nearly all language processing tasks. In contrast, the Divisible Competence Hypothesis maintains that proficiency is the
product of several competencies or distinguishable areas of competence which can be isolated and measured.

More recently the trend has been away from the extremes of the two hypotheses, with most researchers agreeing that the true position lies somewhere in between. While there is a recognized general language ability, language skills have a tendency to be developed and specialized to different degrees and/or at different rates. Carroll (1983), for example, recognizes a general language ability but in addition asserts that language skills can be separately recognized and measured.

Bachman and Palmer (1983:168) sum up the issue:

We feel that we have demonstrated strong support for the distinctness of speaking and reading as traits, and thus reject the unitary trait hypothesis of language proficiency. The two causal models . . . however, indicate a sizable portion of communality in all measures, and thus lead us to reject the completely divisible trait hypothesis as well.

Integrative Testing

With the identification of a global language proficiency came a new emphasis in testing--integrative tests. Cloze and dictation tests are recognized by many (cf. Stump 1978, Oller 1979, Cohen 1980) as among the best of this type. A cloze test is a type of fill-in-the-blank test constructed by randomly deleting words from a prose passage and replacing them with blanks. A dictation test, on the other hand, consists of a passage presented orally at a
C. Psychiatry normal rate with pauses at natural boundaries. The student is required to reproduce the passage in written form.

Cloze and dictation tests are referred to as integrative tests because they are not aimed at a single skill but rather at a general language proficiency. They call into play the internalized "grammar of expectancy," which Oller (1973) has argued is the chief mechanism underlying the skills of thinking, understanding, speaking, reading, and writing. Oller (1979) points out that dictation and cloze require linguistic competence which involves the ability to relate sequences of linguistic elements to their appropriate contexts.

Students demonstrate several language skills on integrative tests. Cohen (1980) identifies three types of knowledge required in order to complete a cloze passage correctly: linguistic knowledge, textual knowledge, and knowledge of the world. One or more of these play a part in the selection of a correct response.

Hinofotis (1980) reports that cloze and dictation tests correlate highly with listening comprehension and reading comprehension subtests, both of which are regarded as highly integrative. Stump (1978) states that cloze and dictation tests correlate strongly not only with each other, but also with other widely used language tests. In addition, he cites a significant correspondence of these tests with the Lorge-Thorndike Intelligence Test and the Iowa Test of Basic
Skills. His conclusion is that all are measuring essentially the same thing—global language proficiency.

Cloze Testing Research

The cloze procedure of testing has been the subject of considerable research in the last fifteen years. Several researchers have addressed the question of whether cloze tests can be substituted for more complicated ESL testing procedures while still retaining the same quality of information.

Hinofotis (1980) found that cloze tests scored by the exact-word method correlated with TOEFL at .71 and with CESL at .80. When scored by the acceptable-word method, correlations rose to .79 with TOEFL and .84 with CESL. Cohen (1980) cites a study in which a cloze test was correlated with Foreign Services Oral Interview (FSI) ratings. Correlations of cloze with FSI ranged from .81 to .84. He attributes this high correlation to the fact that both methods call for integrative performance: filling in a cloze blank and speaking both require the student to sort through his store of linguistic knowledge of grammar and vocabulary.

A considerable body of research exists that provides evidence that cloze tests correlate highly with virtually every other type of language test and with nearly every language skill. Researchers are now turning to analysis of
the construct validity of cloze tests in an effort to determine what each blank actually tests. Bachman (1982), for example, used factor analysis to study the trait structure of a cloze test. He found that a cloze passage with rational deletions is capable of measuring both the syntactic and discourse level relationships in a text.

Cloze Test Construction and Administration

Stump (1978) details the procedure for constructing a cloze test. The first step is to select or write a self-contained passage of 350-450 words which is as close to the criterion language style as possible. Second, every \textit{n}th word is systematically deleted.

Aitken (1977) suggests that optimum results are achieved by using a deletion pattern of not less than every fourth word or more than every tenth word. More frequent deletions do not leave enough of the text intact to allow successful reconstruction. Less frequent deletions necessitate a very lengthy document because of the need to include enough blanks to insure reliability.

Bachman (1982), however, notes that random deletion of every \textit{n}th word may be one cause of inconsistent results on cloze testing since random deletion ignores both syntactic and semantic relationships in a text. Alderson (1982:211) says that "perhaps the principle of randomness needs to be abandoned in favor of the rational selection of deletions,"
based upon a theory of the nature of language and language processing."

With the rational deletion procedure a strict nth word deletion pattern is not maintained. Proper nouns, words that invite too many replacements, and low-frequency words lacking common synonyms, are not deleted. "Key" words that provide context for the passage and words that are parts of unfamiliar idioms are also left intact. With these considerations in mind, words are deleted in the vicinity of every nth word. Even with a fixed-ratio (random) formula it is possible to control the type of word deleted by simply editing or rearranging the passage.

Passage difficulty is affected by vocabulary and structure; however, scaling a passage to the level of students does not seem to be of critical importance in ranking them. Oller (1973) reports that a cloze test which was calculated to underestimate to a large extent the level of the students to which it was administered still functioned quite effectively. He concludes that this finding indicates that the cloze procedure has a rather high tolerance to changes in difficulty. Klein-Braley (1983) states that tests may vary in difficulty, but they will still rank examinees in the same order.

As for scoring the test, spelling should not be a factor, as long as the word is recognizable. However, the word must be grammatically correct. For example, a verb that is
in the wrong tense, or a noun that is singular instead of plural, would be scored as incorrect.

Two methods of scoring have been tried successfully. Exact-word scoring requires the examinee to supply the same word that was in the original passage, while acceptable-word scoring gives credit for any word that is contextually appropriate. Researchers have shown that the two methods correlate with each other very highly. Although scores obtained by the acceptable-word method are higher, actual performance is the same.

Stubbs and Tucker (1974) scored a test both ways and found a correlation of .97, which is significant at the .001 level. Hinofotis (1980) achieved the same results and concluded that there is little appreciable difference in information provided by the different scoring methods.

Students are not expected to achieve high scores on most cloze tests. Cohen (1980) points out that a score of forty-three percent on an exact-word cloze is comparable to a score of seventy-five percent on a standard multiple-choice test of reading comprehension.

Reliability and Validity

The reliability and validity of cloze testing have also received careful attention. In a study by Oller (1973), only 7 items out of 150 failed to discriminate significantly. This showed a high degree of internal consistency, with
reliability coefficients ranging from .80 to .98. Validity was established by a correlation of .89 with the TOEFL.

Streiff (1978) points out that cloze is a criterion-referenced measure which allows comparison of a test-taker's performance with a standard level of performance in relation to text of varying levels of difficulty. The standard or criterion level can be established with reference to scores of native speakers or to those of ESL students who have achieved a satisfactory level of competence. Comparing second-language performance with a such a criterion level of performance provides a further indication of the validity of the test.

Brown (1983) reports that when the Kuder-Richardson 21 (KR-21) formula is used to analyze the reliability of cloze tests, the results are substantially and consistently lower than with other methods of analysis. He attributes this lower reliability coefficient to the fact that KR-21 assumes that all correlations are equal and that all items have the same difficulty. These assumptions are quite broad and are actually rarely met.

Yet in spite of the fact that the KR-21 formula may be expected to underestimate the reliability of cloze tests, it is still the most commonly used measurement since it requires only the mean, standard deviation, and number of items for calculation.
Cloze Test Variations

Several variations of the cloze test have been tried successfully, including oral and multiple-choice versions. Hinofotis and Show (1980) experimented with a multiple-choice cloze test in an effort to lessen the subjectivity of acceptable-word scoring, in which word acceptability is determined by the person who scores the tests. Under each blank in the cloze passage, four words were given as possible choices. The incorrect responses (distractors) were chosen from the most frequent incorrect responses made by native speakers on an open-ended version. High-frequency responses which were grammatical in terms of short-range phrase structure, but not contextually acceptable in relation to the full text, were not used.

Results of the cloze test in both the open-ended form and in the multiple-choice form were then correlated with the CESL test. On the open-ended form, exact-word cloze correlated with CESL at .71 and acceptable-word cloze correlated at .74. The correlation between the multiple-choice cloze and CESL was .63. Although the differences between the open-ended scores and the multiple-choice scores may appear fairly large, they were not significantly different (p > .05). This lack of significant difference indicates that the multiple-choice cloze is providing similar information to the open-ended cloze. However, reliability ratings of multiple-choice cloze have been lower than those
of open-ended cloze. Hinofotis and Show (1980) suggest that test length may need to be increased in order to improve the multiple-choice reliability level.

The New Cloze Test (NCT), developed by Hisama (Hisama 1981), purports to measure those aspects of English language proficiency needed by speakers of other languages in order to study at English-speaking schools. An effort was made to minimize the effect of vocabulary on test scores by limiting words deleted to those which are on the Dale List of 3,000 most frequently used words and by replacing difficult words with synonyms of a lower level of difficulty. Complex grammatical structures which only appear in written forms were not used, and potentially misleading grammatical structures such as double negatives were also avoided.

Hisama found that the reliability of the NCT clearly exceeded that of CELT and Reading for Understanding (RFU), a remedial reading test which is a modified version of the Science Research Associates Reading Improvement Placement Test. The NCT also achieved a higher correlation with the Michigan Test than either CELT or RFU.

Cloze tests have proved to be an effective method of proficiency testing both in the standard open-ended version and in the oral and multiple-choice variations. The open-ended version, however, is the easiest to construct and administer while still yielding high correlations with other standardized tests.
Problems and Prospects

One factor that test makers must continually take into consideration is the problem of guessing. Some tests, due to the way they are made up, are more vulnerable in this area than others. Cziko (1980) addresses the problem of guessing and points out that a test must discriminate between low and high ability examinees if it is to be an effective test. Because discrete-point tests are often composed of multiple choice questions, a student can possibly pass an item by chance. He calls this a "gift." Conversely, a high scorer may fail an item well within his competence because of a cleverly disguised distractor. He labels this a "gyp." Whenever a "gift" or a "gyp" occurs, the sensitivity of the test to individual differences decreases.

Hisama (1981) suggests that the problem of guessing among ESL students on multiple choice questions is magnified because students have widely differing backgrounds and their test taking behavior varies greatly. These factors tend to make multiple choice questions less useful in ESL testing.

In contrast to discrete-point tests, "gifts" are rare on cloze tests because the examinee must know a word before he can correctly fill in a blank. A guess based on knowledge is the only one likely to succeed. "Gyps," however, can occur on cloze tests scored by the exact-word method.
because an answer that is perfectly acceptable contextually may be marked wrong. For this reason, some prefer to use the acceptable-word scoring method. The fact that a student may not get credit for an acceptable word is not actually a significant drawback for exact-word scoring, however, since research has shown that the two methods produce virtually the same results in ranking students.

To summarize, the literature shows that the cloze test has great promise in the area of proficiency testing. It is easily scored, takes little time to administer, is inexpensive, and can be designed by an amateur. It is particularly useful in situations where finances are limited, standardized tests are unavailable, or the language program is small.

If standardized results are desired, a standardized cloze test such as the NCT can be used. Use of such a test is more efficient than use of a discrete-point test such as the TOEFL in terms of time, money, and personnel required for administration. Another alternative is to develop and standardize a cloze test by comparing scores of non-native speakers to those of native speakers and to those of ESL students who have achieved a satisfactory level of performance.

Stubbs and Tucker (1974:241) state that "this technique [cloze] constitutes a powerful and economical measure of English-language proficiency for non-native speakers, as
well as a useful diagnostic tool for the classroom teacher." Aitken (1977:66) puts it even more plainly: "The cloze procedure is not a panacea for all ESL testing problems; however cloze tests yield more 'miles per gallon' of sweat spent in test construction than most ESL teachers realize."

This survey of the literature supported the choice of an open-ended cloze test since a cloze test could accurately measure English proficiency and would best meet the needs of a low-budget ESL program. A few standardized cloze tests are on the market, but a single-page test which includes a low-level section would better meet the needs of the Brookhaven ESL program than any of the tests currently available. Construction of such a test and norming it to Brookhaven's four-level program was the subject of this research.
Chapter III

PROCEDURES

This research attempted to measure the English language proficiency of ESL students at Brookhaven Community College in Farmers Branch, Texas, and the Intensive English Language Institute of North Texas State University in Denton, Texas. A three-section open-ended cloze test was developed and used to measure proficiency. This chapter describes the research with reference to the population, selection of the sample, the instrument, design of the study, and procedures for the collection and analysis of data.

Population

The population of this research consisted of all the ESL students in Continuing Education at Brookhaven Community College and all students at the Intensive English Language Institute at North Texas State University, plus a small group of native speakers.

ESL instruction in the Continuing Education Division at Brookhaven is provided on four levels. Placement of continuing students is the responsibility of the ESL teacher, whereas assessment and placement of new students is done by the Continuing Education staff after interviews with them during registration.
Students at Brookhaven attend two weekly classes for a total of three to four hours of instruction per week. Most classes are held at night for the benefit of students who work, and classes include many older adults of long-term residence in the U.S. Some students are learning English with the goal of enrolling in college; most, however, simply want to be able to communicate.

In addition to the ESL classes in Continuing Education at Brookhaven, there is an ESL class in the credit division. Students in this class must have achieved a score of at least 525 on the TOEFL or must be Texas residents. These students were subjects for the preliminary version of the test.

Students in the Intensive English Language Institute at Denton attend classes five half-days a week. The majority have come to the United States fairly recently, have higher education as a goal, and are of college-age. Denton offers instruction on five levels, with the addition of a Level Zero for students with no proficiency at all in English. Many students make sufficient progress to enable them to complete the program in one year. Level placement is based on standardized test scores coupled with teacher evaluations.

Selection of Data-Producing Sample
The population consisted of 189 students in ESL Continuing Education classes at Brookhaven college, 69 students
in the Intensive English Language Institute at Denton, and
16 native speakers. Students from all four levels of in-
struction at Brookhaven were included in the sample. At
Denton, where class participation was at the discretion of
the teacher, students from Levels Two, Three, and Five par-
ticipated. Native-speaking volunteers were found in an
office and a church library.

The Instrument

The instrument used to measure English proficiency was
a written cloze test. Written cloze, as opposed to oral
cloze or dictation, was chosen for several reasons.

1. Both oral cloze and dictation require uniform
spoken presentation for standardized results. This necess-
itates tape recording and playback equipment which adds to
both the expense and the complexity of administration.

2. An oral test is less practical than a written test
for testing students individually, since it requires that an
administrator be present during the entire test.

3. The scoring of a written cloze test by the exact-
word method is easier than the scoring of an oral cloze or
a dictation passage.

The instrument was made up of three 125-150 word cloze
passages: elementary, intermediate, and advanced (i.e.,
beginning college material). Difficulty was determined by
word length, word frequency, sentence length, grammatical/
syntactic structure, and the amount of cultural knowledge required. Passages were devised by this author with these criteria in mind.

Rather than using a fixed ratio approach of word deletion, a rational deletion ratio of approximately five words per blank was used. As previous studies have suggested (Cohen 1980 and Alderson 1983), words that were key to the context, or that could be replaced by too many different words, were not deleted. Each of the three passages contained twenty-four blanks, for a total of seventy-two.

The test was printed on the front and back of a single legal-size sheet of paper which was then folded in half (see Appendix A). The first page gave instructions for taking the test, plus a sample passage for students to complete. This was followed by the three passages of increasing difficulty.

Thirty minutes were allotted for the test—a time limit being warranted by the fact that response time is a factor in language proficiency (Cohen 1980 and Carroll 1983). Most were ready to hand in their tests well within this time period, either having finished, or having filled in as much as they were able.

Procedure for Collecting Data

Written permission to conduct this research was request
ed from the Vice President of Instruction at Brookhaven
and from the Director of the Intensive English Language Institute at Denton. Each school sent a set of guidelines and questions to be answered. In response to these communications the research proposal was outlined and basic issues were addressed (see Appendices B & C). Subsequently, permission was received from both schools to include their students in the research.

A preliminary version of the cloze test was given in December, 1984, to 13 students in the credit ESL class at Brookhaven. The classroom teacher administered the test using the instructions which were supplied for that purpose (see Appendix D).

As a result of this pilot version, several modifications were made in the test. These included changes in words deleted, vocabulary, instructions, and reduction in number of blanks on each section from twenty-five to twenty-four (see Appendix E).

The revised version of the cloze test was given to students in ESL classes at Brookhaven and Denton during the first week of the spring semester of the 1984-85 school year. Each ESL teacher administered the test to his or her class. Students who were absent when the test was given were not included in the research. The test was taken by 189 students at Brookhaven and 69 students at Denton, for a total of 258 students. Later in the spring 16 native speakers also took the test.
Design of the Study

This research was based on an ex-post-facto design. Students were classified both as to instructional level and as to school. Their degree of English proficiency was determined by a three-section cloze test. Students from four levels at Brookhaven participated, while students from only three levels at Denton were included. No one from Level One or Level Four at Denton participated in the study. Since Level Four is the top level at Brookhaven and Level Five is the top level at Denton, Level Five at Denton was coded as Level Four for purposes of comparison.

Four types of statistical analysis were utilized in this research: ANOVA, Pearson Product-Moment Correlation, t-Test and a Multiple Comparison procedure, the Scheffe Test. The Scheffé test was used because of its advantage in not being limited to samples of equal size and to simple pairwise comparisons. Also of interest were the homogeneity of variance scores included in the statistical procedures.

Procedures for Analysis of Data

Tests were scored by two methods. With the exact-word method, each word that matched the key received one point. On each section of the test some blanks could be filled by two words that were equally acceptable in the context. Rather than avoiding use of these blanks, both words were listed on the key as correct (see Appendix F).
Subtotals were calculated for each of the three sections, and these were then added to give a total score. The tests were scored a second time using acceptable-word scoring. With this method, any word that was contextually appropriate received credit. Totals for each section, as well as an over-all total, were derived.
CHAPTER IV

PRESENTATION OF DATA

This research dealt with English proficiency in relation to level placement of ESL students in two schools. English proficiency was determined by a three-section cloze test. Of the 258 students who took the test, 16 were not able to fill in any blanks at all, 40 attempted only the first section, 60 worked on two sections, and 142 tried to fill in all three sections. Most of those unable to respond at all were in Level One at Brookhaven.

The purpose of this research was to determine whether or not significant differences existed between mean scores on the cloze test relative to class level. Total scores were analyzed as well as scores for each section. Also the homogeneity of variance of classes at each school was studied in order to see what effect the method of class placement had on class makeup.

In order to establish the validity of the cloze test as a measure of English proficiency, CELT scores were correlated with cloze scores in those cases where students had taken both tests. In addition, scores of students were compared with those of native speakers as another measure of validity.
All calculations were done with the Statistical Package for Social Sciences (SPSS) at the North Texas State University Computer Center. Each hypothesis was restated in the null form for testing, and the level of significance was reported for each test. The level of significance below which a hypothesis would be rejected was arbitrarily set at the .05 level.

Prior to testing the hypotheses, the exact-word and the acceptable-word methods of scoring were compared to see if scoring method would significantly affect test results. A Pearson product-moment correlation between exact-word totals and acceptable-word totals indicated that the two methods produced almost identical results ($r = .99, p < .0001$). This finding is consistent with other studies reported in the literature (Stubbs and Tucker 1974 and Hinofotis 1980).

Since the correlation between the two scoring methods was so high, exact-word scoring, with its ease and objectivity, was chosen as the preferred method. Exact-word scores, then, were used to test each of the four research hypotheses. Thus, total scores cited in the data, as well as scores for Sections A, B, and C, all refer to exact-word scores.

The rest of this chapter is devoted to a discussion of the findings on the four research hypotheses. To aid in the discussion of results, statistical summary tables for each hypothesis are presented.
Hypothesis 1

H1: There will be significant differences in total test scores between class levels at both schools.

Because levels do not correspond exactly at the two schools, each school was treated separately in testing this hypothesis. In spite of the differences in levels, however, the means for Levels 3 and 4 at the two schools were very similar (see Table I). No students from Level 1 at Denton participated in the study; hence that cell in the matrix was left blank. The mean of Level 2 at Denton was higher than the corresponding mean at Brookhaven, but that was to be expected since Brookhaven has only one level below Level 2 and Denton has two, both a Level 0 and a Level 1.

<table>
<thead>
<tr>
<th>Level</th>
<th>School</th>
<th>N</th>
<th>X</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Brookhaven</td>
<td>81</td>
<td>2.79</td>
<td>5.37</td>
</tr>
<tr>
<td>2</td>
<td>Brookhaven</td>
<td>58</td>
<td>15.32</td>
<td>12.35</td>
</tr>
<tr>
<td></td>
<td>Denton</td>
<td>21</td>
<td>18.76</td>
<td>6.61</td>
</tr>
<tr>
<td>3</td>
<td>Brookhaven</td>
<td>43</td>
<td>25.81</td>
<td>11.67</td>
</tr>
<tr>
<td></td>
<td>Denton</td>
<td>20</td>
<td>25.80</td>
<td>4.35</td>
</tr>
<tr>
<td>4</td>
<td>Brookhaven</td>
<td>7</td>
<td>40.43</td>
<td>8.48</td>
</tr>
<tr>
<td></td>
<td>Denton</td>
<td>28</td>
<td>40.71</td>
<td>7.71</td>
</tr>
</tbody>
</table>

Native  16  59.13  5.49
Hypothesis 1 was restated in the null form and was then tested by means of a Pearson product-moment correlation. Correlation of total test score with level for the Denton sample showed a significant correlation ($r = .82, p < .0001$). Brookhaven's correlation was lower, but was equally significant ($r = .74, p < .0001$). These results indicate that these cloze test scores correlated highly with class level. However, a significant correlation alone is not an adequate basis for accepting the research hypothesis. To determine if scores were significantly different between levels, a one-way ANOVA was utilized.

<table>
<thead>
<tr>
<th>Source</th>
<th>D.F.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>2</td>
<td>6213.5</td>
<td>3106.7</td>
<td>78.2</td>
<td>.0001</td>
</tr>
<tr>
<td>Within Groups</td>
<td>66</td>
<td>2622.7</td>
<td>39.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>68</td>
<td>8836.2</td>
<td></td>
<td></td>
<td>.238</td>
</tr>
</tbody>
</table>

An examination of Table II shows that there were differences between levels which were significant at the .0001 level. In addition, the Cochran's C value indicates that the equal variance assumption was met, thereby justifying a parametric measure.

To pinpoint the areas of significant difference, a Scheffé test was performed. This information is presented in Table III.
### TABLE III
**SCHEFFÉ TEST FOR TOTAL SCORE**
**DENTON, LEVELS 2, 3, 4**

<table>
<thead>
<tr>
<th>N</th>
<th>Mean</th>
<th>Level</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>18.76</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>25.80</td>
<td>3</td>
<td>*</td>
</tr>
</tbody>
</table>
| 28  | 40.71| 4     | *     | *

(*) Denotes pairs of groups significantly different at .05 level.

An examination of Table III shows that there were significant differences between all class levels included in the Denton sample. These findings further support Hypothesis 1.

Hypothesis 1 was tested a second time in relation to the Brookhaven sample by means of a one-way ANOVA. These results are presented in Table IV.

### TABLE IV
**ANALYSIS OF VARIANCE FOR TOTAL SCORE**
**BROOKHAVEN, LEVELS 1, 2, 3, 4**

<table>
<thead>
<tr>
<th>Source</th>
<th>D.F.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>3</td>
<td>21073.7</td>
<td>7024.6</td>
<td>75.5</td>
<td>.0001</td>
</tr>
<tr>
<td>Within Groups</td>
<td>185</td>
<td>17300.5</td>
<td>93.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>188</td>
<td>38374.2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cochran's C = .39, p = .007

Table IV reveals that significant differences in cloze scores existed between the levels at Brookhaven as well. Cochran's C, however, indicates that the assumption of the homogeneity of variance between the groups was not valid for
this sample. Because the equal variance assumption was not met, it was necessary to utilize a non-parametric measure: the Kruskal-Wallis one-way ANOVA. Data from this procedure are given in Table V.

<table>
<thead>
<tr>
<th>Mean Rank</th>
<th>Cases</th>
<th>Level</th>
<th>Chi-Square</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>49.32</td>
<td>81</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>112.71</td>
<td>58</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>145.64</td>
<td>43</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>176.79</td>
<td>7</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>189</td>
<td></td>
<td>113.94</td>
<td>.0001</td>
</tr>
</tbody>
</table>

Corrected for Ties 115.64 .0001

Chi-Square, when corrected for ties, was significant at the .0001 level, confirming that there were significant differences between levels at Brookhaven. A Scheffé test was used to locate specific areas of significant difference. Results of the Scheffé test are presented in Table VI.

<table>
<thead>
<tr>
<th>N</th>
<th>Mean</th>
<th>Level</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>81</td>
<td>2.79</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>15.32</td>
<td>2</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>25.81</td>
<td>3</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>40.43</td>
<td>4</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

(*) Denotes pairs of groups significantly different at .05 level.
As at Denton, the Scheffé test showed that all class levels at Brookhaven were significantly different at the .05 level. Because ANOVA and Scheffé scores revealed significant differences on mean cloze test scores between all levels at both schools, we can accept Hypothesis 1 in the research form and reject it in the null form.

Hypothesis 2

H2: Classes at Denton will be significantly more homogeneous than classes at Brookhaven.

The two schools could not be directly compared with each other statistically as to degree of homogeneity. When computing Analyses of Variance, however, both schools were compared to the general population to determine if they met the assumption of homogeneity implicit in the procedure. The resulting Cochran's C scores for the two schools can be compared, along with their degree of significance, thus making it possible to infer relative homogeneity.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Cochran's C</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denton</td>
<td>69</td>
<td>.45</td>
<td>.238</td>
</tr>
<tr>
<td>Brookhaven</td>
<td>189</td>
<td>.39</td>
<td>.007</td>
</tr>
</tbody>
</table>

Table VII summarizes the homogeneity of variance figures from ANOVA statistics in Tables II and IV for the two schools. The table shows a Cochran's C value for Denton
which is not significant. This lack of significance indicates that test scores did not violate the assumption of homogeneity of variance between groups: students in classes at Denton were not significantly different from the general population.

The Cochran's $C$ value for Brookhaven, however, is significant at the .01 level. This significance reflects a violation of the assumption of homogeneity and indicates that classes at Brookhaven had a greater variance within them than could be expected by chance.

The fact that Denton classes were within the expected range of homogeneity, while Brookhaven classes were not, is an adequate basis for rejection of Hypothesis 2 in the null form and acceptance of it in the research form.

Hypothesis 3

$H_3$: There will be significant differences in section test scores between class levels at both schools.

Pearson product-moment correlations of scores on all three sections of the test were performed in relation to class level for each school. Results are shown in Table VIII. Examination of this table shows that all correlations between section and level were significant at the .0001 level at both schools. Yet while every section correlated highly with level, no section at either school correlated as highly with level as the total score did. These figures
indicate that although scores of individual sections were
good measures of English proficiency, the total score was a
better measure.

In order to determine which sections discriminated be-
tween which levels, and to what degree, the data for each
school was examined in more detail.

### TABLE VIII
PEARSON PRODUCT-MOMENT CORRELATIONS
OF TOTAL SCORES WITH LEVEL
DENTON AND BROOKHAVEN

<table>
<thead>
<tr>
<th>School</th>
<th>Section</th>
<th>N</th>
<th>r</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denton</td>
<td>A</td>
<td>69</td>
<td>.70</td>
<td>.0001</td>
</tr>
<tr>
<td>Denton</td>
<td>B</td>
<td>69</td>
<td>.72</td>
<td>.0001</td>
</tr>
<tr>
<td>Denton</td>
<td>C</td>
<td>69</td>
<td>.76</td>
<td>.0001</td>
</tr>
<tr>
<td>Denton</td>
<td>Total</td>
<td>69</td>
<td>.82</td>
<td>.0001</td>
</tr>
<tr>
<td>Brookhaven</td>
<td>A</td>
<td>189</td>
<td>.72</td>
<td>.0001</td>
</tr>
<tr>
<td>Brookhaven</td>
<td>B</td>
<td>189</td>
<td>.69</td>
<td>.0001</td>
</tr>
<tr>
<td>Brookhaven</td>
<td>C</td>
<td>189</td>
<td>.55</td>
<td>.0001</td>
</tr>
<tr>
<td>Brookhaven</td>
<td>Total</td>
<td>189</td>
<td>.74</td>
<td>.0001</td>
</tr>
</tbody>
</table>

### TABLE IX
SECTION SCORES FOR DENTON

<table>
<thead>
<tr>
<th>Level</th>
<th>Section</th>
<th>N</th>
<th>X</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>A</td>
<td>21</td>
<td>8.38</td>
<td>3.84</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>21</td>
<td>7.19</td>
<td>2.99</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>21</td>
<td>3.19</td>
<td>2.89</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>20</td>
<td>10.95</td>
<td>2.80</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>20</td>
<td>9.55</td>
<td>2.01</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>20</td>
<td>5.40</td>
<td>2.35</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>28</td>
<td>15.82</td>
<td>3.15</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>28</td>
<td>13.61</td>
<td>2.85</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>28</td>
<td>11.28</td>
<td>3.21</td>
</tr>
</tbody>
</table>
Information on section scores for Denton is presented in Table IX. Data in this table show that students in each level at Denton scored the highest on the elementary section (A) and the lowest on the advanced section (C), with scores on the intermediate section (B) falling somewhere in between.

To determine definitively the degree of effectiveness of section scores in differentiating between levels, a one-way ANOVA and Scheffé test were calculated for each section. These procedures revealed that in some cases section scores were not sensitive to differences between levels to a significant degree. Because section scores did not prove to be as effective in discriminating between levels as total scores, data regarding section scores have not been presented in detail, but have been summarized in the text instead.

ANOVA results on section scores for the Denton sample indicate that all three sections of the test showed significant differences between levels. In addition, variance was within the normal range. However, the Scheffé procedure reveals that while Section A discriminated between Levels 2 and 4 and between Levels 3 and 4, it failed to discriminate between Levels 2 and 3 to a significant extent. Section B showed significant differences between all levels. Section C, like Section A, discriminated between Levels 2 and 4 and between Levels 3 and 4, but did not distinguish between
Levels 2 and 3 to a significant degree. Thus the Scheffé test reveals that for the Denton sample, only Section B was able to differentiate between all levels of proficiency to a significant extent.

Data on section scores for Brookhaven are presented in Table X. This table shows that at every level mean scores were highest on Section A and lowest on section C. These results are similar to results for Denton. A comparison of standard deviations of mean scores on each section from Tables IX and X reveals that the standard deviations of Brookhaven were consistently larger than those of Denton. This finding is in line with Hypothesis 2.

<table>
<thead>
<tr>
<th>Level</th>
<th>Section</th>
<th>N</th>
<th>X</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>81</td>
<td>2.22</td>
<td>3.78</td>
</tr>
<tr>
<td>1</td>
<td>B</td>
<td>81</td>
<td>.43</td>
<td>1.33</td>
</tr>
<tr>
<td>1</td>
<td>C</td>
<td>81</td>
<td>.14</td>
<td>1.11</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>58</td>
<td>7.86</td>
<td>5.52</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>58</td>
<td>5.44</td>
<td>5.20</td>
</tr>
<tr>
<td>2</td>
<td>C</td>
<td>58</td>
<td>1.79</td>
<td>3.28</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>43</td>
<td>13.49</td>
<td>4.86</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>43</td>
<td>8.60</td>
<td>5.06</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>43</td>
<td>3.72</td>
<td>4.19</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>7</td>
<td>17.14</td>
<td>2.79</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
<td>7</td>
<td>13.57</td>
<td>4.42</td>
</tr>
<tr>
<td>4</td>
<td>C</td>
<td>7</td>
<td>9.71</td>
<td>2.81</td>
</tr>
</tbody>
</table>
An analysis of Brookhaven section scores was done by means of a One-way ANOVA to see how well they discriminated between class levels. On all three sections the results were significant at the .0001 level for the Brookhaven sample, just as they were for the Denton sample. However, because section scores were not as effective as total scores in differentiating between levels, results have been summarized in the text rather than being presented in detail.

In testing Hypothesis 1, ANOVA for the Brookhaven sample revealed that total scores failed to meet the assumption of homogeneity of variance. Similarly, ANOVA results for Brookhaven for each section of the test revealed that the assumption of homogeneity of variance had been violated in every case. This finding was to be expected since total score is a product of section scores.

To pinpoint areas of significant difference between level and section scores at Brookhaven, a Scheffé test was performed in conjunction with ANOVA. Scheffé showed significant differences between all class levels on Sections B and C and between all levels except Levels 3 and 4 on Section A. However, in view of the lack of homogeneity of variance, a non-parametric measure would be required in order to validate these results.

Since Sections A and C did not discriminate significantly at all levels at both schools, and no section score discriminated between levels as well as the total score,
further analysis of section scores was deemed to be unproductive.

These results for Denton and Brookhaven show that only Section B was effective in differentiating between all levels at both schools. Therefore the research hypothesis can be retained for Section B but must be rejected for Sections A and C. While Section B did discriminate to a significant extent, it did not correlate as highly with class level as total score did (see Table VIII).

These results indicate that the cloze test is most effective as a three-section test: the total score is a better indicator of proficiency than any section score. Sections A and C did not discriminate significantly in all instances. Section B proved to be an adequate measure of proficiency, but the total score was shown to be a better measure. Because the test is short enough to be administered in a minimum amount of time, the instrument should remain intact as a three-section test.

Hypothesis 4

H4: There will be a significant correlation between cloze scores and CELT scores.

Of the 69 students at Denton, 32 had taken the CELT test within the last five months. A Pearson product-moment correlation between total cloze test scores and CELT scores showed a significant correlation ($r = .79$, $p < .0001$). On
the basis of this correlation the null hypothesis was re-
jected and the research hypothesis was accepted.

Reliability

Reliability was established by use of the Kuder-
Richardson 21 formula. Data from this procedure are given
in Table XI. As indicated in the literature (Brown, 1983),
KR-21 scores are usually somewhat low on cloze tests because
each blank is not necessarily of equal value. Nevertheless,
both samples resulted in high reliability coefficients by
this method.

<table>
<thead>
<tr>
<th>Measure</th>
<th>School</th>
<th>Possible Score</th>
<th>N</th>
<th>X</th>
<th>SD</th>
<th>Reliability Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>KR-21</td>
<td>Brookhaven</td>
<td>72</td>
<td>189</td>
<td>13.3</td>
<td>14.2</td>
<td>.96</td>
</tr>
<tr>
<td>KR-21</td>
<td>Denton</td>
<td>72</td>
<td>69</td>
<td>29.7</td>
<td>11.4</td>
<td>.88</td>
</tr>
</tbody>
</table>

Validity

The correlation of the cloze test and the CELT test in
regard to Hypothesis 4 provided one measure of the cloze
test's validity. Validity was further assessed by means of
a t-Test comparing mean scores of a group of native speakers
with those of students in Level 5 (coded as Level 4 in Table
I for comparative purposes), the highest level at Denton.
Results of the t-Test are presented in Table XII.
TABLE XII
T-TEST
TOTAL CLOZE TEST SCORES
LEVEL 5 DENTON AND NATIVE SPEAKERS

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>( \bar{X} )</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denton - Top level</td>
<td>28</td>
<td>40.71</td>
<td>7.71</td>
</tr>
<tr>
<td>Native speakers</td>
<td>16</td>
<td>59.13</td>
<td>5.49</td>
</tr>
</tbody>
</table>

\[ t = 9.20 \quad p < .0001 \]

An examination of Table XII shows a significant difference between Level 5 students and native speakers. The achievement of significantly higher scores by native speakers is further evidence that the cloze test is a valid test of proficiency in English.

**Norming the Test**

One purpose of the research was to analyze the data and suggest norms for each level of a four-level program. In norming the test, two factors were considered: both means and standard deviations were of central importance. The smaller standard deviations of Denton, where classes were more homogeneous, were used. Since Denton has two levels below Level 2, the mean for Level 2 at Denton was predictably higher than the mean for Level 2 at Brookhaven.

Table XIII gives suggested means and ranges of this cloze test for each class level in a four-level system. Scores are based on the exact-word scoring method.
TABLE XIII
PROPOSED CLOZE TEST SCORE RANGES BY LEVEL
FOR A FOUR-LEVEL PROGRAM

<table>
<thead>
<tr>
<th>Level</th>
<th>Brookhaven X</th>
<th>SD</th>
<th>Denton X</th>
<th>SD</th>
<th>Proposed X</th>
<th>SD</th>
<th>Cloze Score Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.8</td>
<td>5.4</td>
<td>-</td>
<td>-</td>
<td>3.5</td>
<td>3.5</td>
<td>0 - 7</td>
</tr>
<tr>
<td>2</td>
<td>15.3</td>
<td>12.4</td>
<td>18.8</td>
<td>6.6</td>
<td>13.5</td>
<td>5.5</td>
<td>8 - 19</td>
</tr>
<tr>
<td>3</td>
<td>25.8</td>
<td>11.7</td>
<td>25.8</td>
<td>4.4</td>
<td>25.5</td>
<td>5.5</td>
<td>20 - 31</td>
</tr>
<tr>
<td>4</td>
<td>40.4</td>
<td>8.5</td>
<td>40.7</td>
<td>7.7</td>
<td>40.0</td>
<td>8.0</td>
<td>32 - 72</td>
</tr>
</tbody>
</table>

Examination of individual test scores of the Brookhaven sample shows that 85 students scored 4 or less on Section A. Of these students, only 3 went on to score 8 or more on the total test. Thus, 82 of these 85 students would be placed in Level 1 on the basis of the suggested score ranges in Table XIII. The 3 remaining students who went on to score 8 or more to qualify for Level 2, spent a portion of their time on Sections B and C. If they had been given only Section A, they would have concentrated on it and may have gotten a higher score on that section as a result.

In view of these findings, a recommendation was made that new students be given only part A. If they score 5 or more correct, they should be given the remainder of the test. Those who score less than 5 on Part A should be placed in Level 1 with no further testing.
A three-section cloze test was designed as a measure of English proficiency. The test was taken by 69 students at the Intensive English Language Institute in Denton, Texas; 189 ESL students at Brookhaven Community College in Farmers Branch, Texas; and 16 native English speakers.

One purpose of the research was to determine if this cloze test was sensitive to differences in proficiency level at both schools to a significant degree. Both total scores and section scores were analyzed in this regard.

The second purpose of this research was to see if the method of placement of students at both schools affected the homogeneity of class makeup to a significant extent.

The third purpose of this research was to establish the reliability and validity of this cloze test as a measure of English proficiency.

The fourth purpose was to norm the test to a four-level ESL program and to provide suggested score ranges for each level. Since students at Denton had been placed in each level on the basis of proficiency test scores, they were used as a standard for the norming procedure.
The instrument which was prepared to measure English proficiency was a cloze test consisting of three passages—elementary, intermediate, and advanced. Words were systematically deleted from these passages on the basis of a rational deletion procedure and students were asked to fill in each blank with the word that seemed best to them.

For the purpose of statistical analysis, this research utilized analysis of variance, correlation coefficients, a multiple comparison procedure and the t-Test. Through the use of these statistical analyses, the hypotheses were tested for significant differences between the means and for significant correlations.

Research Findings on Hypothesis 1

Hypothesis 1 stated that there will be significant differences in total test scores between class levels at both schools. The two schools were analyzed separately by means of analysis of variance, and significant differences were found at both schools. The Scheffé test was used to locate areas of difference, and it showed that there were significant differences between all levels at both schools. As a result of these findings, the research hypothesis was accepted.

Analysis of variance also indicated that the Brookhaven sample did not meet the assumption of homogeneity of variance, thus entailing the use of a non-parametric measure.
The Kruskal-Wallis analysis of variance confirmed that there were areas of significant difference in the Brookhaven sample.

Research Findings on Hypothesis 2

Research Hypothesis 2 stated that classes at Denton will be significantly more homogeneous than classes at Brookhaven. There was no measure for direct comparison of the two schools as to homogeneity; however, analysis of variance for each school gave a Cochran's C measure of homogeneity of variance. The Denton sample did not differ significantly from the degree of homogeneity expected of the general population, but the Brookhaven sample did. This finding provided support for the theory that the Denton sample was significantly more homogeneous than the Brookhaven sample, and Research Hypothesis 2 was accepted.

Research Findings on Hypothesis 3

Research Hypothesis 3 stated that there will be significant differences in section test scores between class levels at both schools. Research Hypothesis 3 was retained for section B, the intermediate section, but was rejected for sections A and C, the elementary and advanced sections.

Analysis of variance indicated significant differences at Denton on all three sections. A multiple-comparison procedure, the Scheffe' test, confirmed that there were significant differences between the mean scores of Section B
on all levels. Mean scores of Section A and C, however, failed to discriminate between Levels 2 and 3.

The Brookhaven sample showed similar results. Analysis of variance indicated significant differences in mean scores on all three sections. The Scheffé test indicated that Sections A and B were good discriminators on all levels, but Section C did not discriminate between levels 3 and 4. Only Section B produced significant results at both schools and no section score for either school correlated as highly with class level as the total score did. Thus, in spite of the fact that Research Hypothesis 3 was accepted for Section B, the results support leaving the test intact as a three-section measure.

Research Findings on Hypothesis 4

Research Hypothesis 4 stated that there will be a significant correlation between cloze scores and CELT scores. Test scores of the 32 students at Denton who took both tests were correlated. A significant correlation was found and Research Hypothesis 4 was accepted.

Conclusions

The following conclusions are based upon findings of the research:

1. The instrument of this research, a three-section cloze test, is a reliable and valid measure of English proficiency. It correlated highly with a standardized proficiency test,
the CELT test. In addition, native speakers scored significantly higher than ESL students in the highest level at Denton.

2. No single section of the test is as effective as the whole in measuring English proficiency.

3. Exact-word scoring ranks students as reliably as acceptable-word scoring.

4. Classes made up of students placed on the basis of proficiency testing are more homogenous than classes made up of students who have not been tested before placement.

Implications

With the proliferation of ESL programs throughout the country has come the necessity of structuring such programs. As the one-room school has given way to graded classrooms in the school system, so have ESL classes made up of students of varying proficiency been replaced by classes with targeted instructional levels.

A key ingredient to successful ESL program structuring is proficiency testing. In low-budget programs, where time and money are at a premium, a simplified proficiency measure is a necessity.

This research used an ex-post-facto design to find out if the instrument of this study, a three-section cloze test, was sensitive to differences in English proficiency to a significant degree. In this study no attempt was made to
place students or change class makeup on the basis of test results. However, because this research did prove the instrument to be a valid and reliable indicator of English proficiency, it can be used as a tool for level placement in the future.

Brookhaven Community College, North Lake Community College, and Mt. View Community College (all part of the Dallas Community College District) have requested permission to use the test in placing students for the Summer 1985 session. They will each provide information in the form of scores and comments by teachers as to how the test works for them.

A class in "How to Pass the TOEFL" is also a part of Brookhaven's curriculum. One difficulty in the past has been that some students have signed up for the class with a desire to get a high enough score on the TOEFL to enter college, but they have lacked sufficient English proficiency to benefit fully from the course. As a result, they have become discouraged and dropped out. A minimal entry test score would help alleviate this problem. A score of at least 30 on the cloze test is suggested as a requirement for this class. A student who scores lower than 30 would better profit from a regular ESL class.

The Counseling Department of the credit division at Brookhaven has also asked permission to use the test as a part of their assessment of ESL students. In addition,
students in the credit ESL class at Brookhaven took both the Michigan test and the cloze test at the end of the Spring 1985 semester and scores will be made available for correlation. An analysis of these scores will help to further establish the validity of the test.

Recommendations for Further Research

During the last decade there has been considerable research into the construction and scoring of cloze tests. The fact that cloze works as a measure of English proficiency has been well established. What each blank actually tests, however, has only recently become a focal point. Bachman (1982) has done research on the trait structure of cloze test scores to see what type of language knowledge is called into play in filling in a blank. The use of factor analysis to determine what is being tested by individual blanks is a fertile field for research.

As the three-section cloze test of this study is used more extensively, and additional information is received, further refinements might become necessary. If it proves to be an effective tool for level placement, alternate versions of the same type should be developed and validated to be used in conjunction with it.
APPENDIX A

THE INSTRUMENT

A THREE-SECTION CLOZE TEST
In the passage below some words are missing. You are to write the word in the blank that seems to fit best. Please write only one word in each blank. If you write more than one word in a blank it will be marked WRONG. It is a good idea to read the whole passage quickly first. Then go back and write in the easy words. Finally, read it again and write in the harder words. If you aren't sure of a word, GUESS.

Sample passage: (please print)

Joe is sick. He __________ a bad cold. He had to stay home __________ work today. He is __________ to see a __________. He needs to get __________ medicine.

Answers: has from going doctor some
New Shoes

Susan needed some new shoes. Her old shoes were worn
___________. So she ____________ to a shoe ___________.
She saw ____________ kinds of shoes. There ____________
shoes for men, shoes for ____________, and shoes for
children.

Susan found ____________ pair that she liked. She sat
__________ and took her ____________ off. Then she tried
__________ the new shoes. ____________ walked around the
room. ____________ shoes looked nice ____________ they
hurt her _____________. They were ____________ small.

She ____________ the clerk for a ____________ size.
After a ____________ minutes the clerk ____________ her
another pair. Susan put them on ____________ walked around the
room again. ____________ time the shoes felt good. They
__________ just right. She liked them ____________ much
that she kept ____________ on and wore them home.
A Newspaper

When you buy a newspaper you get a lot for your money. It tells you ___________ is happening in your ___________ and in your country. It ___________ tells you international ___________. But a newspaper gives you more ___________ news.

The articles in a ___________ give you information ___________ different subjects, ___________ as gardening, science, and travel. You can ___________ out about the weather ___________ scores of football ___________ and other sports. ___________ funny papers will ___________ you laugh.

A newspaper also helps you ___________ money. It tells you ___________ the sales are. The coupons ___________ the paper will save ___________ money at the store. The paper can ___________ you find a house ___________ buy or an ___________ to rent. It ___________ even help you ___________ a job. Because it helps you in ___________ many ways, a newspaper is a ___________ bargain!
Immigration and Emigration

For thousands of years, people have moved from one country or region to another. Their departure from one region is called emigration. Their entry ______ another region is called ______. This movement differs ______ migration, in which ______ large group of ______ of a whole community ______ from one place to ______.

Immigrants voluntarily move ______ themselves or with their families ______ settle permanently in another ______ or region. They usually travel from one continent to another and ______ in new areas ______ resemble their ______ environments. Most immigration ______ taken place ______ the 1800's.

People emigrate for many ______. They may seek opportunities ______ new jobs or new wealth. ______ may wish to escape ______ religious or political persecution. They may ______ like the government ______ living conditions in ______ homeland, or they may seek ______ freedom or happier surroundings in ______ to live.
APPENDIX B

CORRESPONDENCE, GUIDELINES,

AND PROPOSALS

FOR BROOKHAVEN COMMUNITY COLLEGE
October 24, 1984

Dr. Joy Babb, Chairman
Communications Division
Brookhaven Community College

Dear Dr. Babb:

I have been teaching an ESL class in the continuing education division at Brookhaven since last June. One issue that has come up repeatedly at ESL teachers' meetings as well as on the agenda of the Community College District ESL Task Force, is the need for a satisfactory language proficiency test to use in placement.

With this concern in mind, I did a research paper on methods of proficiency testing for one of my ESL Methods courses this summer. My research convinced me that the Cloze test is the tool best suited to Brookhaven's needs because it is easy to administer, can be completed in a relatively short time, is inexpensive, and can be tailored to discriminate at all levels. The disadvantages of the traditional standardized tests are summed up in the attached research proposal.

Now I am proposing development of a 3 level Cloze test for use in determining proficiency levels of ESL students. Development of this test will be part of my 6 unit thesis project for my Master's degree in ESL. My advisors are Dr. John Crow, founder of the Intensive English Language Institute at NTSU in Denton, and Dr. Dan Robertson, the current director of the institute. Dr. Robertson's area of expertise is educational testing, particularly ESL.

This test, when validated, should meet a real need at Brookhaven, and may meet similar needs at other colleges in the district as well. There is a standardized Cloze test already available but it only tests one level and has the disadvantage of being copyrighted.

I understand that I need approval to do a survey of this type involving students. I would appreciate your sharing this letter and the attached proposal with Russ Mauch, Vice President of Instruction. If you have any questions or if you would like to discuss the project with me, Sue Lichten can easily get in touch with me.

I hope this proposal meets with your approval and I look forward to hearing from you.

Sincerely,

Lauralee Lindholm
437 W. Brooklyn Ave.
Dallas, TX 75208
H. 942-8743
O. 688-2134
Given: The best learning situation is one in which students are of equal proficiency.

Problem: Lack of a tool to determine proficiency levels of ESL students to aid in class placement at Brookhaven Community College.

Standardized Tests: Eg. TOEFL, CELT, Michigan
Expensive
Time consuming
Need a trained administrator
Don't discriminate well on the lower levels - too difficult
Complexity of scoring

Cloze Tests
Short 125-150 word passage - every nth word left blank
Students must correctly fill in the blanks
Have proven to correlate highly with other standardized tests
Very inexpensive - 1 or 2 pages
Can be completed in about 30 minutes
Easy to administer
Scoring is simple
Passages can discriminate on targeted levels

Proposal:
Tool: A cloze test with 3 short passages at beginning, intermediate, and advanced levels
Sample: All ESL students at Brookhaven college in both continuing education and the regular credit program
Method: Class teachers will give the cloze test to all students at the beginning of the spring term. Two weeks later the test will be given again to 1/4 of the students to assure the stability of the test--that subjects get approximately the same score each time they take it. For maximum validity it would be good to give the test to all students at the beginning of the summer session as well.

The test will also be given to ESL students in a program where they have already been placed on levels as the result of performance on a standardized test (Students at the Intensive Language Institute at NTSU in Denton suggested).

Results: Scores of each class will be analyzed to determine the degree of homogeneity within the class--whether students are on the same level.
Two expectancy tables will be developed showing average score ranges of students on each section for each level.
A table would be made giving suggested score ranges on each section for each level of ESL student at Brookhaven.
This test could then be used as a placement tool for ESL students either at the time of registration or could aid the teacher in identifying students who were not on the right level.
I have now had an opportunity to research the issues involved in responding to your proposal to evaluate the Cloze test with our ESL students. I have also reviewed all of the correspondence that has transpired in the past few months regarding the proposal.

In order for me to render an informed decision regarding this proposal, I would be grateful if you and Lauralee Lindholm would respond to the following concerns:

1. Please identify the value that the College will receive from the implementation of the testing proposal.

2. Please identify how the results will be used off-campus.

3. Please identify the amount of time that will be taken away from the instructional process in order for the tests to be administered.

4. Please identify whether or not the ESL student will have the option of participating in the testing process.

In order to expedite the resolution of this issue, I would recommend that you place your responses to these four questions as an addendum to Ms. Lindholm's initial Research Proposal. If you would then send me a copy of the proposal with the addendum attached, I will get a decision back to you without delay.

Thank you for bringing this opportunity to my attention and for your help in clarifying the issues that I have raised.
1. It is the intent of this study to devise and norm a test to aid in ESL placement, at Brookhaven College in particular. Brookhaven was chosen because I have been teaching there and I have seen the need for a placement measure. I understand that Continuing Education is on a limited budget. Most of the standardized tests on the market today are expensive and require a trained administrator and grader. They are also very time consuming. This test requires a minimum of materials and time. Because it is not patented, there would be no charge for using it. Teachers would be allowed to see the students' tests and this would aid in identifying problem areas that need to be covered in class.

2. The test can be given in any setting - either on-campus or off-campus, and either in class or individually. Any teacher or clerical worker can give the test, with the aid of the instructions.

3. I realize that it is important to protect student class time. The test takes only 30 minutes to administer. It is possible that the lower levels would only need to take the beginning section of the test. Once the test has been normed, it would be possible to give it to selected students at the time they apply. This would help in placement and would not take away from class time. Borderline students who are already enrolled could be given the test to help the teacher in making a decision. This could be done in or out of class. Students who successfully complete instruction on a level would not need to take the test to pass to the next level unless the teacher felt a need for it. Sometimes students are placed at too low a level. If the teacher suspects this, the test would help in a determination. I believe that this test takes the least possible time of both teacher and student while still giving valid results.

4. In order to norm the test, it is necessary for all of the students at each level to take the test. To aid in this process, students at the Intensive Language Institute at North Texas State University in Denton are also taking the test. They have been placed on levels after a series of standardized tests. The averages of standardized test scores at each level will be provided for correlation with the cloze test scores. A result of this will be the development of a scale correlating cloze test scores with standardized test scores, such as the Michigan test and CELT. Once the test is normed it will not be necessary for every student to take the test. Participation would be optional, unless the teacher feels the need to evaluate the student, or wants to test the class as a whole. Also if there is question about level placement upon enrollment it would be required. I would like to add that an oral evaluation would be helpful as well.
APPENDIX C

CORRESPONDENCE, GUIDELINES, AND PROPOSALS

FOR INTENSIVE ENGLISH LANGUAGE INSTITUTE

AT DENTON
IELI Guidelines for Research Proposals

1. What is the proposed research topic?
2. What research techniques or instruments will be used?
3. How many students are needed? How much time per student?
4. What directions for administration will be given to cooperating teachers?
5. What feedback will be given to teachers and students? How will this benefit them?
6. What procedures will be followed in order to conform to the following ethical guidelines?
   a. Informed consent. All research should proceed only with the uncoerced, informed consent of the subjects. An exact description of the research project should be given. Subjects should know that they can withdraw at any time.
   b. Deception. Deception is defined as the intentional misleading of subjects to believe that the procedures and purposes of a research project are not what they actually are. Deception should be avoided.
   c. Consequences. It is the researcher's responsibility to minimize all risks to the subjects such as anxiety, lower self-esteem, etc. No risk should be taken unless the resulting benefits warrant it. One way to provide benefit to the subjects is to reveal the results of the study to the subjects for pedagogical purposes.
   d. Privacy. Great care should be taken to preserve the subjects' sense of privacy; they should not be required to reveal any more about themselves or their lives than they wish to. As the concept of privacy varies from culture to culture, the subjects' viewpoint on this question should predominate over the researcher's.
   e. Confidentiality/Anonymity. The subjects should clearly understand that the data collected will be kept confidential, both in its immediate use and in its long-term storage. Numbers, letters, or pseudonyms may be used for this purpose.
   f. Captive Populations. If there is a power differential between researcher and subjects which could operate to the disadvantage of the subjects, the subjects must be considered a captive population. The researcher should be sensitive to cultural differences. The subjects of the research are often newly-arrived. Especially studies which ask subjects to express their personal opinions may be threatening. Further, researchers may need to be cautious about doing studies in the classroom during class time. Students have paid for instruction, not to be subjects in a study; further, they may participate unwillingly out of fear of grade retaliation.

Please complete your proposal with explicit reference to the foregoing guidelines, and submit the proposal with any attachments (instrument, form, etc.) to the Steering Committee for review.
Research Proposal
By Lauralee Lindholm

1. The proposed research topic is the development of an English proficiency test to use in placing ESL students at the proper level of instruction. Current tests (i.e., Michigan, CELT) are time consuming, complicated to administer and score, and expensive. Because of their complexity, they are not very useful in assessing proficiency at the lower levels.

In the low budget continuing education program of the Dallas Community College District there is no place for this type of standardized test. As a result, placement of students has been done almost entirely by means of off-the-cuff assessments by secretaries or teachers. It is possible to transfer a student after classes have begun, but it is difficult because classes meet at various times and locations off campus. There is a strong feeling at this time by the Community College District ESL task force that some type of proficiency test needs to be developed to meet their needs.

2. I have devised a three-level cloze test to test English proficiency. Each passage is 120-150 words long and contains 24 blanks where words have been omitted. Rational deletion has been used with exact word scoring. A time limit of 30 minutes is allowed and nearly all students will finish easily within that time.

3. All Continuing Education ESL students at Brookhaven Community College will be tested at the beginning of the spring 1985 semester. Because these students have not taken any other type of standardized test, it would help very much to also test a student population that has been placed on various levels as a result of proficiency tests.

The goal of the study is to norm the test with suggested score ranges on each part of the test for each level of student. I would like to standardize this test by testing students at the Intensive Language Institute who have taken other standardized tests and then comparing scores.

4. A sheet of instructions for teachers has been prepared. On a dry run of the test with a Brookhaven credit ESL class in December, these instructions proved to be adequate and there were no problems experienced in administering the test.

5. Teachers will be given the final results of the study—along with suggested score ranges for each level. They will also be given the scores of the students. Scores can be given to the students at the discretion of the Institute. The students cannot see their tests as they will be used again.

Goals of the study are to provide a means of testing proficiency and placement. Students can be acquainted with these goals and I think they will appreciate the idea of the development of a rather easy and short test to accomplish these goals.
6. No personal information is asked of any student, other than his name and level. It is the group we are looking at, rather than individual students. The only reason for including students' names is to help the teacher identify problem areas or to give the student his score. I do not have plans for it at this time, but the test could be given again at a later date to measure achievement of students.

Because the test is relatively easy and plenty of time is allowed, I don't believe students will feel threatened or under pressure. I think they understand and agree with the need for determining proficiency levels in relation to instruction.

Teachers could explain the purpose of the study in a general way in their own words or I could prepare a short statement which they could read to the class, prior to their taking the test. It is fine to explain to students that participation is voluntary and in no way affects their performance at the Institute.
APPENDIX D

INSTRUCTIONS ON GIVING TEST
Instructions on Giving Test

Give students a folded test paper and ask them to write their names and levels on the front. Read the instructions to them and be sure that they understand. Especially emphasize that they are to write only one word in each blank. Two words or more will be marked wrong. They are not to change punctuation or words in the passages. They are to fill in the blanks with the sentences exactly as they are.

Ask students to complete the sample passage. When they have finished, read it aloud and ask students to supply the correct words.

Tell them that there are three passages like the one they just did. The passages are longer and the answers are not given at the end. When they finish one passage they should go on to the next. Point out that one passage is on the back of the sheet. Ask if there are any further questions.

Pencil is preferable as they can erase answers if need be. If they use pen they can cross out an answer and write the correct answer next to it. If there isn't room on the line they can write the answer above it.

Allow 30 minutes for the test. Tell them when only 5 minutes remain. Most students will finish easily before the time limit.
APPENDIX E

NOTES TO DR. JOHN CROW ON

REVISIONS IN PRELIMINARY TEST
IMMIGRATION AND EMIGRATION

Range

<table>
<thead>
<tr>
<th># correct</th>
<th># of students</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
</tr>
</tbody>
</table>

Comments: "Usually" was a problem for everyone and gave them difficulty with the rest of the sentence as a result. I am supplying it and reducing the test to 24 blanks.

I allowed "which" for "that" and "more" for "greater."
A NEWSPAPER

Range

<table>
<thead>
<tr>
<th># correct</th>
<th># of students</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
</tr>
</tbody>
</table>

Comments: Jean Conway pointed out that several students had trouble with the word "comics." I had debated whether to use the word "funny papers" or not. Because it is a medium-level test, I changed the word. At least they can analyze it and figure out the meaning.

The last sentence was a big problem for everyone. I reduced the test to 24 blanks and supplied "because."

I included the words "city" for "town" and "big" for "real" as underlined on the accompanying sheet.

What do you think about "what" for "where?"
### NEW SHOES

<table>
<thead>
<tr>
<th>Range</th>
<th># correct</th>
<th># of students</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>2</td>
</tr>
</tbody>
</table>

Comments: Because "tried on" was given in paragraph 2, I can't really disallow it for "put" in paragraph 3. So I chose to change blanks. Also all of the students got "around" right in the sentence with "put" because it was a repeat of an above phrase. So I feel it isn't really a good item for inclusion.

The last sentence was a problem for everyone. I have supplied the word "wore" and I think that takes care of the problem.

I added a blank for "larger" and will also allow "bigger" as an answer. The reason I hadn't included it earlier was because I felt both answers were equally good and with exact word scoring I had to disallow one of them.

I allowed "were" in place of "fit."
APPENDIX F

SCORING KEY
New Shoes

Out went store

Many were women

A down on The feet She but too

Asked few larger, bigger brought and

This fit, were them so
A Newspaper

what town, city
also than

news

about newspaper such find games make

and

The

save in

where, what you
to can, may

help apartment find

real, big
Immigration and Emigration

immigration into another country by moves and people settle that, which

old has since reasons

for not their greater, more

or which
BIBLIOGRAPHY


