1996

Development of parallel cloze tests using cohesion

Steven L. Jenkins
Iowa State University

Development of parallel cloze tests using cohesion

by

Steven Louis Jenkins

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF ARTS

Major: English(Teaching English as a Second Language/Applied Linguistics)

Major Professor: Dan Douglas

Iowa State University

Ames, Iowa

1996

Graduate College
Iowa State University

This is to certify that the Master's thesis of

Steven Louis Jenkins

has met the thesis requirements of Iowa State University

Signature redacted for privacy

TABLE OF CONTENTS

1 OVERVIEW

My thesis is simple: cohesion, the interdependence of textual elements, can be used to build reliable cloze tests. Furthermore, reliable cloze tests built using cohesion are parallel. This paper seeks to explore those hypotheses, and point out why the issues surrounding them are important.

Developing good language tests is difficult. Tests must be both reliable and valid for their interpretations and uses. To be reliable, the potential sources of error in the test itself must be minimized [6]. Reliability is affected by not only the examinee's language ability, but also by test methods, attributes of the test taker not related to the language abilities we want to measure, and random factors [6]. Ideally, only the language abilities that we want to measure are reflected by the test scores, but in practice, this is not the case. Unfamiliarity with a test method, or dislike of a test method, affects the examinee's performance. A test on listening, for example, should not depend on a student's ability to comprehend complex written instructions. Finally, random factors such as lighting, what a test taker had for breakfast, or the temperature in the testing environment can have an effect on scores [6]. In order for a test to be reliable, these influences must be minimized.

For a test to be valid, there must be evidence that the inferences made from the scores are valid. The uses of the tests must be made with validity in mind: the tests must be good measures for the purposes for which we are going to use them. For example, a score on the *Test of English as a Foreign Language* (TOEFL) should not be used in placement decisions for elementary school children.

Also, developing good parallel language tests is even harder than making tests reliable and using the results in valid ways. With parallel tests, not only must the tests be reliable, they must also exhibit a high degree of correlation. This means that measurement of the tests is important. It is not enough to have tests produced by experienced test creators, or have rough guidelines for tests, but there needs to be actual evidence that the tests are parallel. Providing this evidence can be very expensive, thus, a method for test creation that can be shown to produce parallel tests would be very useful.

As universities enroll growing numbers of English as a Second Language (ESL) students, they must be able to accurately and inexpensively measure the English abilities of these students. While many

universities use TOEFL scores in admissions, the time lag between most students taking the tests and arriving at universities is such that another measure needs to be made locally to determine the abilities of new students. Students who have not taken English for quite some time may have problems functioning at an acceptable level in the English speaking classrooms, or they may find that their curriculum was heavily influenced by the TOEFL test, and that their skill set is not quite what is needed by a university. Thus, many universities administer English placement tests (EPTs). These are sometimes haphazard, or possibly dated examinations of varying, even unknown, reliability and validity. In order for these EPT scores to be valid, they must be reliable; however, producing reliable parallel tests is time–consuming and expensive, usually requiring a cycle of test creation, reliability testing, and re-writing. It is hoped, therefore, that the results of this work can be used to alleviate the costs of producing parallel placement tests.

This paper is an attempt to help lay out a framework for developing reliable ESL tests using cloze tests created by examining the cohesion [25] found in texts. The utility of cloze tests of varying types will be discussed, as will cohesion. The paper will also present the results of actually implementing these ideas and testing them on a body of students at Iowa State University.

Cloze tests

"Cloze" tests are created by taking a text, deleting items in the text, and then requiring the takers to insert the deleted items back into the text. There are numerous ways to determine the items to be deleted (for example, random sample, fixed-ratio, and *rational* deletions), and two basic ways to score: exact match and acceptable match. Determining the acceptable match can also be done in several ways.

While cloze tests are not perfect, they can be used to test "knowledge of vocabulary, morphology, syntax and phonology/graphology" [6]. This breadth of language attributes can be important when considering the validity of using test scores for language placement purposes: since a single test measures a wide variety of skills, fewer special-purpose tests can be required. Another justification for using cloze tests is that there is a large body of work on them, which can be drawn upon when examining the construct validity of the test method. Also, cloze tests are simple to create. It is fairly simple to write a computer program that would create a fixed-ratio cloze test, allow an examinee to take the test, then grade the exam using exact match scoring. Thus, producing and administering cloze tests could be done automatically and inexpensively, which is of great benefit when dealing with large numbers of tests and test takers. While these justifications do not really respond to the critics of cloze tests, it is hoped that the reliability found in the actual tests will contribute to the acceptance of the scores for these tests as

valid for placement purposes.

Cohesion

"Cohesion occurs where the INTERPRETATION of some element in the discourse is dependent on that of another"[25]. By using the work of Halliday and Hasan as a basis for measuring cohesion in a text, it is hoped that parallel cloze tests can be prepared that have greater reliability than those made without taking cohesion into account. The greater reliability, as noted above, can help in acceptance of the scores for validity purposes, but it cannot answer all the problems that the cloze test has.

While Halliday and Hasan demonstrate several types of cohesion (see table 1.1) [25], for the purpose of this work, only lexical repetition and conjunction are used, with lexical repetition being the primary focus. Lexical repetition has been chosen for practical reasons: it is easy to measure automatically (i.e. with a computer program). Other types of cohesion are more difficult to measure. For example, determining pronoun reference is beyond the capabilities of a simple computer program. As is clear from table 1.1, lexical repetition is only a small part of cohesion. Other types of cohesion could have been used, but their usage is beyond the scope of this work as they would require more sophisticated measuring tools. Also, by using only one cohesive device to create the cloze tests, other sources of measurement error (or unreliability) can be factored out, leaving a fairly noise-free measuring tool.

Table 1.1 Types of Cohesion

Type	Examples
Reference	pronomials, demonstratives, definite articles, and comparatives
Substitution	nominal, verbal, and clausal
Ellipsis	nominal, verbal, and clausal
Conjunction	additive, adversative, causal, temporal, correlative, summary
Lexical	same item, synonym, superordinate, collocation

Overview of procedures

The procedure followed in this project was as follows.

- creation of computer program to analyze lexical repetition in texts
- examination of texts using this program, to determine suitability
- creation of tests

- administration of tests
- grading
- analysis of results

Rest of paper

In the second chapter of this paper, the theoretical frameworks for cloze testing and cohesion are laid out, and in the third, the actual method used to create the language tests is examined. The fourth chapter presents the results of the experiment while the fifth concludes the paper and points out some areas for future directions. The first two parts of the appendix contain the two cloze tests used in this study, and the last part contains the Scheme code of the program used to produce frequencies of words in the texts.

2 THEORETICAL FRAMEWORKS

Cloze background

The cloze test was developed by W. L. Taylor [66] to measure readability of texts, not measure overall language competency. The ESL community began using it to a small extent before Oller [44], but he is widely acknowledged as the foremost proponent of using cloze to test ESL proficiency.

In [46], Oller claims that cloze tests measure the underlying competence of language, not only performance. Others, however, have disputed this claim [1]. Also, varying methods of test scoring and deletion method have produced slightly differing results. This chapter will attempt to explore four aspects of cloze tests: deletion method, scoring method, correlations with other tests, and underlying skills that cloze tests measure.

Oller recommends that cloze tests have a text length of between 250 and 500 words, and that deletions be done by deleting every nth word (where n typically is between 5 and 10 words). He found that using a more flexible scoring method than requiring an exact match of the deleted word produced better correlations with other measures of English language proficiency for non-native speakers. He also found that the nth word deletion method (fixed-ratio) produces a more general measure, although Chapelle and Abraham [12] contradict this. Also, in [46] and [34], Oller (and others) found that cloze tests correlate most highly with dictations, but that they correlate well with many other tests as well.

Other researchers have questioned the usefulness of cloze tests. Alderson [1] has examined cloze tests to determine the effects of deletion rate, text, and scoring procedure in fixed-ratio cloze tests. He found that there is some interaction between text and deletion rate, but that no exact formula exists. He also found that the scoring method that correlated most highly with the baseline test is the semantic equivalent, not the exact match. Another interesting conclusion he draws is that the cloze only tests items in the immediate environment, not at the clausal level; however, he excludes such devices as lexical repetition, anaphora, and conjunctions.

Bachman [4] attempts to understand what cloze tests measure. He found that "support is found for the claim that cloze tests can be used to measure higher order skills–cohesion and coherence–if a

rational deletion procedure is followed." It should be noted that in his work, semantically acceptable scoring was used.

Deletion method

There are two primary methods for determining items to delete: fixed ratio, and the rational deletion method. The fixed ratio, as noted above, keeps a constant number of words between deletions. The rational deletion method, though, requires that the test creator provide some reason for deleting an item. Bachman [5] attempted to study the differences among tests created with fixed-ratio and rational deletions. He also attempted to come develop criteria for classifying items in a cohesive framework, as doing so has been "both difficult and subjective." He raises five questions that are quite useful:

1. Can obvious and practical criteria which are theoretically justified be developed for selecting deletions?
2. Are there any differences in the proportions of different item types deleted by fixed-ratio and rational deletion procedures?
3. Are there any differences in performance on cloze tests constructed by fixed-ratio and rational deletion procedures?
4. Are there any differences in performance on different types of items?
5. Can a rational deletion procedure yield a "better" test of specified components of language proficiency than a fixed-ratio deletion procedure?

The results from that research show acceptable reliabilities (.608 to .862), and quite high correlations (.620 to .848) to several other measures of ESL. He found that by using the "hierarchical structure of written discourse as a criterion," practical criteria for deletion could be found. He also found that a rational cloze can be constructed so that the test developer can control what items the test is measuring, important for content validity.

Chapelle and Abraham [12] found that the actual cloze method used does not have a significant impact on reliability. In particular, they found that the rational cloze tests correspond strongly with other tests, while fixed–ratio correspond most poorly. The impact for automatically generated cloze tests is obvious: the simplest method (fixed–ratio) is not the best choice.

Scoring method

The methods used to score cloze test vary from exact match (only allowing the exact item) to syntactic equivalent (allowing any word of acceptable part of speech). In the middle lies perhaps the most interesting: semantic equivalent. In semantic equivalent scoring, acceptability is determined by native speakers judging the replacement to be valid or invalid. Many researchers (including Alderson [1]) have found semantic equivalent grading to be the best method for scoring non-native speakers.

Correlations

Hanania and Shikhana [26] studied the relationships among a cloze test, standardized ESL test, and written composition test. Their goal was to determine if a cloze test could replace a written composition test for placement purposes. They found that the cloze tests had very high reliability (KR-21 coefficients of .92 to .98) and high correlations to the local placement test (which was determined to have a high correlation with several standardized ESL exams). Their cloze tests were fixed-ratio (but made according to rational decisions), and scored with exact scoring. The correlation to the written composition was .68, and it was found that either two of the three tests could be used to predict the results on the third.

Oller, of course, has found that cloze tests correlate well with many other tests [44]. Alderson, however, found correlations varying wildly: between .25 and .91. However, he does not present reliability measures for the test he uses as a baseline, thus throwing some question about the utility of his results. Other researchers ([12],[4],[5]) have found that cloze tests, in general, correlate quite well with other measures of language proficiency.

What cloze tests measure

While Oller has made the claim that fixed-ratio, exact-match cloze tests are good measures of general language proficiency, many researchers (including this one) are hesitant to accept a single such test as a valid measure for placement purposes. Alderson[1] in particular believes that such cloze tests are only testing low-level features, not more complex features such as cohesion. Bachman[4][5] asserts that rational cloze tests can actually test specific features. His evidence is quite convincing, both in terms of theoretical model and empirical evidence. His framework for examining test questions is very useful for test developers, and implies that cloze tests measure whatever the test creator designs them to measure.

Some have examined field dependence-independence and cloze tests. Among those are Stansfield & Hansen [61] and Chapelle [13]. They found that field independence is an important variable in cloze tests, which corroborates the assertions that the cloze is a measure of more general language proficiency and not low-level features. However, since field independence-dependence is not a linguistic feature, but rather a more general cognitive ability, its impact on cloze test results is somewhat two-edged. If cloze tests are actually testing cognitive strategies rather than linguistic features only, then cognitive strategies either need to be included in the model of communicative competence used, or the effects of cognitive strategies need to be taken into consideration.

Cloze conclusion

In conclusion, cloze tests seem to possess the necessary reliability and content validity for at least assisting in making placement decisions. Cloze test construction and grading seem to be simple enough to do reliably and cheaply. However, there are important implications for automatically creating and scoring cloze tests, namely that the simplest methods (fixed ratio deletions and exact match scoring) do not produce the highest correlations with other tests. Thus, it is expected that this research will produce lower correlations than are possible with using rational deletions and semantic equivalent scoring.

CALT

We conclude this chapter with a brief discussion of Computer Aided Language Testing (CALT). The utility of CALT in this research is quite limited, as a computer is only used to produce frequency tables for words and sentences in text. However, the possibilities for more extensive CALT are obvious: if a computer program could

1. find texts
2. eliminate texts that will not produce good cloze tests
3. produce cloze tests from the remaining texts
4. administer the tests and
5. grade the tests

then the cost of testing will decrease greatly. However, making the jump from current research in CALT to the above scenario will require a lot of work. In particular, producing cloze tests and grading the

tests will be difficult. The second step, elimination of unacceptable texts, is only slightly easier because a test creator could specify the range of acceptable features that are required. For example, a good text should consist of at least 500 words, have at least 70 words that occur with a frequency greater than 3, and that some percentage of the 70 words should occur at a certain level in a specific corpus (e. g., Brown's corpus). The difficult part here is in determining the acceptable features for a rational cloze. Determining (automatically) the higher level linguistic features in a text is still an unsolved problem [41].

The other two steps are even more difficult. Producing cloze tests relies on determining the linguistic relations of words in the text, and grading the texts is problematic also. Exact match grading, of course, is trivial, but a better model with a rational cloze is to use semantic equivalents, which, of course, requires some degree of Natural Language Understanding (NLU) in a computer program. There might be some success to be found in using *schema* to determine the acceptability of texts, and then relying on the schema to help determine the semantic roles in the text. This approach, however, still requires human intervention, as Morris and Hirst point out[41]. Thus, CALT has yet to become a reality for automatic generation, administration, and grading of cloze tests.

CALT does offer the technology to administer adaptive tests, though. If a battery of tests were developed, ranked according to difficulty, then the tests could be administered to students in a careful fashion. Unlike "traditional" Computer Adaptive Tests (CATs) that select individual questions from a database based on the test taker's performance, an adaptive test using cloze tests would select the appropriate cloze test. As Laurier mentions, though, CAT is not a panacea[39]. Producing a battery of tests would be very expensive, and the psychometrics of a computer administered test do not match up with most "real world" language situations, although in a university setting, understanding and producing English on a computer is often one way non-native speakers interact with university officials (e. g., English composition instructors).

3 TEST METHOD

Test creation

The tests used in this study were created by selecting random texts from freshman level composition texts. This source was chosen, as students entering a university are expected to be reading at that level. The texts selected were then run through a program (written in Scheme, see appendix B) written by the author to chart out the frequency and occurrence of words in the text. The program produces a table with columns representing each sentence, and rows for each word in the text. A piece of sample output is shown in table 3.1. A text was judged adequate if it had enough lexical repetition that 35–50 items could be deleted in the text, keeping at least 4 words between deletions with each item to be deleted occurring at least three times in the text. An attempt was made to ensure that the distribution of deleted words was even (e. g. , that the word "and" was not deleted more often than any other word). The initial plan was to delete only *content* words in the text, or non-function words, but it was soon seen that texts of intermediate length (200-500 words) did not provide enough lexical repetition of content words to produce a cloze test of reasonable (with more than twenty deletions) length.

It should be noted that these criteria are easily met by many texts. However, some texts in particular did not meet them. The shorter writings, predictably, did not contain enough words to be used. Somewhat unexpectedly, though, some writings of more polished writers did not meet the requirements. In particular, E. B. White's writings did not contain enough lexical repetition. While this might not be surprising to the community that studies cohesion [59], it is useful to the language teaching community as a whole. That the more polished writers use different cohesive devices can have some interesting implications: that is, writers at different levels use different kinds of cohesion. In freshman composition courses, often the emphasis is placed on transitional elements, which are primarily the conjunctions of Halliday and Hasan. In lower level courses, students study ellipsis and types of reference (often under the rubrics of subject–verb agreement or pronoun agreement). If the types of cohesion were to be studied more closely in connection with language learning, a better understanding of what cloze tests measure might be found. In addition, if very complex models of cohesion were used to develop tests, a

beneficial backwash might be created, encouraging educators to incorporate the ideas of Halliday and Hasan, and research on discourse in general, in their classes.

The texts chosen were *If I'm so Smart, How Come I Flunk All the Time?* [60], and *The Background of Experience* [67] from a freshman composition text([23]). The full texts of both of these are included in the Appendix. Some samples from the output of running the first text through the computer program that produced some simple counts of frequencies per sentence is produced in table 3.1. The table is arranged by total frequency, token, and then number of occurrences per sentence. Each row shows the data for the word shown in the first column. The columns labeled 1 through 6 show the frequencies of that word in each of the first six sentences. Only the data for the first six sentences is shown. The predominant source of repetition in both texts is by the function words, or the non-content words– words that do not directly contribute to the content of the essay. It was originally hoped that the function words could be filtered out, and only content words used as items for deletion, but the lack of lexical repetition within the texts prevented this. In the first text, there were only 25 words that met the criteria necessary for deletion (i.e., occurred more than 3 times). This is obviously a problem when trying to construct cloze tests with 50 deletions. The solution chosen by the author was to delete content words then function words as necessary to increase the length to a more appropriate size.

The second text had a slightly different distribution but the same essential problem, as can be seen in table 3.2. In addition, it was a much shorter text, only 576 words long versus 1076[1]. The cloze test made from this text had only 37 deletions. The inclusion of this text in the study is evidence of the difficulty in finding appropriate texts. In future work, greater care would be taken to ensure that the texts chosen were of more similar size. It would be possible to create our own texts, but since one of the goals is to automatically create cloze tests, having to create the text would be undesirable.

Overall, a different framework for design might be used in future works to at least select the items for deletion based on more information, i.e., produce a more rational cloze test. Also, more careful selection of texts would be necessary, a problem for all cloze tests. This work, at least, explores using a more careful and precise measure in determining what texts could make good tests.

Student selection

Students were selected from the pool that had taken the English Placement Test (EPT) here at Iowa State University. This population consists of all international students whose native language is

[1] These counts were obtained using the computer program `wc`.

Table 3.1 Word Counts from Text1

Total	Word	Sentence 1	Sentence 2	Sentence 3	Sentence 4	Sentence 5	Sentence 6
67	the	1	0	2	1	1	2
36	to	0	0	1	1	0	0
24	and	0	1	1	0	0	0
11	students	0	1	1	0	0	1
7	teachers	0	0	0	0	0	0
7	scientists	0	0	1	0	0	1
7	homme	0	0	1	0	0	0
7	grades	0	0	0	0	0	0

Table 3.2 Word Counts from Text2

Total	Word	Sentence 1	Sentence 2	Sentence 3	Sentence 4	Sentence 5	Sentence 6
33	the	5	2	1	0	3	3
26	of	3	0	3	2	1	2
18	we	0	0	1	3	2	0
7	rule	1	0	2	1	0	1
7	blue	0	0	0	0	0	2
5	see	0	0	0	0	0	1
5	background	1	0	1	0	0	0
4	scene	0	0	0	0	0	0

not English[2]. At the beginning of their first semester at Iowa State, these students are expected to take the EPT to determine what, if any, ESL courses they will need to take. The remedial courses available include a low level composition class, a higher level class for undergraduates, a higher level class for graduate students, a remedial reading class, and a remedial listening class.

The students selected were drawn from those currently in English 101B, 101C, 101D, and 104 (Freshman Composition). The students from 104 were chosen as they were students who had shown a sufficient level of proficiency that they did not need to take additional ESL courses.

Test administration

The tests were given during students' regularly scheduled classes, and participation was optional. Students were assured that participation would have no effect on their grade in the course. A total of 95 students elected to participate. They were given 50 minutes to complete both sections of the test. It should be noted that almost all finished before the allotted time was over.

[2]The group of international students does not include those students from Puerto Rico, as it is officially a territory of the United States.

Grading

Exact scoring was chosen, as it is the type of grading most amenable to automatic scoring. A blank and an incorrect answer both received a mark of zero, while a correct answer received a mark of one. While it may be preferable to have scored differently, exact scoring is, of course, the simplest. For more sophisticated scoring, some type of Natural Language Processing (NLP) system would need to be used, as a simple thesaurus–type checker would most likely prove unacceptable (see [41]). A system like that, though, would involve a great deal more research. Another interesting idea would be to use the idea of clozentropy to produce grading criteria for the exams. This would be more expensive in terms of test creation time than *pure* NLP, but would probably give a greater degree of confidence in the scores (reliability), and hence, the validity.

Test analyses

Analyses were done as follows: EPT test 1, EPT test 2, test 1, and test 2. The first EPT test is a listening test, and the second a reading. As not all students had scores for the EPT tests (since the EPT scores of those students who took the EPT test during previous semesters were not included), further breakdowns were made. Scores for the EPT and cloze tests were used to measure the correlations, and the scores for those students who did not take the EPT during the current semester were only used to measure the reliability of the cloze tests.

Various statistical measures for reliability and parallel tests were used to do the analyses of the tests, as can be seen in the next chapter.

4 TEST RESULTS

My thesis, that cohesion can be used to build reliable and parallel cloze tests, holds true for the research done for this work, to a certain extent. Both tests created using cohesion show strong reliability, but the correlations are not very good; in fact, one of the correlations is actually negative. These results imply that better tests need to be created but that there is potential in using cohesion (and cloze) to build reliable parallel tests.

Basic results

According to Gulliken[24], parallel tests have equal means, variances, and intercorrelations. Table 4.1 examines some of the necessary statistics. Note that the maximum score for the first test is 50, and the maximum for the second is 37.

Table 4.1 Simple Statistics for the Cloze Tests

Test	Mean	Standard Deviation	Minimum	Maximum
Test 1 ($N = 25$)	16.32	9.99	0	26
Test 2 ($N = 25$)	14.72	5.57	1	21
Test 1 ($N = 94$)	17.19	6.93	0	32
Test 2 ($N = 94$)	14.14	5.75	0	26

The ($N = 25$) group is the subset of the ($N = 94$) group that also has EPT scores. The scores for both sets of students are included to show that the subset is a good sample of the whole on both tests. In other words, the correlation coefficients calculated later should also apply for the larger sample, even though coefficients were only calculated for the smaller group. The means and standard deviations for both are similar for both groups, which is good. However, the low minimums and maximums are not good: they imply that the tests were too difficult for students. Since the EPT is to be taken by students who have already been accepted by the university and achieved fairly high TOEFL scores, it should be the case that their scores should be similar to that of native speakers (i.e., near-perfect). This is obviously not what the results show; however, this problem can at least be partially understood by the

grading policy of only accepting exact matchings. A more lenient policy (e. g., allowing synonyms or misspellings) would have resulted in higher scores.

In order to see exactly which questions were too difficult or too easy, the next section will examine each question in detail.

Item analysis

Test 1

Table 4.2 shows the p-values[1] for test 1. Each row in the table shows p-values for five questions, and the first column serves as a counter for which question is in the second column (e. g., the third row, third column shows the p-value for question 18). As is clear, several of the questions had no correct responder (i. e., a p-value of 0.0) , and many items (24) had p-values below the range usually considered acceptable ($.33 \leq p \leq .67$). This is not good, as almost half of the items on the test would be considered "too hard" for the intended audience. Also, there are a few items (4) with p-values above the range, which is not too bad, as there should be some questions on the test that almost all takers can answer. This section will analyze each item that has an out of range p-value to help determine what the question is testing, and to see if a different scoring method (e. g., semantic equivalent) would have perhaps resulted in more acceptable scores. It should be noted that it is to be expected that semantic equivalent scoring will help most on content words and that, overall, using more lenient scoring should raise the scores. However, doing semantic equivalent scoring requires some subjectivity on the part of the graders (or some level of sophistication in producing possible answers: e. g., clozentropy), which eliminates that option for this research.

It should be noted that there are 17 content words and 33 function words on this test, so the test is predominantly on function words. Based on previous research (as discussed in chapter two), we would expect this test to be well-correlated with other language tests, and that it primarily tests low-level linguistic skills. The text of the test can be seen in Appendix A.

The questions with low p-values are 4, 8, 9, 11, 13, 16, 19, 22, 23, 24, 27, 28, 32, 34, 35, 36, 38, 39, 42, 43, 45, 46, and 50. Those with high values are 12, 20, 25, and 40. As there are fewer of the latter, they will be examined first.

Question 12 (token *and*, p-value .85) is a function word that is part of a common expression: *and so on*, thus it is to be expected that students would find this question easy.

[1] A p-value is the probability that a test taker has correctly answered the question.

Table 4.2 p-values for test 1 ($N = 94$)

Item	p-value	p-value	p-value	p-value	p-value
1	.34	.43	.46	.29	.62
6	.62	.55	.17	.12	.36
11	.23	.85	.21	.53	.44
16	.31	.66	.40	.13	.78
21	.57	.01	.14	.16	.69
26	.72	.26	.04	.45	.46
31	.46	.09	.50	.04	.04
36	.17	.64	.00	.00	.73
41	.39	.12	.14	.44	.20
46	.04	.67	.55	.39	.00

Question 20 (*color*, .78) is a content word, but the cue *blue* is immediately after it, so again, it is not unusual that students would find this one easy.

Question 25 (*of*, .69) is a function word, and its p-value is only slightly high. Its function does not really depend on its lexical repetition, but rather on the students' understanding of prepositions.

Question 40 (*we*, .73) is a function word, but it does depend on the context for students to interpret. They must retain knowledge across sentences that the author is using the first person plural to serve as a subject. The (rough) semantic equivalent *you* is unacceptable because the following clause within the sentence makes it clear that the first person is being used; therefore, the students must do some non-linear reading (perhaps even going back and changing the answer) to find the correct response.

Thus, the easier items test primarily low-level linguistic features and do not measure higher-order abilities.

We will now examine the more difficult items in order of occurrence.

Question 4 (*the*) is a function word that has a p-value of .29, which is very close to the acceptable range. It has several acceptable semantic equivalents, including *any*. Consequently, it is probably a good item for our test, even though the score is slightly low.

Question 8 (*rule*) is a content word and has a p-value of .17. Several semantic equivalents include *idea* and *concept*. Allowing these as acceptable responses would have increased the frequency of correct responses greatly. It should be noted that this content word is testing a relatively low-level linguistic feature: vocabulary. As such, accepting a semantic equivalent would be very valid.

Question 9 (*blue*) is a content word testing higher level linguistic features and has a p-value of .12. The students must be able to maintain the topic from the previous two clauses (previous sentence) in

order to answer this correctly. Semantic equivalent grading would not have helped this question.

Question 11 (*of*, p-value .23) is a function word that tests prepositions, a low-level linguistic feature.

Question 13 (*formulate*, .21) is another content word, and again tests vocabulary, as several semantic equivalents exist (e. g., *create*, *develop*, *see*, or *make*).

Question 16 (*the*, .31) is a function word with a score very near the acceptable range. The article *an* would be a reasonable alternative, and allowing it would make this question fall within the acceptable range.

Question 19 (*do*, .13) is another function word with acceptable alternatives that would make this question have an acceptable p-value.

Questions 22, 23, 24, 27, 28, 32, 34, 35, 36, 38, 39, 42, 43, 45, 46, and 50 are summarized in table 4.3 and will not be discussed in detail. Note that *type* refers to either content word or function word, and *level* refers to the question testing either low-level or higher-order linguistic features.

Table 4.3 Deleted items for Test 1 (in order of deletion)

question	token	p-value	type	level	alternatives?
22	background	.01	content	high	yes
23	law	.14	content	both	yes
24	that	.16	function	low	yes (*which*)
28	our	.04	function	low	no
32	but	.09	function	low	no
34	or	.04	function	low	yes
35	it	.04	function	low	no
36	scene	.17	content	high	yes
38	but	.00	function	low	yes (*though*)
39	about	.00	function	low	yes
42	if	.12	function	low	yes
43	on	.14	function	low	yes
45	turn	.20	content	low	yes
46	scene	.04	content	low	yes
50	as	.00	function	low	no

Test 2

Table 4.4 shows the deleted items for the second test. The full text is available in Appendix A. Of the 37 deletions, 28 items are content words, and 9 are function words. There is also a minimum of repeated items deleted. Tables 4.5 and 4.6 show how the test takers fared on each question. The formats of these is the same as table 4.2, each row showing the p-value for five questions, with the first column

serving as a reminder of the contents on the second column. The results show that no questions were correctly answered by all takers, and that no questions were missed by everyone. While these points are secondary to the goals for the research, they do reinforce the conclusion that the test is too hard. As mentioned in table 4.1, the minimum is 1 and the maximum is 21. But a high score of 21 out of 37 is not really very good because it may imply that the intended audience is not the audience that the test would measure the most appropriately (i.e., with the most validity). Another very important issue is that the p-value of many questions is outside the customary acceptable range ($.33 \leq p \leq .67$)[52]. In fact, of the 37 questions, only 14 fit within that range (for $N = 25$), although a few others are within .05. For $N = 94$, it is even worse: only 10 meet this criteria. This is further evidence that the level of difficulty is not calibrated correctly for the test group.

Table 4.4 Deleted items for Test 2 (in order of deletion)

flunking	teachers	scientists	teachers	boys
and	Dr	they	or	weren't
to	students	the	instructor	their
bright-eyes	that	scaredy-cats	eyes	questions
the	class	school	grades	dummies
knowledge	training	bright-eyes	dummies	teacher
eyeballing	Homme	laboratory	classroom	students
hand-raising	work			

Table 4.5 p-values for test 2 ($N = 25$)

Item	p-value	p-value	p-value	p-value	p-value
1	0.64	0.04	0.04	0.55	0.68
6	0.79	0.64	0.36	0.96	0.21
11	0.55	0.21	0.85	0.00	0.47
16	0.47	0.51	0.85	0.79	0.66
21	0.09	0.13	0.09	0.21	0.43
26	0.00	0.19	0.09	0.00	0.32
31	0.17	0.55	0.23	0.43	0.04
36	0.34	0.43			

Table 4.6 p-values for test 2 ($N = 94$)

Item	p-value	p-value	p-value	p-value	p-value
1	0.23	0.11	0.78	0.02	0.05
6	0.61	0.67	0.83	0.62	0.31
11	0.94	0.27	0.57	0.29	0.86
16	0.01	0.47	0.45	0.54	0.85
21	0.74	0.68	0.07	0.13	0.09
26	0.18	0.43	0.01	0.27	0.05
31	0.05	0.40	0.24	0.53	0.24
36	0.49	0.05			

From figure 4.1, we can see the number of correct responses per question. In the next two subsections, we will analyze the responses according to deletion type, with the analysis of deleted function words followed by the analysis for deleted content words.

Function words

Questions 6, 8-11, 13, 15, 17, and 21 are all composed of deleted function words. These are summarized in table 4.7.The word *weren't* is a helping verb, and *to* is an infinitival. The others are conjunctions and references. The question with the lowest score is question 10, *weren't* with a frequency of 29. All of the others have a score at or above 40 (out of a possible 94). This implies that perhaps these deletions test low-level skills. If so, then a more appropriate cloze test might be constructed deleting only function words, instead of attempting to use content words to create some type of content validity. These results correspond with what Lado [38] and others have seen concerning performance on items that test function words.

Table 4.7 Test 2 function words

Word	Cohesive classification	Frequency
and	conjunction	57
they	reference	78
or	conjunction	58
weren't		29
to		88
the	reference	54
their	reference	81
that	reference	44
the	reference	70

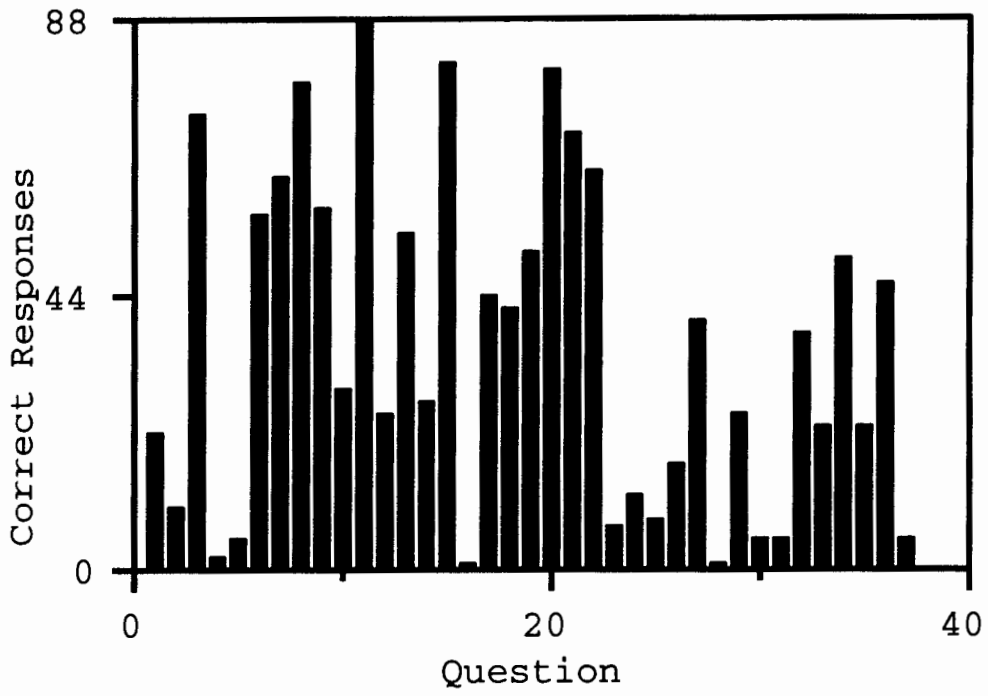

Figure 4.1 Item responses for test 2 ($N = 94$)

Content words

In this section, only those questions whose frequency of correct response is less than 40 are discussed (unless otherwise noted), as these are the most interesting.

The most difficult questions (frequency of one) were 16 and 28. Both required a response of *bright-eyes*, and only one student (not the same) answered each correctly. It should be noted, though, that numerous students provided semantically equivalent responses; e. g. , *students* or *they.*

The next most difficult question was the fourth deletion, *teachers*, with a frequency of two. It required readers to keep the relationships among *scientists*, *students*, and *teachers* clear in the passage. Clues were also available in the following clause; hence, it tests (slightly) the non-linear reading skills of the students. Even though the non-native speakers had poor performance, no native speaker among those that pre-tested the test missed this question. This is surprising, as others [44] have found that very proficient non-native speakers and native speakers have similar performance levels.

The next lowest frequency is seen with questions 5 (*teachers*), 30 (*teacher*), 31 (*eyeballing*), and 37 (*work*), each having 5 correct responses.

Question 23 (*school*, frequency 7) mainly tests cross-clause, but inter-sentence material. It also requires examinees to look ahead in the text, and thus tests non-linear reading. A possibly correct response (*laboratory*) is ruled out if the test taker looks ahead to the rest of the sentence.

Question 25 (*dummies*, frequency 8), required higher-level understanding of the textual content at first glance. However, given responses of *bright-eyes* and *scientists*, it is clear that this question is also testing low-level vocabulary: *biased against* provides a strong clue that the answer should be *dummies*. Even this clue, though, did not help most students.

The second question (*teachers*, 10) showed a clear problem: answers of *students*, *experiment*, or *scientists* were common, and although incorrect, they demonstrate an attempt at using context beyond the clause to solve the problem. This is similar to what Dollerup has seen with students on the Sprogtest [17] in that the students are guessing using incorrect cues.

Question 24 (*grades*) had a frequency of 12, much like the second question, and it tests inter-clausal performance. Had we used semantically equivalent scoring there could have been two possible answers for this question: *grades* and *performance*. Some other answers given are *quizzes* and *participation*. All of these could be acceptable using different scoring methods.

Question 26 (*knowledge*, 17) could also benefit from semantic equivalent scoring: *improvement* in particular would increase the frequency of correct responses. This item tests high-level language abilities and does not depend much on particular low-level features (as contrasted with question 25).

Question 1 (*flunking*) had a frequency of 22, but inexact scoring would have produced much higher results, with frequent answers of *failing*, *staying*, or *poor*.

Questions 33 (*laboratory*) and 35 (*students*) both had correct responses 23 times, and both have acceptable semantic equivalents that would have increased the scores.

Questions 12 (*students*) and 29 (*dummies*) both had a frequency of 25, which is a p-value of .27. Thus, they are just below the acceptable range for a well-calibrated test. Both of these have semantically equivalent forms that would have put them well within the acceptable range (e. g. , *students* and *boys* would be acceptable for number 12, and *students*, *boys*, or *dummies* would be acceptable for number 29).

Question 14 (*instructor*) had a frequency of 27, and an acceptable alternative would have been *teacher*, for example.

Question 32 (*Homme*) had a frequency of 38 (p-value of .40), but there were no semantically equivalent answers: some responses were *students* or *more*, which are syntactically equivalent but do not meet the semantics of the text. Thus, it is clear that this question does test higher level linguistic

features, and that semantically acceptable alternatives would not have changed the score much, but as the p-value is in the acceptable range, there is no need for altering this item.

The question with the highest frequency of correct response is item 18 (*scaredy-cats*), with a frequency of 42. This question, like question 32, tests the higher level skills of students (requiring test takers to recall the taxonomy of students that the researchers developed).

Item analysis conclusion

In conclusion, the content words, if graded according to more lenient criteria, would have exhibited a higher set of p-values, but the problem of automatically grading the responses prevents test-designers from using semantically equivalent scoring more frequently. In particular, the difficulties restricted this researcher from using it, even though it would have clearly improved the scores of students, perhaps making the tests correlate more highly. Another problem with determining what linguistic features items are measuring is that the categorization of questions is not always straightforward (as mentioned elsewhere and seen in this discussion). Thus, the possibilities of automating an inexact process are quite difficult.

Reliability

Table 4.8 shows that the Guttman split-half reliability[2] coefficients for the two tests are quite good. Also note that Test 2 had an odd number of questions, but I only counted the first 36. Both sets of tests, the complete set and the set with the correlating EPT scores, have very close reliabilities. These reliabilities are unsurprising, as many others using the cloze method have found high reliability.

Correlations

The more important statistic for this research is the correlation with other tests, in particular with the EPT. The means and standard deviations are also important in being able to verify that the method chosen builds parallel tests. As table 4.1 shows, the means, standard deviations, and variances are close but not equal. Computing the intercorrelations requires computing correlation coefficients for a third, presumed parallel, test. In our case, of course, we want to use the EPT as the baseline.

Table 4.9 shows the Pearson correlation coefficients and the $Prob > |R|$. The most positive result is that the two cloze tests (TEST1 and TEST2) have a positive correlation. Also, they both have a

[2] $Reliability = 2(1 - \frac{s^2_{h1} + s^2_{h2}}{s^2})$ where s^2_{h1} and s^2_{h2} are the variances of the two halves.

Table 4.8 Guttman Split-Half Calculations

Variable	Variance	Reliability
Test 1, first half ($N = 25$)	19.00	0.71
Test 1, second half ($N = 25$)	12.49	
Test 1, total ($N = 25$)	48.81	
Test 1, first half ($N = 94$)	18.00	0.68
Test 1, second half ($N = 94$)	13.77	
Test 1, total ($N = 94$)	48.0	
Test 2, first half ($N = 94$)	7.51	0.82
Test 2, second half ($N = 94$)	11.77	
Test 2, total ($N = 94$)	32.64	
Test 2, first half ($N = 25$)	9.03	0.79
Test 2, second half ($N = 25$)	9.11	
Test 2, total ($N = 25$)	30.07	

Table 4.9 Correlation Coefficients ($N = 25$)

	EPT1	EPT2	EPTSUM	TEST1	TEST2
EPT1	1.00000	0.19020	0.92147	−0.22368	0.30582
	0.0	0.3625	0.0001	0.2824	0.1371
EPT2	0.19020	1.00000	0.55662	0.04443	0.43084
	0.3625	0.0	0.0039	0.8330	0.0315
EPTSUM	0.92147	0.55662	1.00000	−0.17171	0.42926
	0.0001	0.0039	0.0	0.4118	0.0322
TEST1	−0.22368	0.04443	−0.17171	1.00000	0.41344
	0.2824	0.8330	0.4118	0.0	0.0399
TEST2	0.30582	0.43084	0.42926	0.41344	1.00000
	0.1371	0.0315	0.0322	0.0399	0.0

positive correlation with EPT2 (the reading test). This second result is also exciting, as it shows that the two cloze tests show similarity to each other and to an external, perhaps reliable, test. However, since no statistics for reliability exist for the EPT tests, using them as a baseline to measure the cloze tests is not completely convincing.

The one very negative result is with the listening portion of the EPT and the first cloze test. This is not very surprising, as a strong correlation with a listening test is not expected from written tests, especially those that deal with production, although there might be some discourse–level skills that would be similar (e. g. , prediction [44]). Table 4.10 shows some more detail that may be useful. Oller [46] notes that cloze tests have "consistently correlated best with measures of listening comprehension." His analysis is different from our results.

Table 4.10 Statistics for the EPT

Test	Mean	Standard Deviation	Variance
EPT1	31.8800	7.1199	50.6933
EPT2	22.4000	3.3291	11.0833

The data here imply that most students performed at approximately the same level on the reading portion. With a perfect score of 30, a mean of 22.4, and a variance of 11.0833, the test does not discriminate very highly among the students. However, since (presumably) all of the students had achieved satisfactory scores on the TOEFL prior to taking this exam, the statistics imply the current EPT does a reasonable job. It should also be noted that the cutoff for the scores is 30 for the listening (out of a possible 50), and 18 for the reading (from a possible 30), both 60% of the maximum. Students that score below those percentages are required to take one (or more) of the remedial English courses. The numbers in 4.10 imply that a better cutoff for the listening portion of the test might be 24 instead of 30, as 24 is approximately one standard deviation below the mean. However, feedback from the instructors for the listening course does not support changing the test significantly.

Conclusion

These results imply that good tests could be constructed using these ideas. If tests could be made simply by using some computer program as a filter, and the tests had a reasonable level of reliability, then those needing large numbers of tests would greatly benefit. Also, those needing tests for ESP could use this to help measure language abilities in the needed area by using typical required readings as the texts on which the tests are based. While the limitations of cloze tests are clear, having cloze tests based on texts drawn from actual texts that the students may encounter can be useful. This could provide a degree of content validity for the non-TESL specialist that the scores obtained might have more validity. It would also be useful to attempt to create cloze tests across disciplines, based on readings that educators in specific areas consider standard and see if there is a correspondence across disciplines, but within research areas. For example, selections from a standard reading list from chemistry might be used to create tests for new chemistry graduate students, while texts from economics could be used for the economics students. These two sets of scores could be looked at to help determine reading levels required by the different disciplines, and might enable departments to set their own requirements for entering students in addition to TOEFL scores.

5 CONCLUSION

Overall, rational cloze tests built using cohesion can provide a degree of reliability needed for producing parallel cloze tests. This is not surprising, given the high reliabilities others have seen from the cloze. However, the correlations of these tests to the EPT and each other does not conclusively demonstrate that these are parallel tests; thus, this research is still a work in progress.

Some effects of the following variables need to be explored and perhaps changed to produce more highly correlated tests:

- number of deletions
- length of the underlying text
- reading level of the underlying text
- method for determining items to be deleted
- scoring method

While many of these have been explored extensively for fixed-ratio cloze tests, examining them for rational cloze tests is still an open area. It is hoped that this work could contribute something to that area of research. The work of Bachman [5], in particular, helps provide a framework that might prove even more fruitful in producing parallel cloze tests.

Another clear result is that more complex measures of cohesion are needed. While it is not clear that more complex measures would directly result in higher reliability, it is a reasonable expectation. Lexical repetition is so primitive that other types of cohesion should be considered when determining items for deletion. It could be quite useful, for example, to use lexical repetition only for the main ideas of the text, and use conjunction to get a baseline of language ability. A test constructed in such a manner could be used for both ESP and ESL. However, the problem of automatically measuring cohesion is not easy but requires more sophistication from computer programs than is currently available. However, current CALT research does provide some interesting ideas: in particular the idea of using immediate

feedback and "second chances" to grade tests[30]. If such tests could be automatically produced and graded, then the costs would be minimal, thus making good measures of language proficiency more widely available. However, the results of this thesis show that such a system is not yet available.

In conclusion, there are still many questions to be answered, but cohesion might help produce cloze tests that are parallel enough to be produced quickly, easily, and cheaply.

APPENDIX A TEXTS

The Background of Experience

The familiar saying that the exception proves the rule contains a good deal of wisdom, though from the standpoint of formal logic it became an absurdity as soon as "prove" no longer meant "put on trial." The old saw began to be profound psychology from the time it ceased to have standing in logic. What it might well suggest to us today is that, if a rule has absolutely no exceptions, it is not recognized as a rule or as anything else; it is then part of the background of experience of which we tend to remain unconscious. Never having experienced anything in contrast to it, we cannot isolate it and formulate it as a rule until we so enlarge our experience and expand our base of reference that we encounter an interruption of its regularity. The situation is somewhat analogous to that of not missing the water till the well runs dry, or not realizing that we need air till we are choking.

For instance, if a race of people had the physiological defect of being able to see only the color blue, they would hardly be able to formulate the rule that they saw only blue. The term blue would convey no meaning to them, their language would lack color terms, and their words denoting their various sensations of blue would answer to, and translate, our words "light, dark, white, black," and so on, not our word "blue." In order to formulate the rule or norm of seeing only blue, they would need exceptional moments in which they saw other colors. The phenomenon of gravitation forms a rule without exceptions; needless to say, the untutored person is utterly unaware of any law of gravitation, for it would never enter his head to conceive of a universe in which bodies behaved otherwise than they do at the earth's surface. Like the color blue with our hypothetical race, the law of gravitation is a part of the untutored individual's background, not something he isolates from that background. The law could not be formulated until bodies that always fell were seen in terms of a wider astronomical world in which bodies moved in orbits or went this way and that.

Similarly, whenever we turn our heads, the image of the scene passes across our retinas exactly as it would if the scene turned around us. But this effect is background, and we do not recognize it; we do not see a room turn around us but are conscious only of having turned our heads in a stationary room.

If we observe critically while turning the head or eyes quickly, we shall see, no motion it is true, yet a blurring of the scene between two clear views. Normally we are quite unconscious of this continual blurring but seem to be looking about in an unblurred world. Whenever we walk past a tree or house, its image on the retina changes just as if the tree or house were turning on an axis; yet we do not see trees or houses turn as we travel about at ordinary speeds. Sometimes ill-fitting glasses will reveal queer movements in the scene as we look about, but normally we do not see the relative motion of the environment when we move; our psychic makeup is somehow adjusted to disregard whole realms of phenomena that are so all-pervasive as to be irrelevant to our daily lives and needs.

If I'm so smart, How come I Flunk All the Time?

Can twenty flunking students of varying intelligence raise their math and English a full year's level in only thirty working days?

Dr. Lloyd Homme, chief of a special educational "fix-it" laboratory in Albuquerque, New Mexico, said Yes and put teams of behavioural scientists together with the flunking students to work on the problem. Any available technology could be used– teaching machines, programmed instruction, computer- assisted methods–to cram a year's knowledge into the boys.

Were the experiments a success? The scientists said Yes, but the students said No. When grades were measured using standardized tests under strict laboratory conditions, marks went up more than one year on the average. Meanwhile, back at school, the students were still barely passing, at best. "The experiment was fine for the scientists. They proved their theory on paper and made a name for themselves, but most of us were still flunking in class," remarked one seventeen-year-old.

The only clue to the mystery was this common remark: "The teachers ignore us–they've got it in for us."

At first the scientists on the team thought the complaint was just sour grapes and told the boys to work harder. When grades still failed to rise, the scientists felt there might be some truth in what the young team members were saying. Not that teachers were to blame, necessarily, but there still might be some negative bias. "You should see what goes on in class!" said the boys.

"The only thing to do was to take them up on it, go into the classroom with them and see what was holding back their grades," said Dr. Homme.

Hence, bearded behavioural scientists ended up in the back row of math and English classes and made observations about the behavior of students and teachers. Homme was surprised to discover that two simple actions made the difference.

"With few exceptions, our students acted like dummies," said Dr. Homme, "even though we knew they were ahead of the rest in knowledge. They were so used to playing the class idiot that they didn't know how to show what they knew. Their eyes wandered, they appeared absent-minded or even belligerent. One or two read magazines hidden under their desks, thinking, most likely, that they already knew the classwork. They rarely volunteered and often had to have questions repeated because they weren't listening. Teachers, on the other hand, did not trust our laboratory results. Nobody was going to tell them that 'miracles' could work on Sammy and Jose."

In the eyes of the teachers, students seemed to fall into three groups. We'll call them: bright-eyes, scaredy-cats and dummies.

Bright-eyes had perfected the trick of:

1. "eyeballing" the instructor at all times, even from the minute he entered the room.
2. never ducking their eyes away when the instructor glanced at them.
3. getting the instructor to call on them when they wanted without raising their hands.
4. even making the instructor go out of his way to call on someone else to "give others a chance" (especially useful when bright-eyes themselves were uncertain of the answer).
5. readily admitting ignorance so as not to bluff–but in such a way that it sounds as though ignorance is rare.
6. asking many questions.

Scaredy-cats (the middle group)

1. looked toward the instructor but were afraid to let him "catch their eyes."
2. asked few questions and gave the impression of being "under achievers."
3. appeared uninvolved and had to be "drawn out," so they were likely to be criticized of "inadequate participation."

Dummies (no matter how much they really knew)

1. never looked at the instructor.
2. never asked questions.
3. were stubborn about volunteering information in class.

To make matters worse, the tests in school were not standardized and not given nearly as frequently as those given in the laboratory. School test scores were open to teacher bias. Classroom behavior of students counted a lot toward their class grades. There was no doubt that teachers were biased against the dummies. The scientists concluded that no matter how much knowledge a dummy gained on his own, his grades in school were unlikely to improve unless he could somehow change his image into a bright-eyes. This would mean...

1. Look the teacher in the eye.
2. Ask questions and volunteer answers.

"Teachers get teacher training in how to play their roles. Why shouldn't students get student-training in how to play bright-eyes?" asked Homme. Special training sessions where held at the laboratory. Dummies were drilled in eyeballing and hand-raising, which, simple as they sound, weren't easy to do. "I felt so square I could hardly stand it," complained one of the dummies. "That was at first. Later, when I saw others eyeballing and hand-raising and really learning more, I even moved my seat to the front. It flipped the teacher out of her skull. She couldn't get over it."

Those who found eyeballing especially difficult were taught to look at the instructor's mouth of the bridge of her nose. "Less threatening to the student," explained Homme. "It seems less aggressive to them."

Unfortunately, not all of the dummies were able to pick up new habits during the limited training period. Some learned in the laboratory but couldn't do it in the classroom. These became scaredy-cats–at least a step up. But for the majority, grades improved steadily once they got the hang of their new techniques. The students encouraged and helped each other to hand-raise and eyeball.

Teachers' comments reflected the improvement. "There is no doubt that students involvement was increased by the program and as a result grades went up."

By way of advice to others wishing to improve their own eyeballing and hand-raising, student Jose Martinez suggests: "Don't try to do it all at once. You'll shock the teacher and make it rough for yourself. Begin slowly. Work with a friend and help each other. Do it like a game. Like exercising with weights–it takes practice but it's worth it."

Homme agrees. "In fact, results are guaranteed for life," he says.

APPENDIX B SOURCE CODE FOR ANALYSIS PROGRAM

```
;;; This file contains Scheme code for producing word count statistics
;;; by word and by sentence.  Note that it is not designed for efficiency
;;; (as it puts the entire file into memory as a list)

(define get-prefix
  (lambda (strng)
    (let ((ls (reverse (string->list strng))))
      (list->string (reverse (cddddr ls))))))

(define copier
  (lambda (in-file out-file)
    (let ((input-file (open-input-file in-file))
          (output-file (open-output-file out-file)))
      (letrec ((copier (lambda (ch)
                         (if (not (eof-object? ch))
                             (begin
                               (write-char ch output-file)
                               (copier (read-char input-file)))))))
        (copier (read-char input-file))
        (close-input-port input-file)
        (close-output-port output-file)))))

(define file->list
  (lambda (file)
    (let ((in-file (open-input-file file)))
```

```
      (letrec ((reader
                 (lambda (ls)
                   (let ((ch (read-char in-file)))
                     (if (eof-object? ch)
                         ls
                         (reader (cons ch ls)))))))
        (reverse (reader '())))))

(define list->file
  (lambda (ls file)
    (let ((output-file (open-output-file file)))
      (letrec ((copier (lambda (ls)
                         (if (not (null? ls))
                             (begin
                               (write-char (car ls) output-file)
                               (copier (cdr ls)))))))
        (copier ls)
        (close-output-file output-file)))))

(define analyze
  (lambda (file)
    (let* ((file-list (map char-downcase
                           (remove-if-not (file->list file) used?)))
           (sent-list (map break-into-words (get-sentences file-list)))
           (word-list (insertsort (remove-duplicates (flatten sent-list))))
           (num-sentences (length sent-list))
           (word-data (map (lambda (ls)
                             (cons (total ls) ls))
                           (map (lambda (word)
                                  (map (lambda (sent)
                                         (num-occurs word sent))
                                       sent-list))
```

```
                                word-list)))
          (output-file (open-output-file (string-append
                                          (get-prefix file) ".csv"))))
      (letrec ((print-data (lambda (word data)
                             (display (car data) output-file)
                             (display "," output-file)
                             (display word output-file)
                             (for-each (lambda (data)
                                         (display "," output-file)
                                         (display data output-file))
                                       (cdr data))
                             (newline output-file))))
        (display "Total," output-file)
        (display "Word" output-file)
        (for-each (lambda (num)
                    (display "," output-file)
                    (display num output-file))
                  (one-to-n-ls num-sentences))
        (newline output-file)
        (for-each2 print-data word-list word-data)
        (close-output-port output-file))))

(define (one-to-n-ls num)
  (letrec ((aux (lambda (index acc)
                  (if (zero? index) acc
                      (aux (sub1 index) (cons index acc))))))
    (aux num '())))

(define (num-occurs a ls)
  (letrec ((aux (lambda (ls acc)
                  (cond
                    ((null? ls) acc)
```

```
                    ((string=? a (car ls))
                     (aux (cdr ls) (add1 acc)))
                    (else (aux (cdr ls) acc))))))
    (aux ls 0)))

(define for-each
  (lambda (proc ls)
    (if (not (null? ls))
        (begin
          (proc (car ls))
          (for-each proc (cdr ls))))))

(define for-each2
  (lambda (proc ls1 ls2)
    (if (not (null? ls1))
        (begin
          (proc (car ls1) (car ls2))
          (for-each2 proc (cdr ls1) (cdr ls2))))))

(define total
  (lambda (ls)
    (apply + ls)))

(define flatten
  ;; Messes up the order of the words...
  (lambda (ls)
    (letrec ((aux (lambda (ls acc)
                    (cond
                      ((null? ls) acc)
                      ((pair? (car ls)) (append (aux (car ls) '())
                                                (aux (cdr ls) '())
                                                acc))
```

```
                    (else (aux (cdr ls) (cons (car ls) acc)))))))
      (aux ls '())))))


(define remove-duplicates
  (lambda (ls)
    (if (null? ls) '()
        (cons (car ls)
              (remove-duplicates (remove-all (car ls) ls string=?))))))


(define (remove-all item ls pred)
  (letrec ((aux (lambda (ls acc)
                  (cond
                    ((null? ls) acc)
                    ((pred item (car ls))
                     (aux (cdr ls) acc))
                    (else (aux (cdr ls) (cons (car ls) acc)))))))
    (aux ls '())))


(define break-into-words
  (lambda (ls) ;; list of chars
    (letrec ((aux (lambda (ls current acc)
                    (cond
                      ((null? ls)
                       (reverse (cons (reverse current) acc)))
                      ((and (char=? (car ls) #\space) (null? current))
                       (aux (cdr ls) '() acc))
      ((char=? (car ls) #\space)
       (aux (cdr ls)
                            '()
                            (cons (reverse current) acc)))
                      (else (aux (cdr ls) (cons (car ls) current) acc))))))
      (map list->string (aux ls '() '())))))
```

```
(define (used? ch)
  (cond
    ((char-alphabetic? ch) #t)
    ((eof-object? ch) #t)
    ((char=? ch #\-) #t)
    ((char=? ch #\') #t)
    (else (char-member? ch '(#\space #\newline #\? #\! #\.)))))

(define remove-if-not
  (lambda (ls pred)
    (letrec ((aux (lambda (ls acc)
                    (cond
                      ((null? ls) acc)
                      ((pred (car ls)) (aux (cdr ls)
                                            (cons (car ls)
                                                  acc)))
                      (else (aux (cdr ls) (cons #\space acc)))))))
      (reverse (aux ls '())))))

(define member?
  (lambda (item ls)
    (cond
      ((null? ls) #f)
      ((equal? item ls) #t)
      (else (member? item (cdr ls))))))

(define get-sentences
  (lambda (data-list)
    ;; RETURNS: a list of lists (sublists being sentences)
    (letrec ((aux
              (lambda (data current sent-list)
```

```
                (cond
                  ((and (null? data) (null? current)) sent-list)
                  ((null? data) sent-list)
                  ((or (char-alphabetic? (car data))
                       (char=? (car data) #\')
                       (char=? (car data) #\space))
                   (aux (cdr data) (cons (car data) current) sent-list))
   ((char=? (car data) #\newline)
    (aux (cdr data) (cons #\space current) sent-list))
                  ((end-punct? (car data) data)
                   (aux (finish-sentence (cdr data))
                        '()
                        (cons (reverse current) sent-list)))
   ((char=? (car data) #\-)
    (if (hyphenated? data)
(aux (cdr data) (cons (car data) current) sent-list)
(aux (cddr data)
     (cons #\space current) sent-list)))
   ((or (char=? (car data) #\.)
(char=? (car data) #\!)
(char=? (car data) #\?))
    (aux (cdr data) current sent-list))
                  (else (error "Found something I shouldn't have: "
                               data))))))
      (reverse (aux data-list '() '())))))

(define (hyphenated? char-ls)
  (and (char=? (car char-ls))
       (not (null? (cdr char-ls)))
       (char-alphabetic? (cadr char-ls))))

(define (finish-sentence ls)
```

```
  (cond
    ((null? ls) '())
    ((char=? (car ls) #\newline) (cdr ls))
    ((and (char=? (car ls) #\space)
          (null? (cdr ls))) '())
    ((and (char=? (car ls) #\space)
          (char=? (cadr ls) #\newline)) (cddr ls))
    ((and (char=? (car ls) #\space)
          (char=? (cadr ls) #\space)) (cddr ls))
    (else (error "Should never reach this point in finish-sentence"))))

(define (end-punct? ch ls)
  (if (char-member? ch '(#\. #\? #\!))
      (or (null? (cdr ls))
          (char=? (cadr ls) #\newline)
          (and (char=? (cadr ls) #\space)
               (null? (cddr ls)))
          (and (char=? (cadr ls) #\space)
               (char=? (caddr ls) #\space))
          (and (char=? (cadr ls) #\space)
               (char=? (caddr ls) #\newline)))
      #f))

(define (char-member? ch ls)
  (cond
    ((null? ls) #f)
    ((char=? ch (car ls)) #t)
    (else (char-member? ch (cdr ls)))))

(define (insertsort ls)
  (if (singleton? ls) ls
      (insert (car ls) (insertsort (cdr ls)))))
```

```
(define (singleton? ls)
  (and (not (null? ls)) (null? (cdr ls))))

(define (insert x ls)
  (cond
    ((null? ls) (cons x '()))
    ((string<? x (car ls)) (cons x ls))
    (else (cons (car ls) (insert x (cdr ls))))))

(define (add1 x) (+ 1 x))
(define (sub1 x) (- x 1))
(define (writeln . args)
  (for-each display args)
  (newline))
```

BIBLIOGRAPHY

[1] Alderson, J. C. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly, 13*(2), 219–27.

[2] Alderson, J. C., Windeatt, S. (1991). "Learner–adaptive" computer–based language tests. *Language Testing Update, 9*, 18–21.

[3] Anderson, N. L., Bachman, L., Perkins, K., and Cohen, A. D. (1991). An exploratory study into the construct validity of a reading comprehension test: triangulation of data sources. *Language Testing, 10*, 41–66.

[4] Bachman, L. (1982). The trait structure of cloze test Scores. *TESOL Quarterly 16*(1), 61–70.

[5] Bachman, L. (1985). Performance on cloze tests with fixed–ratio and rational deletions *TESOL Quarterly, 19*(3), 535–556.

[6] Bachman, L. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

[7] Bachman, L., Davidson, F., and Milanovic, M. (1991). The use of test method characteristics in the content analysis and design of English proficiency tests. Paper presented at the annual Language Testing Research Colloquium. Princeton, New Jersey.

[8] Bachman, L., Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*, 449–465.

[9] Bachman, L., et al. (1990). nThe relationshiop of expert system scored constrained free response items to multiple–choice and open–ended items. *Applied Psychological Measurement. 14*, 151–162.

[10] Bradshaw, J. (1990). Test–takers' reactions to a placement test. *Language Testing. 7*, 13–30.

[11] Butler, C. (1985). *Statistics in Linguistics.* Oxford: Basil Blackwell Inc.

[12] Chapelle, C. A., Abraham, R. G. (1990). Cloze method: what difference does it make? *Language Testing, 7*(2), 121–146.

[13] Chapelle, C. (1988). Field independence: a source of language test variance? *Language Testing 5*(1), 62–82.

[14] Clapham, C. (1991). The effect of academic discipline on reading test performance. Paper presented at the annual Language Testing Research Colloquium. Princeton, New Jersey.

[15] Darnell, D. (1970). Clozentropy: a procedure for testing English language proficiency of foreign students. *Speech Monographs 37*, 36–46.

[16] DeStefano, J. (1990). Assessing students' communicative competence using a linguistic analysis procedure. *Linguistics and Education: an International Research Journal, 2*(2), 127–145.

[17] Dollerup, et al. (1994). "Sprogtest": a smart test (or how to develop a reliable and anonymous EFL reading test). *Language Testing 11*, 65–82.

[18] Douglas, D. ed. (1990). *English Language Testing in U. S. Colleges and Universities.* Washington: National Association for Foreign Student Affairs.

[19] Douglas, D., Chapelle, C. eds. (1993). *A new decade of language testing.* Alexandria, Virginia: TESOL Publications.

[20] Douglas, D. ed. (1995). Developments in language lesting. *Annual Review of Applied Linguistics 15*, 167–187.

[21] Educational Testing Service (ETS). (1991). *Computerized Placement Tests* Princeton, New Jersey: Educational Testing Service.

[22] Farghal, M. (1992). Naturalness and the notion of cohesion in EFL writing classes. *International Review of Applied Linguistics in Language Teaching. 30*(1), 45–50.

[23] Gale, S. (1980). *Readings for Today's Writers.* New York: John Wiley & Sons.

[24] Gulliken, H. (1950). *Theory of Mental Tests.* New York: John Wiley & Sons, Inc.

[25] Halliday, M. A. K., Hasan, R. (1976) *Cohesion in English.* London: Longman.

[26] Hanania, E., and Shikhani, M. (1986) Interrelationships among three tests of language proficiency: standardized ESL, cloze, and writing. *TESOL Quarterly, 20*(1), 97–109.

[27] Heaton, J. (1990). *Classrom Testing.* London: Longman.

[28] Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing, 9,* 1–11.

[29] Henning, G., Johnson, P, Boutin, A, and Rice, H. (1994). Automated assembly of pre–equated language proficiency tests. *Language Testing, 11,* 15–28.

[30] Henning, G., et al. (1993). Computer-assisted testing of reading comprehension: Comparisons among multiple-choice and open-edded scoring methods. In D. Douglas and C. Chapelle (eds.) *A New Decade of Language Testing Research.* Alexandria, Virginia: TESOL Publications.

[31] Horning, A. (1991). Readable writing: the role of cohesion and redundancy. *Journal of Advanced Composition, 11*(1), 135–45.

[32] Hudson, T. (1993). Testing the specificity of ESP reading skills. In D. Douglas and C. Chapelle (eds.) *A New Decade of Language Testing Research.* Alexandria, Virginia: TESOL Publications.

[33] Hughes, A. (1989). *Testing for Language Teachers.* Cambridge: Cambridge University Press.

[34] Irvine, P., Atai, P., Oller, J. W. Jr. (1974). Cloze, dictation, and the Test of English as a Foreign Language. *Language Learning, 24,* 245–252.

[35] Jonz, J. (1987). Textual cohesion and second-language comprehension. *Language learning: A Journal of Applied Linguistics, 37*(3), 409–438.

[36] Jonz, J. (1991a). Cloze item types and second language comprehension. *Language Testing, 8,* 1–22.

[37] Jonz, J. (1991b). What cloze tests measure: A factor analysis with replications. Paper presented at the annual TESOL Convention. New York, New York.

[38] Lado, R. (1986). Analysis of native speaker performance on a cloze test. *Language Testing, 3*(2), 130–146.

[39] Laurier, M. (1991). What we can do with computerized adaptive testing—and what we cannot do! In S. Anivan (ed.) *Current Developments in Language Testing.* Singapore: SEAMEO Regional Language Centre. 244–255.

[40] Lovejoy, K., Lance, D. (1991). Information management and cohesion in the study of written discourse. *Linguistics and Education: An international research journal, 3*(3), 251–273.

[41] Morris, J., Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics, 17*(1), 21–48.

[42] Moss, P. (1994) Can there be validity without reliability? *Educational Researcher, 23*(2), 5–12.

[43] Myers, G. (1991). Lexical cohesion and specialized knowledge in science and popular science texts. *Discourse Processes: a multidisciplinary journal, 14*(1), 1–26.

[44] Oller, J. W. Jr., Conrad, C. A. (1971). The cloze technique and ESL proficiency. *Language Learning, 21*, 183–196.

[45] Oller, J. W. Jr. (1972). Scoring methods and difficulty levels for cloze tests of proficiency in ESL. *Modern Language Journal, 56*, 151–158.

[46] Oller. J. (1973). Cloze tests of second language proficiency and what they measure. *Language Learning, 23*, 105–118.

[47] Oller, J. W. Jr. (1978). *Language Tests in School.* London: Longman.

[48] Olsen, L. (1989). A discourse–based approach to the assessment of readability. *Lingustics and Education: An International Research Journal, 1*(3), 207–231.

[49] Perkins, K. (1992). The effect of passage topical structure types on ESL reading comprehension difficulty. *Language Testing, 9*, 163–172.

[50] Reid, J. (1992). A computer text analysis of four cohesion devices in English discourse by native and nonnative writers. *Journal of Second Language Writing, 1*(2), 79–107.

[51] Rekart, D., Dunkel, P. (1992). The utility of objective (computer) measures of the fluency of speakers of English as a second language. *Applied Language Learning, 3*, 65–85.

[52] Reynolds, T., Perkins, K., and Brutten, S. (1994). A comparative item analysis study of a language testing instrument. *Language Testing, 11*, 1–14.

[53] Rost, D. (1993). Assessing the different components of reading comprehension: fact or fiction. *Language Testing, 10*, 79–92.

[54] Sang, F., Schmitz, B., Vollmer, H., Baumert, J., and Roeder, P. (1986). Models of second language competence: a structural equation approach. *Language Testing, 3*, 54–79.

[55] Sasaki, M. (1991). A comparison of two methods for detecting differential item functioning in an ESL placement test. *Language Testing, 8*, 95–111.

[56] Schils, E., van der Poel, M, Weltens, B. (1991) The reliability ritual. *Language Testing, 8*, 125–138.

[57] Shohamy, E. (In press). Language testing and SLA interfaces: the case of discourse. In L. F. Bachman and A. D. Cohen (eds.) *Interfaces between SLA and language testing research.* Cambridge: Cambridge University Press.

[58] Sigott, G, Koberl, J. (1993). Validating the X–test. *Language Testing Update 14*, 53–58.

[59] Simpson, P. (1992). Teaching stylistics: analyzing cohesion and narrative structure in a short story by Earnest Hemingway. *Language and Literature: Journal of the Poetics and Linguistics Association, 1*(1), 47–67.

[60] Slack, C. (1980). If I'm so smart, how come I flunk all the time? In Gale, S. ed. *Readings for Today's Writers*, New York: John Wiley & Sons.

[61] Stansfield, C., Hansen, J. (1983). Field dependence–independence as a variable in second language cloze test performance. *TESOL Quarterly, 17*(1), 29–38.

[62] Steffensen, M. (1986). Register, cohesion, and cross-cultural reading comprehension. *Applied Linguistics, 7*(1), 71–85.

[63] Stevens, V. (1991) Strategies in solving computer–based cloze: is it reading? Paper presented at the annual TESOL Convention. New York.

[64] Stotsky, S. (1983). Types of lexical cohesion in expository writing: implications for developing the vocabulary of academic discourse. *College Composition and Communication, 34*(4), 430–445.

[65] Taylor, C., Buck, G. (1994). TOEFL 2000: Language testing in the future. Paper presented at the annual TESOL Summer Institute. Ames, Iowa.

[66] Taylor, W. (1953). Cloze procedure: a new tool for measuring readability. *Journalism Quarterly, 33*(42), 8.

[67] Whorf, B. (1980). The background of experience. In Gale, S. ed. *Readings for Today's Writers*, New York: John Wiley & Sons.

[68] Yang, A. (1989). Cohesive chains and writing quality. *Journal of the International Linguistic Association, 40*(1–2), 235–254.

[69] Young, R., Perkins, K., Brutten, S. (1993). Designing computer–adaptive tests: A practical solution. Paper presented at the annual TESOL Convention. Atlanta, Georgia.