

AN ENGLISH GRAMMAR PROFICIENCY EXAM
FOR NATIVE SPANISH SPEAKERS

George Keeler
Joan Walker

Universidad de las Americas, A. C.
Cholula, Mexico

This paper has a threefold purpose: first, to describe the development of an English grammar proficiency examination for native Spanish speakers at the Universidad de las Américas, A. C. located in Cholula, Puebla, Mexico. Second, to examine under what circumstances it may be advisable to develop one's own proficiency examination instead of using a standardized examination, as concluded by this investigation. Third, to share with colleagues, who are regular classroom ESL/EFL teachers and not test specialists, suggested steps to be followed and problems to be aware of when developing a grammar examination.

There are five sections in this paper. The first briefly describes the role of English at the Universidad de las Américas. The second explains why it was decided to develop such an exam. The third presents the steps followed in the construction of the exam. The fourth discusses statistical analysis and usability. The final section offers concluding statements with suggestions and insights gained.

1. The Role of English at the Universidad de las Américas

The fundamental goal of the Language Department of the Universidad de las Américas is to produce functionally bilingual students (English/Spanish) at both the undergraduate and graduate levels, whether they are native Spanish speakers or native English speakers. A native Spanish speaking student who does not have functional mastery of English upon entering the University must plan to acquire that proficiency during the course of his studies. One of the graduation requirements for a native Spanish speaker is that he pass a proficiency exam offered by the Department of Languages which certifies that he is functionally bilingual. Thus, a student must plan to take sufficient coursework in English to prepare him to pass a proficiency exam. Proficiency in English follows the basic philosophy of the University: to establish a bilingual-multi-cultural environment.

It was decided that a proficiency exam should test a student's working knowledge of English as needed to function within the University and had to be

based on the student's previous classroom exposure to English. The latter was decided upon as most of the students at the University do not acquire English by spending extended periods of time in English speaking countries. For the purposes of the Language Department, a "proficiency exam" is a measure of whether the student has acquired sufficient knowledge of English to perform in and pass his academic coursework that is taught in English at the University.

2. Why a "Special" Examination Was Developed

After reviewing certain standardized proficiency exams such as the Michigan Test of English Language Proficiency, Structure Test-English Language (STELL), English Achievement Series (EAS), and the Comprehensive English Language Test (CELT), it was concluded that none of these exams met the needs of the U. D. L. A. Since the General English Division of the Language Department deals with a large majority of native Spanish speakers, it was considered essential to focus on those obvious and most commonly used grammatical structures that present particular difficulty for these native Spanish speakers.

Most standardized English exams are developed to test at random the whole grammatical continuum of the language in order to avoid giving any kind of advantage to one non-native speaker from one linguistic and cultural background over another non-native speaker from another linguistic and cultural background. In other words, by taking no specific language group into account, an attempt is being made to minimize linguistic bias. This type of exam is sound for programs which exist in English speaking countries where English is taught to people from diverse language backgrounds.

This type of exam failed to coincide with the assessment of proficiency for the University because of what was felt to be the omission of indispensable items and the inclusion of unessential material. Additionally, a working proficiency in English at the Universidad de las Américas is different from proficiency in English in English-speaking countries and is, therefore, somewhat different from the English tested on standardized exams.

3. Construction of the Exam

3.1 First Pretest. The project was begun by establishing the most important English grammar points that should be tested in order to meet the needs of the University. First, a memorandum was sent to teachers in the Language Department, including Spanish teachers, asking each to list those areas of English grammar he felt presented particular difficulty for native Spanish speakers. In addition, a list of items was constructed by reviewing structural patterns that are found in texts for teaching English grammar to speakers of other languages. With this information items were written for the first pretest. A detailed outline of the test content was not written, nor

was "the percentage of items to be written around each problem" indicated as suggested by Harris (Harris 1969:25). However, three to five items were written around forty-seven grammar points compiled from the teachers' suggestions and the list constructed from reviewing structural patterns that are found in grammar books.

Each item--"an individual problem on a test" (Lyman 1963:191)--was written in the form of a dialogue with a local setting using informal spoken as opposed to formal spoken or written language. According to Harris, "The dialogue form provides more context and therefore may make the problem somewhat clearer" (Harris 1969:26). Steps were taken to make each item different from a question-answer type dialogue. An attempt was made to be clear and concise and to follow Harris's suggestion: "The second part of the dialogue should sound like a natural response to the first part" (Harris 1969:30). Distractors were developed which were considered to be unnatural to the situation for native English speakers, and therefore not social or regional variants of English. Many distractors which were included were incorrect responses considered to be frequently made by native Spanish speakers (See Appendix I).

In organizing the first pretest, some of the items were eliminated due to the unnatural language of the dialogue. Finally, it was decided that a total of 141 items were to be included on the first pretest.

Forty-nine Spanish-speaking students took the first pretest. Represented were students from the basic, intermediate, and advanced level English courses at the University as well as individuals who had taken or were taking academic courses offered in English at the University. The purpose of this wide range of students was to determine whether the items were either too difficult or too easy.

The results of this pretest were subjected to statistical analysis and a copy of the pretest was submitted to each of the full-time English teachers in the Language Department for constructive criticism. Based on their suggestions and the statistical results, modifications of individual items and distractors were made, and new items were written to be included on the second pretest. The altered and unaltered items which were not discarded were evenly distributed among the three forms of the second pretest. The new items were included on all three forms of the second pretest (See Appendix II: A, B, C and D).

One of the objectives from the beginning of the project was to develop three final forms of the exam each being the same length, having approximately the same degree of difficulty, and testing the same grammar points. The reason for developing three distinct final forms was to decrease the probability of test sensitivity on the part of the student who must take the exam more than once and the outside possibility of theft both of which could

jeopardize the reliability of the exam.

3.2 Second Pretest. After the three forms were prepared, they were administered again to students from the three different levels of English courses in the General English Division. One hundred students took this second test.

Once again, based on statistical analysis, the new and old items and their distractors were evaluated. Most of the new items were kept with little alteration for what would make up the final three forms of the exam. The second pretest presented the opportunity to check again the difficulty of each item from the first pretest and to determine whether the altered distractors were better than the previous ones. Some of the distractors were kept as they appeared on the second pretest, and others were changed to the way they had appeared on the first pretest (See Appendix II: E and F).

With duplicate, and in some cases triplicate, data on each item the final three forms of the examination were organized. First, a list was made, categorizing the different items according to the grammar points to be tested. Once this was completed it was decided, based on what was considered to be the most important grammatical points, how many items testing a particular point were necessary to have on each form of the final exam. Each test form had the same number of items for each grammar point. Some items were discarded because they were either too easy or too difficult. The degree of difficulty of each remaining item of the same type was checked to determine the distribution on the three forms. When there were three items of similar difficulty, testing the same grammar point, one went on each form. If there were only two, then the item with the better degree of difficulty went on two forms and the third on the last form. Special care was taken not to put the "easiest" items, as determined by degree of difficulty, all on one form. When there was only one item with an excellent degree of difficulty, it was placed on all three forms.

After deciding on the content and organization for these three forms typing began. Care was taken to space the items evenly, allowing for easy readability. The items were not arranged from easy to more difficult. Although the forms were checked to see that there was no observable pattern to the correct answer positions, it was not checked to determine whether each correct answer position was used an equal number of times.

3.3 Final Forms of the Examination. After the final forms of the exam were typed, they were given to a class of native English speakers in order to check the distractors, the language used in the items, as well as to locate any typographical errors. Some good suggestions were received as to the wording of some of the items and distractors, and a few typographical errors were found; alterations were made accordingly. Each of the final forms has sixty test items with one correct answer and three distractors.

3.4 Instruction Page. The instruction page for the pretests was written

in Spanish to minimize the possibility of misinterpretation of the test procedures. However, the instruction page for the final forms of the exam was written in English, as it was felt that any student declared eligible to take this exam should be able to read the instructions in English. Both in the Spanish and English forms of the instruction page, instructions were written in order to be brief and easily understood. Also, all the essential points related to writing instructions were followed that are suggested in books on testing: 1) giving sufficient examples, 2) indicating how to mark the answers, 3) instructing whether or not guessing is recommended, and 4) indicating the length of time allowed for the test.

3.5 Answer Sheet. The answer sheet for the final forms of the exam is printed for sixty items. Care was taken in spacing, so that a student will not become confused when marking his answers. Space is provided at the top of the answer sheet for personal data. A special space is blocked out and labeled for the two example items. The answer sheets are designed to be corrected by the University's computer or by punched keys.

3.6 Establishing the Standard for Proficiency. The final three forms were given to a control group of native Spanish-speaking students who had studied English at the University and taken academic courses taught in English but had not had lengthy exposure to the English language by either living or traveling extensively in an English-speaking country. This step was taken to establish the standard to be used for proficiency; in other words, to determine the minimum passing score on the exam. The standard was set by the average attained by these students.

4. Statistical Analysis and Usability

Before going into the details of the statistical analysis that was conducted, it would be beneficial to point out a few facts before deciding that statistical analysis is too complicated to use in evaluating a test. It is not necessary to be an expert in statistical analysis to evaluate a test. A basic understanding of statistical analysis is sufficient. There is ample literature available which simplifies statistical analysis. Some basic statistics are required in order to evaluate the effectiveness and dependability of a test. In addition, when constructing a test with more than one form it is necessary to compute basic statistics so that each form is equally reliable in testing an individual's proficiency.

"Validity", "reliability", and "usability" are the three essential features of this type of test. Validity and reliability can be evaluated by means of statistical analysis as well as by other factors not related to statistics. The basic statistical measures that need to be calculated to determine validity and reliability of a test should consist of the following: range, mean, standard deviation, reliability coefficient, standard error of measure, general degree of difficulty and item degree of difficulty. The formulas for calculating these measures and how they relate to validity

and reliability can be found in any book on basic statistics or educational measurement.

4.1 Validity. The validity of a test refers to whether the test is "appropriate in terms of our objectives" (Harris 1969:13) - that is, whether or not the test is a true measure of what it is intended to measure. Experts differ in terms of the number of types of validity that should be determined, but the following types of validity mentioned are those on which most experts agree are vital.

Three kinds of validity that can be evaluated without the use of statistical data are content validity, curricular validity and face validity. Content validity consists of a good "sampling of a specific universe of content" (Ebel 1969:437). After a specified collection of all the elements to be considered on a test, for example English grammar, has been identified, specific parts of that collection are selected (Shao 1974:99). In order to determine whether a test has content validity or not, a professional judgment must be made as to whether the "sampling" adequately and proportionately represents the "universe of content". It is suggested the professional judgment of an "outside" expert, someone not involved in the preparation but familiar with the subject of the test, be sought to evaluate the content validity of a test (Ebel 1965:438). Curricular validity "---is determined by examining the content of the test itself and judging the degree to which it is a true measure of the important objectives of the course, or a truly representative sampling of the essential materials of instruction" (Ebel 1965:437). Familiarity with the specific objectives of a particular program will allow the test developer to check to see that the contents of the test cover the objectives of his program. Face validity "refers . . . to what the test appears to measure" (Ebel 1965:437). That is, a brief examination of the test will indicate whether the test measures what it is designed to measure.

The U. D. L. A. exam was subjected to evaluation on all three of these kinds of validity. With reference to content validity, the "universe of content" of the exam is English grammar as needed to function within the Universidad de las Américas. The "sampling" includes, aside from indispensable areas of English grammar such as verbs, subject-object pronouns and adjectives, those areas which were felt to present particular difficulty to native Spanish speakers, such as distinguishing between the use of "make" and "do" and the use of articles, prepositions, etc. As noted earlier, in order to assure curricular validity, a memorandum was sent to several members of the English teaching faculty as well as to Spanish teachers asking for their suggestions as to what should be put on an English grammar proficiency exam for the University. Incorporation of their suggestions gave the test the curricular validity it needed. Then an examination of the pretest exams and the final forms of the exam was made to verify that each of the exams appeared to be an accurate measure of English grammar.

A fourth type of validity which requires the use of statistical data is criterion-related validity (also referred to as empirical validity), which "refers to the relationship between test scores and a criterion, the latter being an independent and direct measure of what the test is designed to predict" (Ebel 1965:437). It is established by a coefficient of correlation, which is similar to the coefficient used for determining a test's reliability, and through statistical comparison of the test scores with the values or scores on an outside variable, such as classroom grades or scores on another test. While developing the U. D. L. A. exam, the criterion-related validity was not determined although it is being established now.

4.2 Reliability. The reliability of a test refers to the "consistency or stability of a test" (Lyman 1963:194), whether a test obtains the same scores or results regardless of the testing conditions (Harris 1969:14). There are several ways to determine the reliability of a test; these are: the test-retest method, the equivalent forms method, the split-halves method, the reliability coefficient, and the standard error of measurement. For best results it is strongly suggested that, if feasible, at least two of these methods should be used to determine a test's reliability. However, beforehand one must have a thorough understanding of the implication of their use. Not all of these methods were used for the U. D. L. A. exam for the following reasons. Because of the alterations, additions and deletions that were made on the second pretest, it was not feasible to use the test-retest method. The equivalent forms method which calls for giving the same student two forms of a test that are of equal difficulty was not used because there was no indication at the time that the different forms of the exam were parallel or equivalent. The split-halves method, which involves "splitting a test... to score the odd-numbered items and the even-numbered items separately," and comparing the results (Ebel 1965:413), could have been used with the first pretest but would not have worked with the three forms of the second pretest since each of the three forms had too few items for this method to work effectively. Therefore, in order to determine the test's reliability, it was decided to use the reliability coefficient and the standard error of measurement, each of which is calculated by a specific formula.

"A reliability quotient of 1.00 would indicate a test is 'perfectly' reliable. A quotient of zero would denote a complete absence of reliability" (Harris 1969:16). Most test constructors are satisfied with a coefficient of approximately 0.90 (Ebel 1965:421). Using the Kuder-Richardson formula 20, we obtained the following results: on the first pretest the coefficient was 0.90; on the second pretest the coefficients were 0.93 for Form A, 0.92 for Form B and 0.88 for Form C. All of these coefficients indicated that the pretests were reliable.

"All educational measurement contains some degree of error" (Harris 1969:17), which means that a student given the same test a second time under exactly the same conditions of the first testing will

probably obtain a different score than that obtained the first time. If a test's reliability coefficient never reaches 1.00, it indicates that there is some room for error (Ebel 1965:421). Therefore, the standard error of measurement (SEM), which indicates the range of variation of a score needs to be calculated. Tables in some books on testing list what should be the SEM of a test with a particular number of items. The fact that the SEM is smaller than that indicated on the table is one indication of the high reliability of a test. The first pretest had a standard error of measurement that is average for a test of 141 items. On the second pretest Form A's standard error was 2.68, Form B's was 2.88 and Form C's was 2.81. The fact that these results are slightly low for a test of fifty-four to fifty-six items (the normal SEM for a test with this number of items is 4) is a further indication that the second pretest was highly reliable.

In constructing the final forms of the exam, two statistical measures proved to be important: the general degree of difficulty and the item degree of difficulty, which is one of the three steps of item analysis.

4.3 General Degree of Difficulty. The general degree of difficulty, also known as the mean difficulty of the items, is the proportion of correct responses to the total number of items on the exam. Harris defines a test as easy if the degree of difficulty is between 70 percent and 90 percent and difficult if the degree of difficulty is between 50 percent and 70 percent (Harris 1969:144-45). The degree of difficulty for the first pretest was 66.8%. The degree of difficulty for each form of the second pretest was: Form A = 74.4%, Form B = 65.5% and Form C = 72.0%, equalling an average of 70.6%. These results indicate that the grammar proficiency exam tends to be slightly difficult, a good characteristic for a test that is meant to separate the proficient from the non-proficient.

4.4 Item Analysis and Difficulty. As explained previously, after each pretest an item analysis was carried out in order to eliminate those items which proved too difficult or too easy, "as not contributing significantly to the measurement function of the test" (Harris 1969:105). There are three steps involved in item analysis to completely determine an item's difficulty; these are, item degree of difficulty, item discrimination and effectiveness of the distractors (Harris 1969:105-07).

The item degree of difficulty indicates which items should be discarded because they are either too easy or too difficult. The degree of difficulty is calculated by dividing the total number of individuals who take the test into the number who chose the correct answer for each item. As suggested by Harris, those items with a degree of difficulty greater than 92 percent (which are considered to be too easy) and less than 30 percent (which are considered to be too difficult) were eliminated (Harris 1969:105). Eight of the 141 items from the first pretest and four of the nineteen new items which were added to the second pretest were discarded.

4.5 Usability. Usability refers to practical factors such as "cost,

ease of scoring, time required" (Lyman 1963:21), ease of administration and of interpretation of results, and involves no statistical evaluation. The cost of the U. D. L. A. test is minimal because the tests and answer sheets can be printed and scored at the University. Scoring is done by the University computer, which is programmed to correct the test as well as print out statistical results in raw scores and percentiles. This allows for ease in interpreting the results. During the second pretest, note was taken of the length of time required by the majority of the students to finish the test; this determined the time limit for the final test. The test can easily be given in one of the large rooms at the University.

5. Summary

5.1 Conclusions. The U. D. L. A. English grammar proficiency exam was designed to measure a student's competence in English grammar. Its function is to separate the proficient students from the non-proficient.

The principal aim in carrying out this project was to design a test with three forms that would be reliable, valid, usable, and, therefore, effective as an English proficiency exam for the University. The relatively high reliability coefficients and the degrees of difficulty of the pretests showed that the test items were consistent and stable, and the final test appeared to be neither too easy nor too difficult and effectively separated proficient students from the non-proficient. However, in order to continue to improve upon the reliability of the items of the test, the item discrimination and the item difficulty are currently being calculated for all three forms of the final test. The test was found to have content, curricular, and face validity. The criterion-related validity of the test is being evaluated by comparing scores on the exam to grades students receive in their English classes. In terms of what has been accomplished in the time that has been spent working on the exam, it is felt that the project was very successful.

5.2 Insights Gained. The following are some of the most important insights gained from carrying out this project:

1. Perhaps the greatest factor in developing a reliable and valid test is time, which is a limited commodity for all full- or part-time faculty members as well as full- or part-time students.
2. It becomes difficult due to time and scheduling to assemble all the same subjects to take all the pretests.
3. It is rather difficult to write a sufficient number of dialogue type items covering the same grammar point that employ natural language and are neither too easy nor too difficult.
4. It is not always so easy to establish a reliable external variable for determining criterion-related validity.

5. It is essential to have brief and easily understood instructions with examples indicating the procedures for the test.
6. It is recommended that the test be accompanied by a manual explaining in detail how to administer the test.
7. Statistics are vital for developing such a test and statistical analysis is not difficult to learn and apply.

5.3 Suggestions. The following are suggestions for someone who is considering developing a proficiency examination:

1. Review the available standardized examinations to see if one may fit or be adaptable to the needs of the institution where it is to be used.
2. Thoroughly research the different formats of examinations and their development, along with methods for determining test reliability and validity.
3. In selecting a format for an exam, be sure that it is usable in terms of cost, ease of administration, and of interpretation of its results as related to the particular program in which it is to be used.
4. Write items that use two or more grammar points in order not to create the need of writing so many individual items, if you opt for the multiple choice type of exam.
5. Make a detailed outline of the test content indicating the percentage of items to be written around a particular grammar point. This would serve as a checklist for test items, eliminating the possibility of having too many items for one point or not enough for another, and would be useful as a reference to insure that an important point has not been omitted from the test.
6. In typing the pretest(s) and the final form(s) of the test, be certain to space the items evenly to make for easy readability (See Appendix I).
7. Be sure that the instructions for the test are clear and precise and include examples so as to minimize the possibility of misinterpretation of the test procedures.
8. Take care to see that there is sufficient space between numbered items and answers on the answer sheet so that a student will not become confused marking his answers.
9. There should be no observable pattern to the correct answer positions. This will minimize the probability of guessing.

10. It may be desirable to arrange items from easy to more difficult to have a progressively more difficult exam.
11. Before running the first pretest, submit a copy of the test to the scrutiny of an outside "expert" not involved in the preparation of the test but familiar with the subject matter of the test. This would eliminate the need to add, delete or change items on a second pretest.
12. Prior to each pretest administration, give the test to a group of native English speakers to be proof-read. This will insure that there are neither grammatical nor typographical errors on the test.
13. Select a good representative sample of subjects to take the pretest, and if two pretests are planned, there should be a control group to take each pretest.
14. Use as many of the methods for determining test reliability as feasible. These will clarify the consistency of the test.
15. It is necessary to evaluate the curricular, content, and face validity of the test and try to determine the criterion-related validity by correlating the scores on the test to some established external variable.

BIBLIOGRAPHY

- Best, Jeanelle and Hyin, Donna. 1976. STEL: Structure Tests-English Language. Rowley, Mass.: Newbury House Publishers, Inc.
- Briers, Eugene. 1973. "Cross-Cultural Biases in Language Testing", Focus on the Learner: Pragmatic Perspectives for the Language Teacher. Edited by John W. Oller, Jr. and Jack C. Richards. Rowley, Mass.: Newbury House Publishers, Inc.
- Ebel, Robert L. 1965. Essentials of Educational Measurement. 2d ed. Englewood Cliffs, N.J.: Prentice-Hall, Inc.
- Greene, John A. 1963. Teacher-Made Tests. New York: Harper and Row Publishers.
- Harris, David Payne. 1969. Testing English as a Second Language. New York: McGraw-Hill Book Company.
- Harris, David Payne and Palmer, Leslie A. 1970. A Comprehensive Language Test for Speakers of English as a Second Language. New York: McGraw-Hill Book Company.
- Lyman, Howard B. 1963. Test Scores and What they Mean. 2d ed. Englewood Cliffs, N.J.: Prentice-Hall, Inc.
- Shao, Stephen P. 1974. Mathematics for Management and Finance. 3d ed. Cincinnati, Ohio: South-Western Publishing Co.
- Spaan, Mary. 1972. ELL: English Achievement Series. Ann Arbor, Mich.: University of Michigan Press.
- Spolsky, Bernard. 1972. "Language Testing-The Problems of Validation", in Readings on English as a Second Language. Edited by Kenneth Croft. Cambridge, Mass.: Winthrop Publishers, Inc.
- Upshur, John; Palmer, Leslie; Harris, John; and May, Geraldine. 1962. Michigan Test of English Language Proficiency. Ann Arbor, Mich.: University of Michigan Press.
- Valette, Rebecca M. 1967. Modern Language Testing: A Handbook. New York: Harcourt, Brace and World, Inc.

39. "Did your family sell the cottage?"
"Yes, it's no longer _____."
- ours
 - their
 - our
 - of us
40. "No, that's not yours; it's mine."
"Well, your book looks _____ my book."
- equal to
 - the same as
 - like to
 - the same
41. "I like foreign-made movies."
"_____."
- So I do
 - So do I
 - Too do I
 - Too I do
42. "Who's taller? You or John?"
"I'm taller than _____."
- he
 - him
 - himself
 - his
43. "You are doing well in your history class, aren't you?"
"Yes, _____ I hardly ever go to class."
- in spite of
 - even though
 - in order to
 - on the contrary
44. "What did Jim say?"
"He asked _____."
- what time it was to me
 - me what time it was
46. "I don't like American coffee."
"_____."
- Either do I
 - Neither do I
 - I don't neither
 - I don't too
47. "Nick's talking is distracting."
"I know. I _____ he'd keep quiet for a while."
- hope
 - wish
 - desire
 - want
48. "Do you have any money left?"
"Yes, I _____ have five-hundred pesos."
- already
 - yet
 - still
 - also
49. "Martha plays volleyball very well."
"They say she's _____ on the team."
- the best player
 - better player
 - the most player
 - the better player
50. "What did you do last night?"
"First we went out to dinner; _____, we went to the movies."
- in addition to
 - besides
 - too
 - then
51. "Look at that beautiful car."
"Which one _____ at?"
- you are looking
 - are you look

- c. what time for me it was
- d. me what time was it

- c. are you looking
- d. you looking

45. "That's a beautiful painting."
"Yes, I have _____ seen such a
beautiful piece of art work."

- a. ever
- b. already
- c. never
- d. yet

52. "Did you have a good time at the
party last night?"
"Yes, I really enjoyed _____."

- a. myself
- b. itself
- c. me
- d. self

A. -Items discarded from the first pretest:

	Times chosen	Degree of Difficulty
14. "Were they students at this University last semester?" "No, they _____."		97.96%
a. weren't	48	
b. didn't	0	
c. wasn't	1	
d. hadn't	0	
66. "How does he get everything finished on time?" "He _____ his wife _____ him type up letters."		12.24%
a. has. . .to help	31	
b. has. . .helps	8	
c. has. . .to helping	2	
d. has. . .help	6	

B. -Distractors altered from the first pretest:

	Times chosen	Degree of Difficulty
32. "What time is it?" "Look at the clock _____ the door."		46.97%
a. after*	1	
b. in	8	
c. at	16	
d. over	23	

*the first distractor "after" was changed to "under" for the second pretest. The distractor remained as "under" for one of the final forms of the test.

	Times chosen	Degree of Difficulty
77. "George doesn't want to go with us." "But he said he _____."		51.02%
a. can go	18	
b. could go	25	
c. can have gone*	1	
d. could went	2	

*the third distractor was the only distractor changed. It was changed from "can have gone" to "can have". The distractor remained as "can have" for one of the final forms of the test.

C. -Item altered from the first pretest to be put on the second pretest:

12. "What happened to him?" "His wife _____ a new car."	25. "What's wrong with him?" "His wife _____ their new car."
--	---

- | | |
|--------------------|----------------------|
| a. was just buying | a. wrecked just |
| b. bought just | b. was just wrecking |
| c. just bought | c. just wrecked |
| d. just was buying | d. just was wrecking |

D. -Items added to the second pretest:

"We have a meeting this afternoon."
 "It's _____ five o'clock sharp."

- a. on
 b. at
 c. by
 d. in

"Dinner is ready and waiting on the table."
 "We _____ before the food gets cold."

- a. would eat
 b. should have eaten
 c. would have eaten
 d. should eat

E. -Item added to and then later discarded from the second pretest:

	Times chosen			Degree of Difficulty		
	Forms A	B	C	A	B	C
"What kind of work do you do?" "I'm _____."				86.67%	89.29%	96.43%
a. lawyer	3	3	1			
b. the lawyer	0	0	0			
c. a lawyer	26	25	27			
d. an lawyer	1	0	0			

F. -Distractors that were changed back to the distractors from the first pretest after the second pretest or changed a third time for the final forms of the test:

	Times chosen	Degree of Difficulty
First pretest:		
8. "Mr. Jones is my teacher" " _____ is he _____ ?"		36.65%
a. How. . .like	26	
b. What. . .like	16	
c. Which. . .like	3	
d. That. . .like	2	
Second pretest (Form C):		
9. "Mr. Jones is my teacher" " _____ is he _____ ?"		14.29%
a. How. . .like	22	
b. What. . .like	4	
c. Which. . .like	1	
d. When. . .like	1	

*The last distractor was changed from "That...like" to "When...like" for the second pretest. On the final form (Forms A, B, and C) the item and its dis-

tractors read as follows:

"Mr. Jones is my teacher,"

"_____ is he _____?"

- a. How, .,like
- b. What, .,like
- c. Which, .,like
- d. That, .,like

First pretest

	Times chosen	Degree of Difficulty
91. "Why does Antonio go to Mexico City every weekend?"		59.18%
"_____ visit his girlfriend who lives there."		
a. Thus*	0	
b. Because	17	
c. In order to	29	
d. Since*	0	

Second pretest (Form B)

39. "Why does Antonio go to Mexico City every weekend?"		53.57%
"_____ visit his girlfriend who lives there."		
a. Therefore*	2	
b. Because	11	
c. In order to	15	
d. Because of	0	

*As can be seen the first and last distractors were changed. On the final form of the test a third change was made to the last distractor:

Final form (Form B)

53. "Why does Antonio go to Mexico City every weekend?"	
"_____ visit his girlfriend who lives there."	
a. Therefore	
b. Because	
c. In order to	
d. So	