The development and validation of an institutional reading placement test

by

Rosyati Abdul-Rashid

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF ARTS

Department: English

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

Major Professor: Dr. Dan Douglas

Iowa State University

Ames, Iowa

1996

Graduate College
Iowa State University

This is to certify that the Master's thesis of

Rosyati Abdul-Rashid

has met the thesis requirements of Iowa State University

Signature redacted for privacy

Major Professor

Signature redacted for privacy

For the Major Department

Signature redacted for privacy

For the Graduate College

DEDICATION

To God, the Wise and Worthy of all praise,

Who has given me the ability to complete

this major research of mine.

To my mother, Rugayah,

who always prays for my success.

To a friend dear to my heart, Dr. Janet Hart Heinicke,

who has given me encouragement and moral support

to finish this research study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I. INTRODUCTION

Background of the Study

The Intensive English and Orientation Program (IEOP) at Iowa State University (ISU) was started in 1966. The program, which teaches spoken and written American English, is managed by the College of Liberal Arts and Sciences. Four courses are offered in the IEOP program: grammar, writing, reading development, and communication skills. Within each course, there are five or six levels of instruction, from a novice level to an advanced level.

All students entering IEOP are required to take a placement test at the beginning of the session in which they are enrolled. The test is administered to determine each student's appropriate level of proficiency in five skill areas: listening, grammar, vocabulary, reading, and writing. The results of the test are used to place students into homogeneous classes at their appropriate ability level. This method of grouping students helps make teaching more efficient and learning more effective since students who are grouped in homogeneous ability levels can be taught similar language or learning points (Brown, 1990).

A similar test is given to all IEOP students at the end of each session. There are five sessions per year: two in the spring, one during the summer, and two in the fall. Each session has eight weeks of instruction. The scores obtained from the placement test administered at the end of a session, together with the individual instructor's class assessment of each student, help to guide placement decisions for the following session for continuing students.

The placement test used in IEOP consists of a writing subtest developed internally by the IEOP staff, and four other subtests of the English Placement Test (EPT) developed by the Testing and Certification Division of the English Language Institute at the University of

Michigan. The EPT is specifically developed for use by institutions offering courses in English as a foreign language. This objectively scored test is designed to enable staff to quickly group English as a second language (ESL) students into similar ability levels.

The EPT contains four subtests: listening comprehension, grammar, vocabulary, and reading. There are 100 items in the test: 20 are devoted to listening comprehension; 30 to grammar; another 30 to vocabulary; and 20 to reading. The listening subtest consists of three-option, multiple-choice items while the grammar, vocabulary and reading subtests are composed of four-option, multiple-choice items. The administration time for the complete test is 75 minutes.

The scores obtained in each subtest are converted to a 100 percent scale, and are used to determine a student's appropriate ability level within the four courses offered in the program. Table 1 (source: IEOP coordinator) illustrates the approximate scores on the EPT used by IEOP instructors to place students into different ability level classes.

There are three forms of the EPT that are used in IEOP: A, B, and C. The forms are used alternately to ensure that students do not use the same form within a session or within two consecutive sessions.

Table 1. Minimum scores used to determine placement into IEOP levels

Level		EPT (%)
1	(Beginning)	0 - 25
2	(Intermediate A)	25 - 40
3	(Intermediate B)	40 - 55
4	(Intermediate C)	55 - 70
5	(High)	70-85
6	(Advanced)	85 - 100

The EPT has been used in IEOP for 25 years. The test is indeed an old one. In fact, all three EPT forms used in IEOP were developed in the 1970s. Most of the IEOP faculty strongly feel that the EPT should be replaced with a new placement test. However, it is difficult to find an alternative standardized test that suits the IEOP's placement purposes. What the IEOP faculty can do to resolve this problem is to develop their own test which will be used solely in the program. This internally developed test will be a better placement instrument because such a test is usually more specifically related to the content of the curriculum (Brown, 1990).

However, designing a new placement test is not an easy task because it requires trained and experienced personnel, time, and money (Angelis, 1990). Given these constraints, it is certainly not practical to create new subtests for the EPT all at once. The most practical course of action in this situation is to gradually develop the new subtests. As a starting point, efforts should be taken to develop a new test for one of the four subtests in the EPT. A decision must be made as to which subtest should be replaced first. This decision can be made by determining which among the four subtests in the EPT is considered to be the most problematic placement instrument.

Statement of the Problem

Among the four subtests in the EPT currently used in IEOP, the reading subtest has been found to be the subtest most urgently in need of revision or replacement. There are four reasons as to why this subtest is considered to be an inadequate instrument for placing IEOP students in the appropriate reading class. First, the reading subtest is an old test (i.e., it was developed in the 1970s) and could, therefore, be based on an old or outdated theory of

reading. Second, the construct of the subtest is not known. Third, the subtest is not specifically designed for IEOP students and thus may not be a suitable placement instrument for the range of abilities found in IEOP (Brown, 1984a, 1987, 1990). Finally, the reading subtest is not related in content to the reading course offered in IEOP.

A serious mismatch between what is tested by the EPT reading subtest and what is taught in IEOP reading classes can be discovered through an examination of the format of the EPT reading subtest. The following example illustrates the general form of the items used in the EPT reading subtest:

John drove me to Eleanor's house.
Who drove?

A. I did.
B. John did.
C. John and I did.
D. Eleanor did.

The above example illustrates that the EPT reading subtest consists of sentence-level reading items. Such items are not a representative sample of items used in IEOP reading classes Typical items used by IEOP instructors consist of passages followed by comprehension questions. Therefore, a new reading comprehension subtest consisted of authentic passages needs to be developed to replace the old reading subtest so that placement processes in IEOP can operate more efficiently and effectively.

Another problematic aspect of the EPT is that the test separates vocabulary items from reading comprehension items. That is, in the EPT, vocabulary and reading comprehension items are tested in two separate subtests. Like the reading subtest, the vocabulary subtest contains sentence-level items. The following is an example of vocabulary items tested in the EPT:

I can't _____ you his name, because I don't know it.

A. talk
B. say
C. speak
D. tell

Most taxonomies of reading list comprehension of vocabulary items as one of the important reading skills. This explains why most current tests on reading include both reading comprehension items and vocabulary items. In these tests, vocabulary items are tested in a contextualized way. In other words, the vocabulary items are listed based on a passage and not on a sentence.

In this study, the researcher attempted to develop a more integrative and communicative reading test. In the developed test, reading comprehension items and vocabulary items are tested based on selected passages.

Purpose of the Study

The purpose of this study was two-fold. First, the study was undertaken to develop a new reading comprehension subtest specifically designed to replace the old reading subtest and the vocabulary subtest in the English Placement Test (EPT) currently used in the Intensive English and Orientation Program (IEOP) at Iowa State University (ISU). The second purpose was to validate and evaluate the newly developed reading comprehension subtest. The primary focus of the study was to examine the procedures followed in the development of the new reading comprehension subtest and find justifications for using its scores to make placement decisions.

Research Questions

This study was designed to answer the following research questions:

1. What are the item statistics (i.e., item facility and item discrimination indices) for the pilot and final versions of the new reading comprehension subtest?
2. What are the descriptive statistics for both versions (i.e., pilot and final draft) of the subtest?
3. Is the new reading comprehension subtest valid? More specifically, is there any evidence related to the following types of validity?
   a. *Content validity* - Do the items in the new subtest represent the types of items offered in the IEOP reading course?
   b. *Construct validity* - Does the new subtest measure the constructs on which it is based?
   c. *Criterion-related validity* - Do the respondents' performances on the new subtest match other measures of their abilities?

Assumptions of the Study

The basic assumptions of this study were:

1. The new reading comprehension subtest is a valid measurement device for assessing the reading abilities of IEOP students.
2. Students who participate in this study are representative samples of the population for whom the new subtest is designed.
3. Students who participate in the study will give honest responses.
4. The committee responsible for the development of the new reading subtest is comprised of experienced and trained ESL instructors.

5. The feedback obtained from the IEOP reading instructors is valid.

Delimitations of the Study

The study was subjected to the following limitations:

1. The study was limited to the IEOP students and English 101B students enrolled during spring semester, 1996, at Iowa State University (ISU). The students in English 101B are ISU students who either failed or scored very low in the ISU placement test.
2. The language skill under investigation was limited to reading.

Procedures of the Study

The study followed the procedures as listed:

1. Formulate the problem to be studied.
2. Review the related literature.
3. Identify the population for the study.
4. Develop a new reading subtest as an instrument to be used for gathering data for the study.
5. Gather data.
6. Analyze the data in inferential and descriptive terms.
7. Interpret the results of the analyses.
8. Make conclusions.
9. Outline the implications of the study.
10. Give recommendations for future research.

CHAPTER II. REVIEW OF LITERATURE

Introduction

The central purpose of this study was to develop and validate a new reading placement subtest for the Intensive English and Orientation Program (IEOP) at Iowa State University. The new reading subtest developed in the study was designed to assess the reading ability of IEOP students. This chapter presents a review of the theoretical, pedagogical and research work related to the present study. The review covers four major areas: (a) Test Development Process; (b) Definition of Test Construct; (c) Test Methods; and (d) Test Validation Procedures.

Test Development Process

Brown (1983) defined a test as ". . . a systematic procedure for measuring a sample of behavior" (p. 8). The behavior referred to in this case is the test takers' responses to test items. Inferences of the test takers' characteristics are then made based on the numerically scored responses. Examples of characteristics inferred are intelligence, achievement, attitude, and reading comprehension.

Inferences made from test scores will only be valid if the behaviors exhibited by the test takers on a test designed to measure a particular characteristic or construct adequately reflect the construct. To ensure that the behaviors elicited from a test are a true indication of the ability or construct being measured, the items used in the test must be a representative sample of the relevant content domain. Such items can be appropriately and systematically constructed based on the test's specifications.

A test is "a systematic procedure" because the entire process of test development

comprising item construction, test administration, scoring and interpretation of scores is based on a specific set of prescribed rules or procedures (Bachman & Palmer, in press; Brown, 1983). The procedures followed in the development of a test vary from simple to highly complex ones depending on the situation (i.e., depending on what the test is intended for and who is taking it). However, most test development processes include three major steps proposed by Bachman and Palmer (in press): specification, operationalization, and administration. These steps correspond closely to those suggested by Brown (1983) (see Figure 1).

Test specifications

A test's specifications are the blueprint that assists test constructors in developing a good and useful test. The blueprint consists of a detailed description of the general parameters for the design of a test. Some of the parameters outlined by test specialists (Alderson, 1995; Bachman & Palmer, in press; Brown, 1983) are specifications of the purpose of the test and the nature of the construct the test is intended to measure, the descriptions of the target language use situation and the characteristics of the test takers, the listing of the target language use tasks, and the decision of the test method to be adopted.

Specification of test purpose

Many language testing specialists (e.g., Alderson, 1995; Alderson, Krahnke & Stanfield, 1987) have categorized language tests into several different categories or types based on purposes of the tests. However, Brown (1990) argued that what most language testing specialists have categorized are not the different types of test but the different uses made of test scores. A clearer explanation concerning the test types and purposes is provided

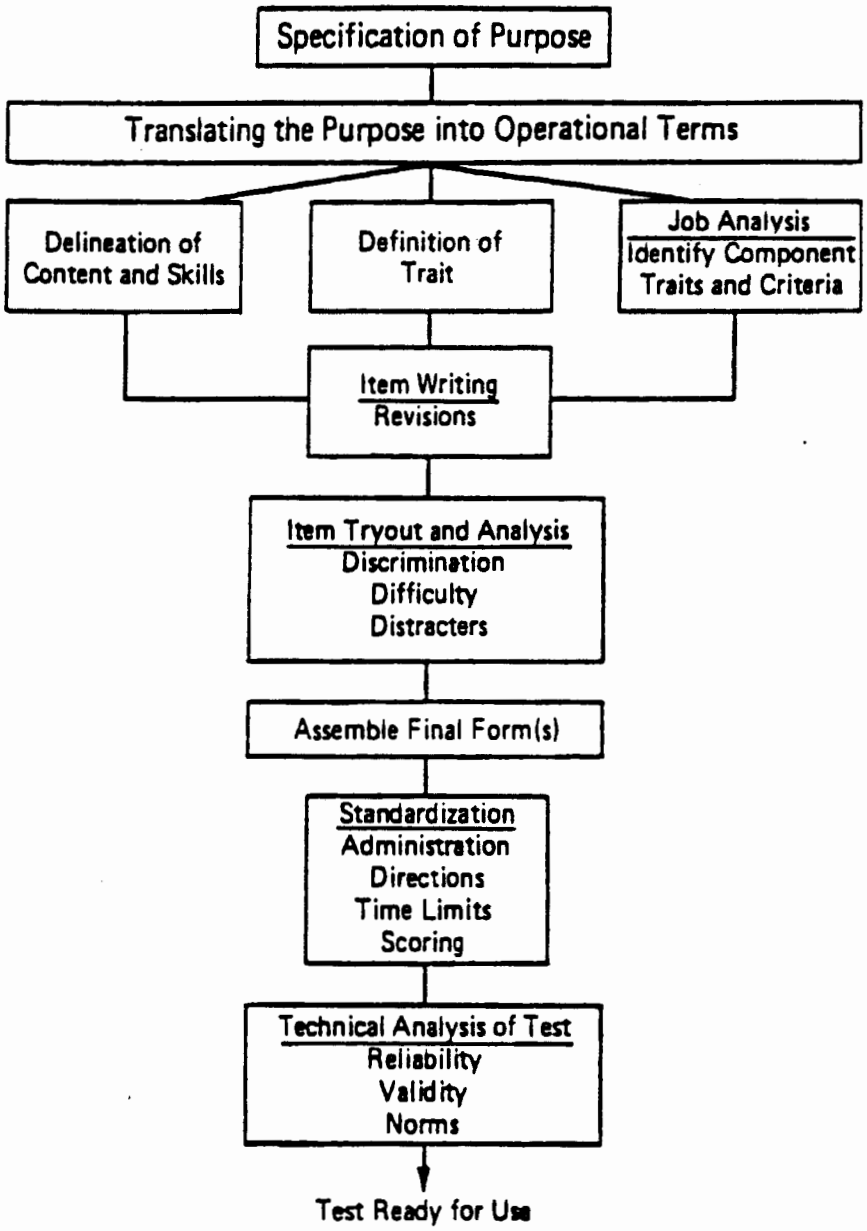

Figure 1. The steps in test construction (Brown, 1983)

by the following discussion on the differences between norm-referenced and criterion-referenced tests.

Norm-referenced versus criterion-referenced tests. According to Brown (1990), language tests are designed to serve one of two purposes: to provide information about a test takers' performance on a test in relation to the performance of other test takers, or to compare the individual test taker's performance against a criterion or a set standard. In other words, all language tests ". . . are either norm-referenced tests (NRTs) or criterion-referenced" (p. 28). The NRTs differ from the CRTs in two main ways: in design and in the interpretation of scores.

The NRTs are designed to measure general language abilities such as English language proficiency and reading comprehension. The purpose of an NRT is to distinguish test takers of different ability levels. The distinction is based on the comparison made between the scores of an individual test taker and the scores of others who sat for the same test. For example, if a test taker scores in the 75th percentile on an NRT, the student can then be said to have performed better than 75 out of 100 test takers who took the test. Two examples of NRTs are the Test of English as a Foreign Language (TOEFL) and a placement test.

The CRTs, in contrast, are designed ". . . to measure well-defined and fairly specific instructional objectives" (Brown, 1989, p. 68). The purpose of a CRT is to determine the extent to which takers have mastered a specified course content or ". . . have developed knowledge on a set of objectives" (p. 68). The level of mastery attained is measured based on the percent of items scored on a CRT. Therefore, if a test taker obtains a score of 75 percent, the test taker can then be claimed to know or master 75 percent of the items tested in the test.

An example of a CRT is a final examination given to students at the end of a course.

In addition to being different in design and score interpretation, the NRT and CRT are also different from one another in terms of score distribution and the test takers' knowledge of test items. Table 2 (Brown, 1989, p. 69) provides a summary of the differences between the NRT and CRT.

Table 2. Differences between norm-referenced and criterion-referenced tests

Characteristic	Norm-referenced	Criterion-referenced
Type of measurement	•General language abilities or proficiencies are measured.	•Specific objectives-based language points are measured.
Type of interpretation	•Relative: A student's performance is compared with that of all other students	•Absolute: A student's performance is compared only with a prespecified learning objective.
Score distribution	•There is a normal distribution of scores around a mean.	•If all students know all of the material, all should score 100%.
Purpose of testing	•Students are spread out along a continuum of general abilities or proficiencies.	•The amount of material known or learned by each student is assessed.
Knowledge of questions	•Students have little or no idea what content to expect in the questions.	•Students know exactly what content to expect in test questions.

Source: Brown, J. D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly, 23*(1), 69.

Uses of scores on NRT and CRT tests. The inferences made from test scores can assist test users in making decisions. The four basic decision types are: proficiency, placement, diagnostic, and achievement. The test users can use the scores on an NRT to help them make proficiency and placement decisions, whereas those on a CRT are used to guide them to make diagnostic and achievement decisions.

Proficiency decisions. A proficiency decision is usually made to ensure that individuals have met the standards set for entrance to or exit from a language program or institution. The focus of a proficiency decision is on the individuals' ". . . general levels of language" (Brown, 1990, p. 15) that can be inferred from scores on proficiency tests designed to measure general skills. One example of proficiency tests is the TOEFL, which is used by most institutions in America to determine the English language proficiency of international students prior to their admission to the relevant institutions.

Placement decisions. Test scores can help test users decide how to appropriately place students in classes or courses from which they would benefit most. The distribution of scores of an NRT can guide test users to distinguish students of different ability levels in a specific skill area such as reading. The ability level of an individual student is determined based on a comparison of the student's test scores with the scores of others (i.e., the norm group) who took the same test. Students who are ranked (in ability level) according to their test scores can then be placed in appropriate ability level classes. For example, the students with the top ten scores in a reading test can be put into the advanced reading class.

Placement decisions are based on placement tests designed to assess the individuals' ability levels in some specific skill areas. Thus, placement tests are usually administered at the beginning of a program or course.

Diagnostic decisions. Test scores are also used to identify the specific areas in which individual learners have strengths and weaknesses. The information gained from scores on a diagnostic test enables language learners and instructors to focus their efforts on

areas that can foster achievement. For example, the identification of learners' strengths can lead to efforts taken to further promote the strengths. Conversely, the identification of learners' weaknesses focuses the learners' and instructors' efforts to finding ways to overcome the weaknesses.

Since a highly specific diagnostic test is difficult to design, many test users resort to using proficiency and achievement tests for diagnostic purposes (Alderson, 1995). Diagnostic tests are usually given at the beginning and in the middle of a course. This enables the learners to note their strengths as well as to make adjustments to overcome their weaknesses.

Achievement decisions. Individuals' scores on a language test may reflect their degrees of achievement in language learning. That is, the scores may indicate how much language the individuals have learned. The results obtained from an achievement test designed to assess individuals' mastery of language skills specified in the course objectives can help test users make decisions pertaining to matters such as the award of certificates or diplomas and the granting of permission for continuing study at a higher level.

Achievement tests are administered at the end of a course. The items on the tests usually cover the language areas or skills taught throughout the course.

Specifications of test construct, target language use situation and tasks

Once the decisions on the uses of test scores are made explicit, test constructors must provide a specification of the construct or the characteristic of the ability to be measured in the test under construction. In addition to the construct specification, the domain within which the inferences about test takers' abilities are to be made must also be specified. The

specification of the domain should be complemented with a specification of the tasks to be carried out in the domain.

Specification of test takers' characteristics

The test constructors must also write specifications of the test takers' characteristics. These include the descriptions of the test takers' personal characteristics such as age, gender, religion and nationality, and their educational background. These descriptions may help the test constructors to write appropriate test items. For example, knowledge of test takers' religions and nationalities can prevent test constructors from writing items that have religious or cultural bias.

Operationalization

Operationalization involves translating test specifications into operational terms. In other words, test specifications are used as guidelines to follow in constructing a test. The specifications help test constructors decide on the content of the test, the types of items to include, the length of the test, and the measures to use for scoring the test.

Test administration

A test is administered to test takers either for the purpose of trying out the test items or for the purpose of using the test scores to help make some specific decisions. Some tests are administered for both of the purposes discussed but many tests are given to test takers without even being pre-tested. Alderson (1995), who carried out a survey of EFL examination boards in Britain, found that six out of the twelve boards that responded to questionnaires did not follow any pre-testing procedures. In addition, out of the six boards

that pretested their test items, only three conducted statistical analyses of the pretesting results.

Pretesting and item analysis

The term pretesting refers to the administration of a test before the final or actual use of the test. A pretest or pilot test is usually given to a group of test takers similar in background and level to those for whom the test is designed. The pilot administration can provide test users with some useful information on the quality of test items and administration procedures.

Although a relatively good test with many potentially good items can be developed by a group of experienced and well-trained test developers who follow strict editing procedures, some problems in the test can only be identified by empirical item analyses. Moreover, research by Alderson (1993) on the judgments of language professionals in language testing produced evidence indicating that even experienced language examiners ". . . were unable to predict with any degree of accuracy the difficulty of test items" (Douglas & Chapelle, 1993, p. 56). This finding establishes the need to pretest all test items. For this reason, the researcher of the present study decided to adopt pretesting procedures in the development of a new reading test.

Analysis of the pretest items may help to provide information on two important characteristics of a test item: the item's level of difficulty and its discriminatory power, the extent to which the item can discriminate or distinguish between the more able test takers and the less able ones. The item difficulty and discriminatory indices calculated on the results of the pretest can help test constructors to revise the pretest items. The revision will ensure that

the items used in the final form of a test are effective items.

Pretesting also enables test constructors to check on and improve testing conditions such as the instructions given in the test, the time limit set for the test completion, the physical setting, and the psychological climate during testing. Nevertheless, not all test users will be able to do this because pretesting of items is not always possible and statistical analysis cannot be carried out on the items of some tests such as a writing test.

Analysis of a test

The efficacy of a completed test (i.e., the test in its final form) as a measurement device must be determined before the test is recommended for future use. The test's effectiveness can be examined by performing analyses on the test. Statistical data gathered from the analyses may serve as evidence of the test's reliability (i.e., that the test measures consistently) and validity (i.e., that the test measures what it is intended to measure). A detailed illustration on test analysis procedures is provided in the discussion on test validation.

Defining the Construct of the Reading Test

Research on reading English as a second language (L2) began to flourish in the 1970s. Most of the L2 reading research drew heavily from first language reading research (Grabe, 1991). The three approaches to reading adopted by many L2 researchers are bottom-up, top-down and interactive approaches. The bottom-up approach emphasizes the readers' ability to make use of lower-level processes (letter and word recognition) to comprehend a text (Goodman, 1970). The top-down approach, on the other hand, places more emphasis on the ability to rely on the higher-level processes (reliance on context and prior knowledge).

Schema theory is based on top-down processing. According to this theory, readers comprehend a text by bringing in information, knowledge, emotion, experience, and culture to the printed text (Clarke & Silverstein, 1977). The third approach that an L2 researcher may use is an interactive approach which is actually a synthesis of the bottom-up and top-down approaches. Most current research on L2 reading has adopted the interactive approach.

Many interactive models of reading have been proposed. One of them is the interactive-compensatory model by Stanovich (1980). The assumption underlying Stanovich's model is that reading involves interactions of high and low level processes. According to the model, a reader who is weak at a particular reading level process will try to compensate for the weakness by relying on another process. For example, a reader who is weak at recognizing a particular word will resort to the context of the sentence or paragraph in which the word resides to understand the meaning of the text.

Eskey (1986) referred to the processes discussed by Stanovich as processes for interpreting meanings and identifying forms. Eskey argued that second language readers can comprehend a text fully only if they master two main categories of knowledge: form and substance (see Figure 2). Eskey's framework of "categories of knowledge crucial to reading" corresponds closely to Bachman's (1990) framework of communicative language ability. Using the frameworks of these two language testing specialists and the often cited taxonomies of reading by Barrett (1968) and Davis (1968), the present researcher formulated a model of reading comprehension to serve as the construct for the reading test developed in this study. The model is based on the interactive approach to L2 reading and focused on three abilities:

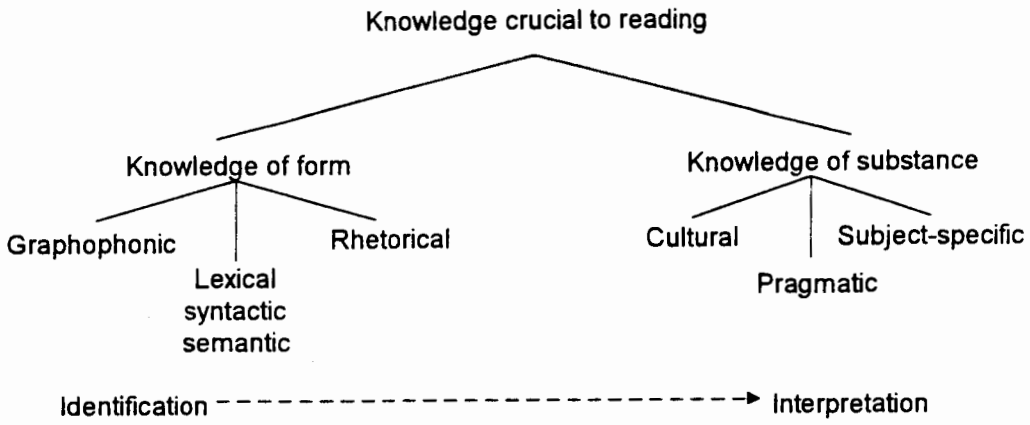

Figure 2. Categories of knowledge crucial to reading (Eskey & Grabe, 1986, p. 18)

1. ability to comprehend a passage;
2. ability to comprehend the meanings and use of vocabulary items in a passage; and
3. ability to infer from a passage.

The ability to comprehend a passage includes several other "sub-abilities" such as the ability to identify the main and supportive ideas, and the ability to identify anaphoric reference. William et al. (1989), who wrote a summary of current issues in reading pedagogy and research, stated that the "sub-abilities" discussed were considered important by many material writers.

The ability to comprehend the meaning of words is another important component of reading comprehension. Language testing researchers such as Alderson and Urquhart (1984a) and Singer (1981) recognized ". . . the need for extensive vocabulary for reading" (cited in Carrell et al., 1988, p. 59).

Several studies in L2 reading have found a strong relationship between vocabulary and reading ability. For example, Barnet (1986), Strother and Ulijin (1987), and Lewis (1987) found vocabulary to be an important predictor of reading ability. In addition, Laufer (1992) found a high correlation between lexical level in a second language and second language reading ability.

Another important component of reading comprehension is inferencing. Making inferences refers to the ability to use information that is not explicitly stated in a reading passage to answer questions based on the passage. Inferencing is listed as one of the important reading skills in most reading taxonomies. A great deal of evidence exists to support including inferencing as a part of the model of reading comprehension. Olen (1985) discovered that good readers have a better ability to answer questions than the poor readers. A similar finding was found by Davey and Macready (1985) and Singer (1988).

Test Methods

When developing a good test, the test constructors must ascertain that the method or format employed in the test can produce statistically sound results and has a minimum impact on the test scores. The validity of a test can be seriously undermined if the test scores are highly influenced by the test's format. Such significantly affected test scores cannot be interpreted as true measures of the test takers' knowledge of the trait the test is purported to measure.

Considerable research in the field of language testing has shown that testing methods can affect test scores and, thus, the general performance of the tests under consideration (Alderson, 1980; Allan, 1992; Bachman, 1985; Shohamy, 1984). For example, a study by

Shohamy (1984) on the impact of the multiple-choice format versus open-ended format on the trait of reading comprehension produced results indicating that different testing methods affect test-takers' scores in different ways, with the low level test takers being the ones most affected by the methods.

Therefore, the task of the researcher in the current study was to find a method that can most accurately assess the trait of reading comprehension. The researcher carried out the task by reviewing research and theoretical works on the strengths and weaknesses of four methods most commonly adopted for a test administered to a large test population: cloze test, open-ended questions, true or false item-response format, and multiple-choice format.

Cloze

The cloze test has been extensively used as a reading assessment instrument in both first and second languages. The popularity of the cloze test among language testers is attributed to its two characteristics: ease of construction and scoring. The cloze test is relatively easy to construct because it only requires selection of texts and systematic deletions of words from the texts. The cloze test items do not need to be pre-tested or revised (Shohamy, 1984).

However, none of the validity research on the cloze managed to produce evidence on the construct measured by the cloze. In other words, what the cloze may actually measure is not yet known. This is probably the reason why the cloze procedure is not popularly used in the reading tests in the 1990s (Cohen, 1994).

Open-ended questions

The open-ended (OE) testing format is the traditional technique used by the British for assessing literacy skills. The format has been very popular in Britain and British colonies. The OE procedure requires respondents to write their responses to reading comprehension tasks. Therefore, this format assesses both reading and writing skills.

Nuttall (1982) listed three advantages of OE procedures. First, OE items are fairly easy to construct because the test writers do not have to create plausible distractors. Second, the procedure can be used for many purposes. That is, it allows test constructors to assess many different reading skills. Third, it forces the respondents to figure out their responses by directly coming to terms with the texts.

The major disadvantage of the OE procedure is that OE items cannot be easily and objectively scored. The scoring of responses is time consuming and may require the involvement of many graders. The scoring procedure may also cause a disagreement among the graders regarding the correct answers and the scores to be assigned to the responses. The differences in scores assigned by test graders can also result in a low reliability for the test that employed the OE procedure. For these reasons, the OE procedure was not chosen to be the format for the test developed in the current study.

True/False item response

A test adopting a true/false item-response format calls for responses that indicate whether the statements presented in the test are true or false. This test format is usually used to assess knowledge of factual information. However, the format is also recommended for measuring higher mental processes such as comprehension, application and problem solving.

The main advantage of a T/F format, as noted by Heaton (1990), is that it lends itself to items that are quick and easy to construct. The format also allows many items to be generated from the same reading text. The T/F format is suitable for a test to be administered to a large population because scoring can be done rapidly, reliably, and objectively.

On the other hand, the T/F also has many shortcomings. For example, there is a high tendency for test takers to cheat and guess on a test that adopts the T/F item-response format. Random guessing can produce a score of 50 percent correct and, thus, limit the scores' range. Based on the limitations of the T/F format, the researcher decided not to adopt this format for the placement test developed in this study.

Multiple choice

The multiple-choice (MC) format is one of the most widely used of the objective test formats. An MC item is made up of a stem and a set of alternatives. The stem usually consists of an incomplete statement or a comprehension question based on a given text. The alternatives or options are the statements from among which the test takers are asked to select the correct or best answer. The alternatives that are considered as incorrect answers are called distractors.

The basic format of a MC test consists of four alternatives. However, some MC tests may use three or five alternatives instead. Cohen (1994) recommended the use of four alternatives over three, arguing that the employment of the former decreases the percentage of getting an item right by chance more than the latter (25% vs. 33%).

Among the available testing methods, MC is the most common format used in American standardized reading tests. On the other hand, MC is also the most controversial

format. Hughes (1989) and Weir (1990) provided a long list of the limitations of MC items, among which are the difficulty of constructing the items, and the facilitation of cheating and guessing.

Nevertheless, advocates of the MC format defend its effectiveness and supremacy over others by stressing the "technical superiority of recognition items" (Brown, 1983, p. 238) in MC tests. Technically, the quality of MC items can be monitored through the statistical analyses of the items (Peirce, 1994). Using item analysis enables test constructors to determine the difficulty and discriminatory power of the MC items. In addition, it also directs the attention of the test constructors to potentially problematic items.

In addition to their "technical superiority", MC items have also been claimed to have the ability to assess higher order skills such as analysis, application, and evaluation (Green, 1975; Harrison, 1983; Marshall, 1971). Nuttall (1982), for instance, acknowledged that MC items, if carefully designed, can serve as a "highly effective instrument for training interpretive skills" (p. 126).

Criticism made about the MC format may have been based on the MC tests that were misused or poorly constructed. Good and effective MC items can, in fact, be written if proper item construction procedures are followed and the relevant item statistics procedure is employed. Guidance in developing MC items is offered in many language testing textbooks (e.g., Harris, 1961; Heaton, 1975; Murphy, 1969).

What attracts most test constructors to use the MC format is the fact that MC items can be easily and objectively scored. This characteristic of the MC format makes it the most suitable format for placement tests that are administered to a large group of respondents. In

view of the many strong points of the MC format, particularly the "technical superiority" of the MC items, the researcher decided to adopt this format for the placement test created in this study.

Test Validation

A validation study of a test involves examination of the validity of the test. Validity is a term that refers to the extent to which a test measures what it is intended to measure (Henning, 1987). Evidence of the validity of a test is needed to support the claim that the inferences made from the scores obtained in the test are reliable or justified.

An example of the inferences made from the test scores is that the scores are a true indication of the test takers' abilities in the construct (to be explained later in this chapter) the test is claimed to be measuring. Based on this inference, further decisions such as placement or diagnosis can be made about the test takers. This implies that validity is very much dependent on the uses made of the test scores.

Since different test users may use some test scores for different purposes (e.g., diagnostic vs. placement purposes) and different test populations produce different scores, what is valid in one test may not necessarily be valid for another (Alderson, 1994); Henning, 1987). For this reason, the validity of a test needs to be constantly and continually examined.

To ensure that a test is appropriately used for the purpose intended, test constructors or users need to establish and demonstrate the validity of the test concerned. This is usually carried out through theoretical and empirical studies that involve the formation of some theoretical frameworks from which some testable hypotheses can be deduced, and the collection of data that can be used to substantiate the generated hypotheses (Brown, 1983).

The procedures for determining the validity of a test can be categorized under three main headings: content validity, criterion-related validity, and construct validity.

Content validity

Content validity is a form of test validation that involves determining the extent to which the content of the test items represents the content domain of interest. For example, if the relevant content domain is reading comprehension, then the test items must serve as samples of knowledge or skills in reading.

The process of content validation usually involves having experts (i.e., people who are knowledgeable in the subject area covered by the test) systematically compare the test items with the subject matter content. If the items are judged by the experts as representative samples of the domain, the test can then be said to be content valid. Judgments of experts can be gathered through interviews (Wall, Clapham & Alderson, 1994), questionnaires or rating scales (Alderson & Lukmani, 1989; Bachman, Kunnan, Vanniariajan & Lynch, 1988; Clapham, 1992) or test review process in which the test items are systematically reviewed by the experts (Peirce, 1994).

Criterion-related validity

Criterion-related (CR) validity refers to how well the scores obtained from the test to be validated are empirically related to the scores on another test that serves as an external criterion measure. The criterion test employed is usually a recognized or a valued measure (i.e., a highly valid test). Evidence of CR validity is achieved in the form of a correlation coefficient computed between the scores obtained from the two tests compared. If the

resultant coefficient is significantly high, the test under validation can then be claimed to be as valid as the criterion test.

If the tests being compared are administered at approximately the same time, the evidence of validity obtained is called concurrent validity. However, if the scores of the test to be validated are compared to those of a test that can predict the test takers' future performance, the validity evidence gathered is termed as predictive validity. In addition to comparing the scores of two tests, CR validity of a test can also be investigated by comparing the test scores with other measures of the test takers' abilities such as their teachers' ratings of their performance (Alderson, 1994).

Construct validity

Construct validity refers to the extent to which a test measures the trait or construct upon which it is based. Ebel and Frisbie (1991) offered an explanation of what the process of construct validation entails:

> Construct validation is the process of gathering evidence to support the contention that a given test indeed measures the psychological construct the makers intend it to measure. The goal is to determine the meaning of scores from the test, to assure that the scores mean what we expect them to mean. (p. 108)

This explanation implies that test scores are used by test constructors to make inferences about test takers' abilities in the domain of interest (e.g., reading) and to examine the potential of the test developed as a measure of the trait under consideration. However, prior to making the inferences, the test constructors must define the theory of the trait, or construct, underlying the test. This is followed by the formation of hypotheses about the expected

behavior of the test takers and the general performance of the test. The hypotheses may then be tested by comparing test scores with the theory of the test.

Since test scores are viewed to be important in determining the construct validity of a test, the scores must then be carefully and thoroughly examined to ensure that they are a meaningful measure of the construct of a test. This can be done through the accumulation of data that "cast light on the meaning of test scores" (Brown, 1983, p. 138). The data can be gathered in a variety of ways, of which three are analyses of the internal structure of the test, correlational studies, and experimental studies.

Internal structure analysis

The nature of the construct measured by a test can be defined through the analyses of the test content and of the relationships between test items. An examination of the internal structure of a test can provide information about the performance of test items in relation to one another, the items' homogeneity, and stability.

The performance and potential of items are indicated by their discriminatory and difficulty indices while the internal consistency of the items is revealed by the reliability coefficient. The indices and coefficient can serve as construct validity evidence. Freedle and Kosten (1993), for example, used predictions of item difficulty of the reading subtest in the TOEFL to argue for the construct validity of the subtest. Harness (1995), on the other hand, used the level of internal consistency of a test (i.e., reliability coefficient of the test) as evidence for the construct validity of the reading placement test developed at Iowa State University.

Correlational studies

Construct validity of a test can be established by correlating the scores on the test being validated with the scores on another test presumed to measure the same construct. A high correlation indicates that the two tests compared are measures of a similar construct. The correlation also provides justification for interpreting the scores on two tests in the same way.

Correlational studies can also be used to establish the fact that tests measuring different constructs are not highly intercorrelated. In evaluating a placement test at the University of Lancaster, Wall et al. (1994) carried out correlations among the subtests in the placement test and between the placement test and other tests taken by the international students prior to their admission at Lancaster. From the intercorrelational study, the researchers found that each of the subtests was tapping into a different construct. However, due to truncated samples, the researchers failed to obtain a high correlation between the scores on the placement test and the scores on other external measures.

Experimental studies

Evidence of construct validity can be gathered through experimental studies in which the researcher can manipulate the variables presumed to affect test scores. The most popular type of experimental studies conducted by language testing researchers is one that involves giving the same language test to native speakers (acting as the control group) and non-native speakers (representing the target group). Argoff and Sharon (1970) obtained evidence of construct validity of the TOEFL by comparing the performance of American students with that of foreign students on the test.

The New Concept of Validity

The new concept of validity was advocated by Messick (1989) who defined validity as ". . . an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores . . ." (p. 13). Thus, Messick rejected the traditional concept of validity which distinguished the different forms of validity evidence (i.e., content validity, criterion-related validity, and construct validity) by classifying them into three different types, each of which can be sufficiently used to justify specific testing purposes.

According to Messick, the various forms of validity evidence should be used integratively to support the interpretations of scores and use of a test as proposed by test constructors or users. The unifying force of the new unitary concept of validity is construct validity which subsumes content validity and criterion-related validity. In other words, an examination of construct validity evidence will lead to an examination of evidence related to content and criterion-related validities because all these different forms of validity evidence are "complements to one another" (Messick, 1989, p. 16).

Current research on test validation, including the present study, is highly influenced by the new view of validity. In such research, evidence of validity is gathered from multiple sources. In this study, the researcher attempted to gather three forms of validity evidence: content validity, criterion-related validity, and construct validity. Content validity of the newly developed reading subtest was examined by determining how well the items in the subtest represent the content of the IEOP reading class. The empirical item analysis was carried out to investigate the effectiveness of the test items. Correlational studies were

conducted to gather construct and criterion-related validity. That is, the studies were carried out to examine the extent to which the new reading subtest measured what is was designed to measure and to investigate whether the respondents' scores on the new subtest matched their scores on other measures of their abilities (i.e., the respondents' scores on the TOEFL and their instructors' class assessment of their reading abilities).

CHAPTER III. METHODOLOGY

Introduction

The central purpose of the study was to develop and validate a new reading comprehension subtest for the Intensive English and Orientation Program (IEOP) at Iowa State University. The focus of the study was on investigating the procedures for developing a good reading subtest that could serve as an accurate placement measurement device. Emphasis was also placed on gathering evidence to support the validity of the newly developed reading subtest.

This chapter provides a detailed description of the subject, instrument, data collection procedures, and analyses conducted on the data. This chapter includes the following subsections: (a) Population of the Study; (b) Development of the Instrument; and (c) Procedure.

Population of the Study

The subjects for this research were 58 Intensive English and Orientation Program (IEOP) students and 26 undergraduates and graduates enrolled in the spring semester, 1996, at Iowa State University. All the IEOP students were international students who were attending reading classes in the program. Ten of the students were new students who had just enrolled in their first session of studying English while the other 48 students had already had formal instruction in English in the IEOP classes for more than one session.

The 48 IEOP students were placed in the reading classes at five different proficiency levels: 13 students were in Intermediate A; 13 in Intermediate B; 10 in Intermediate C; 7 in

High; and 5 in Advanced. On the other hand, the 26 ISU students were in their 101B writing classes. These students had either failed or received low scores in the writing subtest of the ISU Placement Test and were required to take an intensive course in writing. The level of proficiency of these students equaled the advanced level of the IEOP students.

Instrument Development

In order to assess the reading abilities of the subjects, a new reading comprehension test was developed by a test development committee consisted of the researcher, two TESL professors from the English Department at Iowa State University (ISU), an IEOP instructor, and an ESL instructor who used to teach in the IEOP.

The new reading comprehension subtest contains five short reading passages. Each passage is approximately four paragraphs long. The passages were selected in lieu of discrete sentences (i.e., the format of the reading subtest currently used in the IEOP) because they were considered more representative of texts encountered by IEOP students in their reading classes.

The five reading passages were selected from among the 24 passages that were drawn from several ESL reading textbooks. The passages were chosen based on three criteria: (a) the levels of difficulty of the passages; (b) the length of the passages; and (c) the content of the passages. The passages, as judged by the test development committee, span a variety of levels of difficulty, from a passage for beginners to one for advanced learners. Passages that were judged to be politically, sexually, religiously or culturally offensive were not included in the test. The rationale for excluding these potentially offensive passages is that test-takers should be spared from unnecessary anxiety which could affect their performance on the subtest.

The new reading subtest has a four-option multiple-choice format. The format was selected because it lends itself well to the measurement of a variety of reading abilities, some of which constitute the construct of the new reading test. The format also enables a large number of items to be administered within a short period of testing time. This particular aspect of the format can help to increase both reliability and validity of the new subtest. Moreover, the format is also viewed to be most suitable and practical for a test administered to a large number of students (i.e., the multiple-choice format is easy to administer and score).

Procedure

Test development

The test development committee met three times during the spring semester in 1996. During the first meeting, the committee selected 5 passages from the 24 passages that were extracted from several ESL reading textbooks. The selection was based on three criteria: (a) the levels of proficiency for which the passages were intended; (b) the content suitability of the passages for the target group, the test-takers (i.e., IEOP students); and (c) the length of the passages.

The selected passages were then distributed equally among the five test developers on the committee for item development. All of the test developers agreed that the test items to be developed for the passages assigned to them should be based on the construct of the new reading test being developed. That is, the test developers should create items that assess the following abilities:

1. Ability to comprehend the passage;

2. Ability to infer from the passage; and

3. Ability to understand the meaning(s) of the particular word(s) or phrase(s) in the passage.

One week after the first meeting, the test developers met a second time to discuss the items they had developed. In this meeting, the test developers had a lengthy discussion on the items to be included in the new reading test. During the discussion, items that were considered good were selected while those that were judged inappropriate or weak were either rewritten or discarded. The group discussion ended with all test developers reaching agreement on 31 items.

The 31 test items and the selected passages were then converted into a coherent pre-test set. Then a copy of the subtest set was distributed to each of the four IEOP reading instructors for a comprehensive review. The instructors were asked to comment on the appropriateness of the subtest items and to make suggestions on the new items (see Appendix A). This reviewing stage was an important stage in the developmental process of the subtest because it helped the researcher gather some evidence on content validity of the subtest as well as enabled the test developers to re-examine any ambiguous or faulty items that had been overlooked.

The test development committee met for the final time during the second week of March, 1996, to discuss the reviews completed by the IEOP reading instructors. The committee worked through the suggestions and comments given in the reviews, and made some changes to the subtest items where they considered them appropriate. This final meeting resulted in a subtest consisting of 33 items.

Pilot subtest

A pilot version of the new reading subtest, consisting of 33 items, was administered during the second week of April, 1996. Prior to administering the pilot subtest, it was submitted for approval by the Human Subjects Review Committee at Iowa State University to ensure that no unintended improprieties would result from the administration of the test. A copy of the signed approval form is shown in Appendix B.

The subtest was administered to 36 international students: 26 101B students and 10 new IEOP students. These students had similar ability to the IEOP students who were at high or advanced levels. The administration of the subtest took place in the students' classrooms. Since the researcher wanted to estimate the reliability of the test, no time limit was allotted for this pilot subtest (i.e., students were allowed to take as much time as they needed to complete the test).

The students' performances on the pilot subtest were scored at the Test Evaluation Center at ISU. A standardized item analysis was conducted on the scores obtained. The analysis provided the researcher with descriptive statistics and item statistics. The descriptive statistics provided information on the general performance of the subtest.
Table 3 presents the descriptive statistics for the pilot subtest.

As shown in Table 3, the reliability of the pilot subtest was moderate (0.74), indicating that the subtest needs to be further improved. The mean of the subtest was high (25.94), reflecting that the subtest was fairly easy for the students. This finding was expected because none of the students who took the pilot subtest was from the low ability level group.

Table 3. Descriptive statistics for the pilot version of the new reading comprehension subtest

Descriptive statistic	Index
Number of subjects	36
Number of items	33
Maximum score	32
Mean	25.94
Variance	15.66
Standard deviation	3.96
Standard error of measurement in raw scores	2.02
Kuder-Richardson Formula 20 reliability	0.74

The subtest produced a moderately wide spread of scores which was indicated by a moderate standard deviation, variance and reliability. The histogram drawn in Figure 3 shows the distribution of the scores for the pilot subtest.

As shown in Figure 3, the subtest had a negative skewed scoring distribution, with many high-level students clustering to the right side of the histograms. The negatively skewed distribution suggests that the subtest was easy for most of the respondents and also reflects that the range of abilities in the test population was small. A similar distribution was obtained in other studies on test validation that use subjects ranging from high to advanced reading ability levels (Alderson, 1994; Harness, 1995).

The item statistics for the pilot subtest are tabulated in Table 4. The statistics include item facility (IF) and item discrimination (ID) indices. These statistics helped the test development committee to further improve the effectiveness of the subtest. The IF indices provided the committee with information on the difficulty level of individual items. The ID indices, on the other hand, informed the committee about the power of individual items to

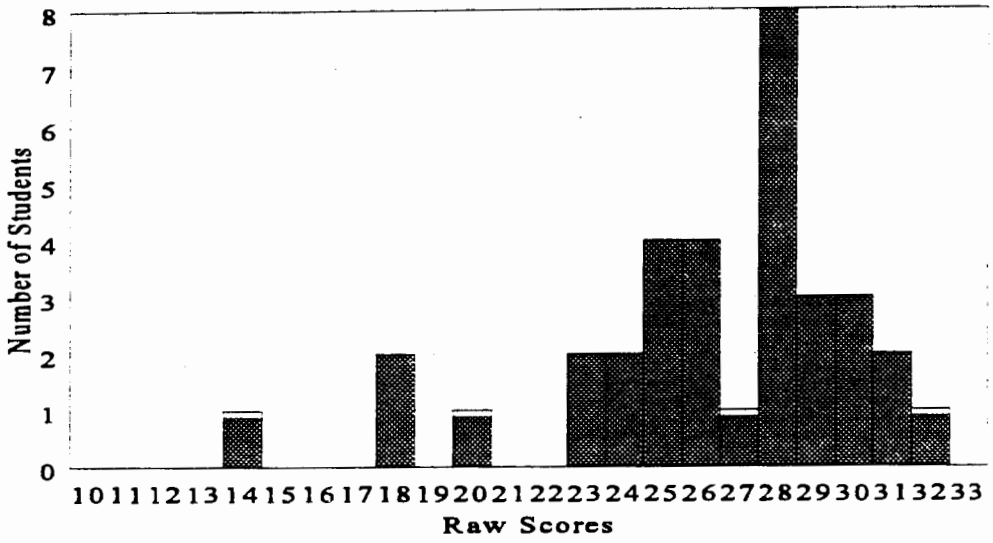

Figure 3. Distribution of scores on the pilot version of the new reading comprehension subtest

discriminate or distinguish proficient students from the less proficient ones. Therefore, the IF and ID indices assisted the committee in making decisions on which items to include, revise or reject.

As shown in Table 4, two-thirds of the items in the pilot subtest are in the extreme range of difficulty (i.e., they have IF values in the range of 80-100%). This finding indicated that the students who took the pilot subtest found most of the test items to be easy. This is not a surprising fact because the majority of the students who took the pilot subtest were considered to be either at the high or advanced level in reading. Nevertheless, a third of the items were in the desired range of difficulty (i.e., they have IF values within the range of 30-70%).

Although most of the items in the pilot subtest seem to be relatively easy, the ID indices of these items do indicate that the items could fairly discriminate high-scoring students

Table 4. Item statistics for the pilot version of the reading comprehension subtest

Item No.	Item facility (IF)	Item discrimination (ID)
1	97	0.34
*2	81	-0.01
3	89	0.40
*4	81	0.03
5	97	0.34
*6	97	-0.00
7	64	0.22
8	67	0.27
9	97	0.13
10	50	0.45
11	100	0.00
12	100	0.00
*13	75	-0.20
14	37	0.47
15	86	0.10
16	89	0.37
17	78	0.45
18	81	0.54
19	42	0.35
20	69	0.45
21	50	0.24
22	86	0.60
*23	97	-0.13
24	89	0.40
25	80	0.53
26	86	0.32
27	60	0.53
28	66	0.55
29	63	0.62
30	91	0.72
31	97	0.69
32	83	0.57
33	94	0.46

*Items that have been adjusted or discarded.

from the low-scoring ones. The majority of the items had satisfactory ID values ranging from 0.22 to 0.72.

To improve the quality of the items in the subtest, items that had either an ID value below 0.20 or a negative ID index (2, 4, 6, 13, and 23) were re-examined. Items with negative ID indices could not be included in the subtest because they have a high tendency to cause good students to miss or answer them wrongly and let poor students answer them correctly.

Upon re-examination, three items (2, 4, and 13) were adjusted, and two (6 and 23) were discarded. This reduced the final version of the subtest to 31 items instead of 33. All of the adjustments that were made involved rephrasing of the distractors in the items.

Following is an example of item 2 which was revised:

Item 2. The word "gestures" in this passage means

a. words used to convey feelings
b. different ways of talking
c. several different languages
d. expressive body movements

For item 2, the item analysis (see Appendix C) indicated that 7 out of 36 students who took the subtest selected distractor (b). The negative ID index of the item (-0.01) reflected that these 7 students were good students who obtained high overall scores in the subtest.

Having examined distractor (b) closely, the test developers found that the distractor could also be a right answer to the question in item 2. That is, the word "gestures" can mean different ways of talking (i.e., body language). The item was, therefore, adjusted by replacing the word "talking" in distractor (b) with "learning" (see Appendix D).

Item 4 was also adjusted because it had a very low ID index (0.03), indicating that it did not discriminate the students well.

Item 4. From this passage, we can conclude that

a. gestures are not important for travelers.
b. gestures are the same everywhere.
c. gestures do not contribute to effective communication.
d. gestures can contribute to effective communication.

The item analysis of the pilot subtest showed that 5 out of 36 students taking the subtest chose distractor (c) over key (d). All or some of the 5 students who answered item 4 wrongly could be good students. What is apparent in item 4 is that the distractor closely resembles the key. Such a close resemblance between options can trap "careless" students. For example, top scoring students who read the options in a hurry may select option (c) instead of (d) without paying much attention to the small difference between these two options. This kind of default must be avoided if the true abilities of the students are to be assessed.

The test developers of the new reading subtest attempted to solve the problem created by item 4 by rephrasing distractor (c) as follows:

c. gestures do not communicate thoughts

The last item revised by the test developers was item 13:

Item 13. What do you expect will happen next in this story?

a. The Peckhams will go on another vacation.
b. Hoa Van Nguyen will throw a bottle into the ocean.
c. Hoa Van Nguyen will write to the Peckhams.
d. The Peckhams will visit Hoa Van Nguyen.

Like item 2, item 13 also had a negative ID index (-0.20). Nine good students who answered item 13 wrongly chose distractor (b) over key (c). A re-examination of distractor (b) revealed that the distractor could be a plausible answer to the question addressed in item 13. The distractor was, therefore, rephrased as follows:

b. Hoa Van Nguyen will throw the paper away.

In the revision process, items 6 and 23 were discarded from the subtest. The decision to exclude these items was based on the items' high IF values (97%) and their negative ID indices (i.e., the items were considered to be very easy and have poor discriminative power). Nevertheless, some poor items (with an ID value below 0.20) were retained in the subtest. These items (9, 11, 12, and 15) were not discarded from the final version of the new reading subtest for three reasons. First, the items were viewed to be important because they probe some aspects of the students' reading abilities, many of which may constitute the construct of the new reading subtest. Second, the test developers presumed that the ID values of the items concerned would increase when they were tested in the final subtest on students with a wider range of abilities. Third, the items were not easily or quickly replaceable within the constraints of time during which the study was conducted.

Final subtest

The final version of the new reading subtest, which was a revised version of the pilot subtest, contained 31 items. This subtest was administered on the second-to-last day of the spring session, to 48 IEOP students of five different proficiency levels. The students took the subtest within their reading class period and were given 50 minutes to complete the subtest.

Analysis

The subjects' performances on both the pilot and final versions of the new reading subtest were scored at the Test Evaluation Center at ISU. Standardized item analyses of both tests were also conducted at the center. The analyses provided the researcher with item statistics and descriptive statistics. The results of the item analysis performed on the pilot test

helped the test development committee to further improve the efficiency and effectiveness of the test, while the analysis completed on the final test enabled the researcher to examine some empirical evidence on the validity of the test.

For construct validation, a correlational study was carried out between the subjects' scores on the new reading subtest and their scores on each of the subtests of the English Placement Test (EPT). To obtain evidence of criterion-related validity, correlations were computed between the subjects' scores on the new reading subtest and their scores on the reading section of the institutional TOEFL that was administered at the end of the session. The scores that were used in the correlational studies were of those students who took the final version of the reading subtest at the end of the spring session. The students' responses to the final version of the new subtest were given to their respective reading instructors to obtain their feedback on the students' performances on the new subtest related to their performance in the reading class: (i.e., to ascertain whether the students' reading test scores matched their respective IEOP instructors' assessments).

Summary

The primary purpose of the study was to develop and validate a new reading placement subtest. The subjects of the study were 84 international students at ISU. The instrument used in the study was a multiple-choice reading subtest developed by the researcher together with four other test developers who had a background in Teaching English as a Second Language (TESL). Pre-testing procedures were adopted in the study. The procedures include (a) pilot testing, (b) item analysis and (c) item revision.

A pilot subtest was administered to 36 international students at ISU. The item analysis conducted on the scores indicated that some of the items needed to be revised. Based on the

analysis, three items were revised and two were discarded.

The revised version of the subtest was administered to 48 international students. The scores on this final version of the subtest were also analyzed statistically to examine the effectiveness of the revision procedures and to demonstrate the validity of the subtest.

CHAPTER IV. FINDINGS AND DISCUSSION

Introduction

The primary purpose of this study was to develop and validate a new reading comprehension subtest for the Intensive English and Orientation Program (IEOP) at Iowa State University. The new subtest was specifically designed to assess IEOP students' reading abilities. The scores on the subtest were to be used for making responsible placement decisions within the reading course offered in IEOP.

The data gathered in the present study were item statistics and descriptive statistics. The results of the analyses of the data were used for two purposes: (a) improving the subtest as a placement instrument and (b) validating the subtest. Three strategies were employed in the investigation of the validity of the subtest: the content, construct, and criterion-related approaches.

The results obtained from the analyses conducted on the scores of the final version of the new reading subtest. They are reported and discussed under four main headings: (a) Item Statistics; (b) Descriptive Statistics; (c) Reliability; and (d) Validity.

Item Statistics

The scores obtained from the final subtest were analyzed using classical norm-referenced item analysis statistics (see Appendix E). Item facility (IF) and item discrimination (ID) indices calculated on the scores of final version of the new reading comprehension subtest are reported in Table 5.

Table 5. Item statistics for the final version of the reading comprehension subtest

Item No.	Item facility (IF)	Item discrimination (ID)
1	90	0.20
*2	75	0.62
3	94	0.13
*4	92	0.28
5	94	0.39
6	54	0.38
7	64	0.37
8	98	0.13
9	54	0.20
10	90	0.47
11	96	0.38
*12	75	0.27
13	29	0.26
14	88	0.27
15	77	0.29
16	65	0.44
17	79	0.50
18	17	0.39
19	73	0.33
20	48	0.42
21	88	0.42
22	73	0.50
23	72	0.46
24	90	0.27
25	52	0.50
26	73	0.40
27	46	0.32
28	83	0.08
29	88	0.24
30	88	0.33
31	83	0.66

*Revised items.

Note: In the final subtest, all item numbers after 5 and 21 have been re-arranged accordingly due to the decision to exclude items 6 and 23 of the pilot subtest.

The item analysis produced many favorable results. As shown in Table 5, 13 items demonstrated an increase in their ID values. All of the three revised items (2, 4, and 12) had higher ID values. Both items 2 and 12 that produced negative ID indices in the item analysis conducted on the results of the pilot subtest, produced positive ID indices in the second analysis (the ID values for items 2 and 12 were 0.62 and 0.27, respectively). The ID value of item 4 had increased significantly, from 0.03 to 0.27. A few other items (10, 11, 20, and 31) also demonstrated significant increases. For example, items 10 and 11 had increased from a zero value (0.00) for item discrimination to an ID value of 0.47 and 0.38, respectively.

The analysis also indicated fluctuations in the percent difficulty of the items (i.e., some items had higher item difficulties while others had lower item difficulties). Ten items demonstrated an increase in their ID values.

Despite these apparently favorable results, there were some undesirable ones that arose from this second analysis. Disturbingly, about half of the items in the final version of the subtest produced lower ID values in the analysis. Nevertheless, the ID values of these items, with the exception of items 3, 8, and 28, were still in the satisfactory range (i.e., the items had ID values above 0.20).

Descriptive Statistics

The descriptive statistics for the final version of the new reading subtest are tabulated in Table 6. The mean raw score of the subtest was high, indicating that the students found the subtest fairly easy. The negative skewed scoring distribution as shown in Figure 4 also indicated that most of the students taking the subtest obtained high scores.

Table 6. Descriptive statistics for the final version of the new reading comprehension subtest

Descriptive statistic	Index
Number of subjects	48
Number of items	31
Maximum score	30
Mean	22.75
Variance	18.31
Standard deviation	4.28
Standard error of measurement in raw scores	2.09
Kuder-Richardson Formula 20 reliability	0.76

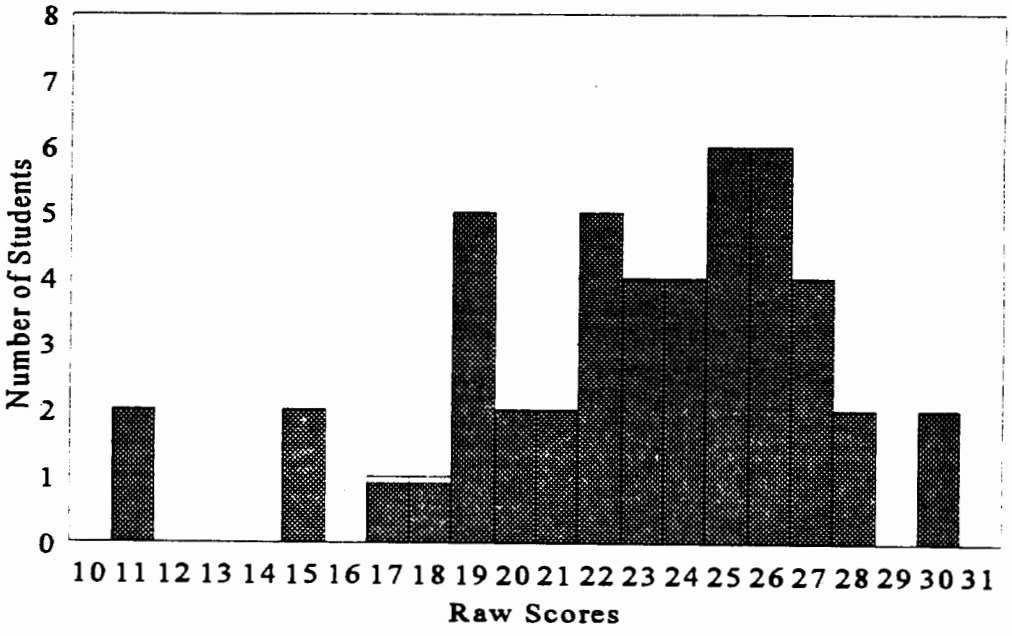

Figure 4. Distribution of scores on the final version of the new reading comprehension subtest

These findings were expected because the students were of the ability levels ranging from intermediate to advanced.

When the score distribution pattern of the final subtest was compared to that of the pilot subtest (see Figure 1 in Chapter 3), it was found that the final subtest had a slightly wider spread of scores than the pilot subtest. This is not surprising because the students who took the final subtest were from a test population of a slightly wider range of proficiency levels than those who sat for the pilot subtest (i.e., only high and advanced level students took the pilot subtest). In fact, this difference was reflected by the standard deviations of both versions of the subtest: the standard deviation of the final subtest was 0.32 higher than that of the pilot subtest (3.96 vs. 4.28).

However, the simple subtraction of the standard deviation of the pilot subtest from the final subtest is misleading. According to Baker (1989), the standard deviations between the two tests can only be correctly compared by dividing the standard deviations by the total scores of the tests. Dividing the standard deviations of both subtests by the total scores will produce the following coefficients of discrimination:

Pilot subtest	3.96 / 33 = 0.12
Final subtest	4.28 / 31 = 0.14

The calculated values (0.12 and 0.14) revealed that the final subtest spread out the students' scores only slightly more effectively than the pilot subtest.

Reliability

As reported in Table 6, the reliability coefficient calculated on the scores of the final version of the new reading subtest was moderate (0.76). This indicates that the subtest needs

to be further improved before it is used to make placement decisions.

Upon comparison, the reliability coefficient of the final subtest was found to be .02 higher than the coefficient of the pilot subtest. This means that after revision, the reading subtest had only a marginal increase in reliability coefficient. This slight increase can be attributed to several factors such as variation in group ability, the length of the subtest, the discrimination and facility values of subtest items, and the increase of errors in the subtest (Alderson, 1994; Hatch and Farhady, 1982; Henning, 1987). A more detailed explanation of the effects of these factors on the reliability coefficient of the new reading subtest follows.

When the sample populations of both versions of the reading subtest were compared, it was discovered that the range of ability of the population in the final subtest was only slightly wider than the one in the pilot subtest. Since the difference between the range of ability of the population in the pilot and final versions of the subtest was only marginal, one may, therefore, expect the final version of the subtest to have only a slightly higher reliability coefficient than the pilot version.

Furthermore, the final subtest was shorter in length than the pilot subtest. That is, the final subtest had 31 items whereas the pilot subtest had 33 items. According to Henning (1987), the length of a test can affect its reliability, i.e., the more items a test has, the higher is its reliability. Since the number of items included in the final subtest was not higher than that used in the pilot subtest, one cannot expect the reliability of the final subtest to be very much higher than the reliability of the pilot subtest.

The statistics in Table 6 also showed that the final subtest had a higher standard error of measurement in raw scores than the pilot subtest (2.09 vs. 2.02). This suggests that there

was more random error in the final subtest which in turn, helps to explain why the final subtest did not have a high increase in reliability.

The amount of error in the final subtest can be reduced by increasing the reliability of the subtest. To increase reliability, the items in the subtest have to be further revised. As indicated in Table 5, most of the items in the final subtest were still within the extreme IF range (71-98 percent) and many did not have a very good discrimination value (.40+). A further improved subtest must, therefore, include more items in the medium range of facility (30-70 percent) and items that have very good discriminative power.

Validity

Three different types of validity were investigated in the study: (a) content, (b) construct, and (c) criterion-related validity.

Content validity

Since the new reading subtest was specifically designed to place IEOP students in appropriate reading classes, its content validity was examined by comparing the content to that of the IEOP reading classes. During the development of the new reading subtest, IEOP reading instructors were asked to review the passages and items developed for the subtest to determine whether they were representative samples of those used in IEOP reading classes (see Appendix A). In their reviews, all of the instructors stated that they preferred the format of the new subtest to that of the English Placement Test (EPT) because the former corresponds to the format used in IEOP reading classes.

In their response to the subtest items, the instructors commented that some of the

items needed to be revised and more items on inferencing and deriving main ideas should be included in the new reading subtest. The IEOP reading instructors also seemed to agree that, in general, the passages and items of the new subtest were similar to the ones they used in IEOP classes (i.e., in general, the content of the new subtest was assessing the reading skills taught in IEOP reading classes). Nevertheless, the level of agreement found among the instructors was not statistically quantified.

Construct validity

To gather evidence of the construct validity of the new reading subtest (i.e., to prove that the subtest measures the language skills it purports to measure), a correlation study was carried out. The basic assumption underlying such a study is that two subtests that are measuring different aspects or skills will not correlate strongly with one another while those that are assessing similar aspects of language will correlate highly.

Table 7 presents the results of the correlations computed between the respondents' scores on the new reading subtest and their respective scores on each of the four subtests in the English Placement Test (EPT) currently used by IEOP. The results showed that the computed correlation coefficients demonstrated that each of the subtests measured a distinct aspect of language (i.e., none of the coefficients was significant enough to show that a pair of subtests measured a similar ability or was based on similar constructs).

A closer examination of the correlation coefficients revealed that there was a moderate relationship between the new reading subtest and the EPT ($r = 0.64$). The coefficients presented in Table 7 also demonstrated that the new reading subtest had the strongest

Table 7. Pearson product-moment correlations between scores on the new reading subtest and scores on the respective four EPT subtests.

	Listening	Grammar	Vocabulary	Reading	Total
New reading subtest	0.33*	0.59**	0.68**	0.56**	0.64**

* significant at .05.
** significant at .01.

relationship with the EPT vocabulary subtest (r = 0.68). This finding is desirable because it indicates that, to a certain extent, there was an agreement between the scores obtained in the two subtests. This agreement, or overlapping, was expected since one of the abilities that the new reading subtest was designed to assess is the comprehension of vocabulary items.

The data gathered also indicated that the correlation between the scores on the new reading subtest and the scores on the EPT listening subtest was low (0.33). This finding is an important piece of evidence of construct validity because it supports the notion that two tests measuring different skills are not expected to be related to one another.

Criterion-related validity

Criterion-related validity involves the demonstration of the degree to which the respondents' scores on the new reading subtest match other measures (external criterion measures) of their abilities. In the present study, criterion-related validity evidence was gathered by computing correlations between the respondents' scores on the new reading subtest and their scores on the reading section of the Institutional TOEFL administered at the end of the semester. For comparison purposes, the correlation between the respondents'

scores on the EPT reading subtest and their scores on the reading section of the institutional TOEFL were also computed. The resultant correlation coefficients are reported in Table 8.

As indicated in Table 8, not all of the respondents taking the new reading subtest or the EPT reading subtest took the Institutional TOEFL. Approximately two-thirds of the 48 respondents who sat for the new reading subtest took the Institutional TOEFL and only 27 respondents took both the Institutional TOEFL and the EPT reading subtest.

The results indicated that the computed correlation coefficients were not high enough to provide convincing criterion-related validity evidence. The coefficients obtained were only within a moderate range. Nevertheless, the resultant correlation coefficients revealed an important piece of information regarding the criterion-related validity of the EPT reading subtest. The correlation coefficient between the EPT reading subtest and the Institutional TOEFL was not high indicating that the EPT reading subtest, like the new reading subtest, was not highly criterion-related valid. In fact, the correlation coefficient concerned was only 0.09 higher than the coefficient obtained between the new reading subtest and the Institutional TOEFL. This slight difference could be due to the fact that the EPT had a high reliability. The test developers of the EPT stated that the reliability of the entire EPT ranged between 0.89

Table 8. Pearson product-moment correlations between scores on the reading section of the Institutional TOEFL and scores on the new reading subtest and the EPT reading subtest

Reading section of the Institutional TOEFL	n	r
New reading subtest	30	0.63*
EPT reading subtest	27	0.72*

* significant at .01

and 0.94 (see Appendix F). Although the developers do not provide reliability data for the individual subtests in the EPT, it is likely that the reliability of the EPT reading subtest is higher than that of the new reading subtest developed in this study.

However, caution must be used in thinking about the correlation coefficients obtained in this study because the samples used were truncated (i.e., not all of the original test population was used in the validation). The effect of using a truncated sample is that it is likely to lower the reliability and validity coefficients.

To gather more evidence on criterion-related validity, the respondents' scores were also returned to the respondents' IEOP reading instructors who were asked to compare the scores to their class assessment of the respondents' reading abilities. The feedback received from the instructors demonstrated that the scores generally matched the instructors' class ranking of the respondents' reading abilities. In other words, the distribution of scores for each IEOP reading class reflected that, generally, students who performed well on the new reading subtest also did well in their reading class while those who did not fare well on the new subtest did not do so in their reading class. However, no computation of validity coefficients was conducted between the respondents' scores and their IEOP reading instructors' class assessment.

Summary

This chapter has been devoted to the presentation and discussion of the results obtained in the study. The results produced by the analyses of the data gathered in the study were used to improve the effectiveness of the newly developed reading subtest and also to demonstrate the validity of the subtest. The item analysis on the scores of the pilot subtest

indicated that some of the items needed to be revised. Based on the analysis, three items were rewritten and two were discarded. The revision was proven to be successful by the results of the item analysis conducted on the scores of the final subtest.

The descriptive statistics for the final subtest indicated that the subtest needs to be further revised before it is used to make placement decisions. The reliability of the subtest was only moderate. In fact, the reliability of the final subtest was only 0.02 higher than the reliability of the pilot subtest. This slight increase in the reliability of the final subtest was due to some changes in (a) the composition of the subtest population, (b) the number of items used in the subtest, (c) the amount of error in the subtest, and (d) the IF and ID values of the items in the subtest.

Several pieces evidence of validity were also gathered in the study. Content evidence was obtained through the analysis of the IEOP reading instructors' reviews of the new reading subtest. The reviews suggested that the passages and items of the subtest were representative samples of those taught to IEOP students. Construct validity evidence was delivered by the computation of correlations between the respondents' scores on the final subtest and their scores on the respective four subtests on the EPT. The correlation coefficients obtained indicated that each of the subtests examined was testing a different ability or skill. Two pieces of evidence on criterion-related validity were obtained through a correlational study and the IEOP reading instructors' evaluation of the respondents' scores. However, these pieces of evidence were viewed to be weak.

CHAPTER V. SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

The present study attempted to develop and validate a new reading placement subtest for the Intensive English and Orientation Program (IEOP) at Iowa State University (ISU). This chapter presents a summary of the study, conclusions drawn from the findings, implications of the findings, and recommendations for future research.

Summary

In the study, a new reading subtest specifically designed to assess the IEOP students' reading ability was developed by a test development committee comprised of members with a background in Teaching English as a Second Language (TESL). The scores of the new subtest were intended to be used for placement purposes. That is, the scores were to be used for placing IEOP students in appropriate reading classes.

Pretesting procedures were adopted in the study. The pilot version of the new reading subtest was administered to international students of high and advanced levels in reading ability. Upon revision, the final version of the new subtest was given to IEOP students, ranging from low intermediate to advanced levels in reading ability.

Statistical data gathered from the subtest versions helped the researcher to improve and validate the new subtest. The item statistics computed for the pilot subtest indicated that most of the items were fairly easy and had a satisfactory discriminative power. However, five items had to be revised due to a very low or negative ID index. The revision was proven to be effective by the item statistics computed for the final subtest. These statistics also showed that the IF and ID indices of the subtest items fluctuated when the items were tested on students with a wider range of reading ability.

The descriptive statistics of the pilot and final subtests revealed that each subtest had a respectably high mean. The reliability coefficients for both subtests were moderate indicating that the subtests need further revision. The correlations computed between the respondents' scores on the final subtest and the scores on each of the four subtests in the EPT produced coefficients reflecting that each subtest was tapping a different ability. The correlations calculated between the respondents' scores on the reading subtest of the TOEFL and their respective scores on the new reading subtest and the EPT reading subtest produced coefficients within a moderate range. The reviews and feedback received from the IEOP reading instructors revealed that the instructors generally agreed that the content of the new reading subtest was a representative sample of that of the IEOP reading course, and that the respondents' scores on the final subtest tended to confirm the instructors' class assessment of the respondents' reading abilities.

Conclusions

The conclusions drawn from the study are of two kinds. The first relates to the procedures followed in the development of the new reading subtest, while the second relates to the reliability and validity of the new subtest.

Conclusions related to the test procedures

This study has demonstrated that test development is essentially a collaborative effort in which trained and experienced test developers work together to produce a test. The result of such a collaboration is the development of a good and reliable test that consists of items that were created based on the experience and judgments of the test developers. In the study,

the item and descriptive statistics of the new reading subtest reflect the success of the collaborative work of five test developers. Most of the items created for the subtest had satisfactory discriminative power (ranging from 0.20 to 0.72) and the subtest's descriptive statistics indicated that the subtest has potential as a placement instrument.

However, there is evidence in the literature that indicates that professional judgments of experienced test developers may not always be accurate (Alderson, 1993; Buck, 1991). For this reason, items generated by all the test developers need to be pre-tested before they are included in the final test administered to the target test population. The results of the present study have indicated that pilot-testing is an essential part of the test development process. The pre-testing procedures followed in the development of the new reading subtest for IEOP provided the test developers feedback on both the test items and the procedures.

The results of the present study have also shown that standardized item analysis is practical and valuable. In the study, statistical analysis of the test items helped the test developers to study the discriminatory power and difficulty level of the test items. The analysis confirmed the predictions that the test developers made about the fairness and suitability of the test items, and pointed out faulty and ambiguous items that had been overlooked. In other words, the analysis enabled the test developers to monitor the quality of the test items and helped them to produce a good and reliable test.

Conclusions related to the new reading subtest

On the basis of the analyses of the data gathered in the study, several conclusions can be drawn about the new reading subtest. The first conclusion is related to the effectiveness and future utility of the subtest. The descriptive statistics reported in Chapter 4 and 5

indicated that the new reading subtest can function effectively as a norm-referenced reading placement instrument if the items of the subtests are further improved.

The second conclusion to be drawn from the results of the study concerns test reliability. The reliabilities of both versions of the new reading subtest were moderate. An interesting test behavior was observed during the examination of the reliabilities of both the pilot and final versions of the new reading subtest. It was observed that the reliability coefficient of the new subtest had only a marginal increase (0.02) when it was administered the second time. Upon a detailed investigation, it was discovered that the small increase was partly due to the composition of the subtest's population. That is, the range of ability of the population in the final subtest was not so much wider than the one in the pilot subtest. This observation leads to the third conclusion—the performance of a test is influenced by the composition of its population.

The final conclusion centers on the validity of the new reading subtest. There is some evidence in the study to suggest that the new subtest is respectably valid. The evidence for construct validity of the new subtest is provided by correlation coefficients computed between the respondents' scores on the new reading subtest and their scores on the four subtests in the English Placement Test (EPT) currently used in IEOP. The coefficients revealed that the aspect of language measured by the new subtest is distinct from those assessed by the respective subtests in the EPT.

The study also demonstrated that the new subtest is relatively content and criterion-related valid. The evidence for these two types of validity was obtained from the IEOP reading instructors. From the instructors' reviews of the new subtest, some consensus was

observed among the instructors regarding the representativeness of the content of the new subtest. Generally, the instructors seemed to agree that the content of the new reading subtest represents that of the IEOP reading course. A similar agreement was discernible concerning the relationship between the respondents' scores on the new reading subtest and the IEOP reading instructors' class assessment of the respondents' reading abilities. That is, there was a strong tendency for the scores to match the instructors' assessments. However, these two pieces of evidence on content and criterion-related validity are considered weak because they were derived from the data that were not quantified. This study has also shown the difficulty of gathering convincing statistical evidence on criterion-related validity with truncated samples.

Implications and Recommendations for Further Research

The implications of the findings of the present study and the recommendations of the researcher for future research will be discussed in relation to the contributions of the study to the Intensive English and Orientation Program (IEOP) and to the field of language testing.

Contributions to the Intensive English and Orientation Program

The findings of this study have value and social implications for IEOP. The study has shown that a respectably reliable and valid reading placement subtest can be developed specifically for use in IEOP. The new reading subtest developed in the study is viewed to be a better placement instrument than the one currently used in IEOP (i.e., the EPT reading subtest). The content of the new subtest is proven to be a representative sample of that of the IEOP reading course. All the IEOP reading instructors preferred the reading passages used in

the new reading subtest to the discrete sentences used in the EPT reading subtest because the reading materials that the instructors introduced in IEOP reading classes consisted of academic and non-academic passages and not of discrete sentences. Most of the items in the new reading subtest were also considered to be content valid because they assessed similar reading skills taught to IEOP students. Based on these facts, it was, therefore, expected that the scores obtained from the new reading subtest will provide better indications of the IEOP students' reading abilities.

Using the scores of the new subtest for placement purposes can thus lead to two positive consequences that have both social and pedagogical implications for IEOP. First, since the students will be placed in classes based on their "true" abilities, they will benefit more from the classes and will feel comfortable learning and interacting with other members of the class of approximately similar ability. Second, the reading instructors can make the teaching and learning of reading skills more efficient and effective because similar language or teaching points can be used in a homogeneous group of students.

However, the new reading subtest needs to be further improved and evaluated if it is to be used in IEOP. There are two reasons for this line of reasoning. The first reason concerns the subtest's reliability and validity. The findings of the study have demonstrated that the reliability of the new subtest is only moderate and some of the validity evidence gathered is not very strong. The second reason relates to the generalizability of the findings of the subtest. The results of the subtest cannot be generalized to other testing setting because what is valid and reliable in one testing setting may not be generalizable to another testing settings (Henning, 1989). Therefore, it is recommended that more studies be carried out

within IEOP to further improve and validate the new reading subtest.

Since the present study has used a relatively small number of items and the majority of the items was not within the significant item discrimination (ID) and item difficulty (IF) index range, a follow-up study should include a larger number of items that have higher ID indices and are within the medium range of difficulty (30-70 percent). Increasing the number of subtest items and the items' ID indices can help to increase the reliability coefficient of the subtest. Including more items with IF values in the range of 30-70 percent in the subtest can also cause the reliability of the subtest to increase because these items usually result in higher item discrimination and thus better distinguish the subtest's population according to their levels of ability.

The reliability of the subtest can also be increased by enlarging the size of the subtest's population and widening the range of ability of the population (Alderson, 1994; Henning, 1987). The size of the population in the current study was small (less than 100 students) and the range of ability of the population was not extensive enough to cover all levels of ability in IEOP. A future study should therefore use a larger number of population of a wider range of reading ability to allow for potentially greater reliability and a wider spread of scores. The advantage of improving the quantity and quality of the items and population of the subtest is that it can increase both the subtest's reliability and validity. The validity of the further improved subtest needs to be examined using better procedures than those used in the present study. For instance, the judgments and assessments of the IEOP reading instructors should, in a future study, be quantified using some new or adopted data collection instrument.

Further revision and evaluation of the new reading subtest will definitely consume time

and require expertise. Therefore, it is recommended that a group of IEOP instructors be assigned to continue improving the quality of the subtest.

Since it has been shown in this study that a respectably reliable and valid subtest can be generated specifically for use in IEOP, additional research should be carried out to develop and validate subtests in the other three skill areas (listening, grammar and vocabulary) so that these new subtests can replace all the EPT subtests currently used in IEOP.

Contributions to the field of language testing

This study contributes to the language testing literature on the development and validation of placement tests. The practical and theoretical issues dealt with in the present study are of interest to the field of language testing. The findings of the study not only confirm some of the findings of other related studies but also add more to the body of knowledge related to test development and validation.

The study also produced findings that help to throw some light on the procedures to be followed in test development. The findings suggest that all test development attempts should include statistical analysis and pre-testing procedures. In addition, the findings of the study also indicate the need to consider a variety of things such as test context, purpose and test population prior to test development.

Since the study has demonstrated the benefit of team work among test developers, it is therefore desirable that future test development be a collaborative effort of trained and experienced test developers. The study has also shown the difficulty of gathering convincing evidence of a test's validity. Future research should therefore focus on finding appropriate or legitimate ways to validate tests.

APPENDIX A. MEMO AND CHECKLIST SENT TO REVIEWERS, AND SELECTED COMMENTS

To:
From: Rosyati Abdul Rashid (graduate student in TESL)
Subject: Evaluating a reading test
Date: March 1, 1996

Currently, I am carrying out a thesis on a validation study of the reading section of the IEOP's placement test. A new reading subtest has been developed by a committee comprised of myself, Dr. Dan Douglas, Dr. Barbara Matthies, Fellicity Douglas and Dan Harness. A pilot test will be carried out during the third or fourth week of March, 1996.

I would like to ask for your opinion and comment on this new test so that further improvement or adjustment can be made before the piloting date. I would appreciate it very much if you could **respond to all the test items** (i.e., by circling the correct answers) and **complete the checklist** attached.

Please return the attached reading passages and the checklist to my mailbox in Ross 206 or put them in a box (with my name written on it) in the IEOP's library (Ross 326), by March 8, 1996.

Thank you very much for your cooperation.

Comments:

"This looks really good to me. My first suggestion is to find a way to include the kinds of questions like "What topic preceded this passage?" or "What is the paragraph following this one likely to be about?" My students always have trouble with those, and the students who do well on them are the best readers

My second suggestion is to find a way to include more inference items. They also separate the best readers.

I am giving you my first reactions. Call me if you have any questions. Excellent passages!

Checklist

1. This passage is typically of what level?

   a. Beginning b. Low intermediate c. High intermediate d. High

2. Please indicate whether there is any problem with the test items, i.e., whether the items are bad or whether the questions and the given answer choices need rewording.

Item No.	Problem

3. Please suggest some new test items.

APPENDIX B. HUMAN SUBJECTS APPROVAL

Last Name of Principal Investigator Abdul Rashid

Checklist for Attachments and Time Schedule

The following are attached (please check):

12. [X] Letter or written statement to subjects indicating clearly:
    a) purpose of the research
    b) the use of any identifier codes (names, #'s), how they will be used, and when they will be removed (see Item 17)
    c) an estimate of time needed for participation in the research and the place
    d) if applicable, location of the research activity
    e) how you will ensure confidentiality
    f) in a longitudinal study, note when and how you will contact subjects later
    g) participation is voluntary; nonparticipation will not affect evaluations of the subject

13. [ ] Consent form (if applicable)

14. [ ] Letter of approval for research from cooperating organizations or institutions (if applicable)

15. [X] Data-gathering instruments unrevised set of passages is attached.The questionnaire has not yet been developed and will be submitted later.

16. Anticipated dates for contact with subjects:

First Contact	Last Contact
3 / 20 / 1996	5 / 10 /1996
Month / Day / Year	Month / Day / Year

17. If applicable: anticipated date that identifiers will be removed from completed survey instruments and/or audio or visual tapes will be erased:

7 / 12/ 1996
Month / Day / Year

18. Signature of Departmental Executive Officer | Date | Department or Administrative Unit

Dale A. Ram | 3/8/96 | ENGLISH

19. Decision of the University Human Subjects Review Committee:

___ Project Approved ___ Project Not Approved ___ No Action Required

X Project approved with the understanding the final questionnaire will be submitted when it is compieted.

Patricia M. Keith | 3/13/96 | PM Keith

Name of Committee Chairperson | Date | Signature of Committee Chairperson

GC: 8/95

APPENDIX C. STANDARDIZED ITEM ANALYSIS OF THE PILOT VERSION OF THE NEW READING SUBTEST

ENGLISH RASTRID IEOP READING EXAM(RESEARCH) SPRING 1996 (CASH

KR-20 RELIABILITY ESTIMATE = 0.74 *** SCORE DISTRIBUTION ***
AVERAGE TEST SCORE = 79%
ERROR VARIANCE = 4.06
STANDARD ERROR OF MEASUREMENT IN RAW SCORES = 2.02
STANDARD ERROR OF MEASUREMENT IN TSCORES = 51.02

NUMBER TAKING TEST = 36
MEAN = 25.94
VARIANCE = 15.66
STANDARD DEVIATION = 3.96
NUMBER OF SCORED ITEMS = 33

SCORE	N	CUM	%ILE	TSCORE	(NUMBER OF ASTERISKS=N)
14	1	1	3	198	!*
15	0	1	3	223	!
16	0	1	3	249	!
17	0	1	3	274	!
18	2	3	8	299	!**
19	0	3	8	325	!
20	1	4	11	350	!*
21	0	4	11	375	!
22	2	6	17	400	!**
23	2	8	22	426	!**
24	2	10	28	451	!**
25	4	14	39	476	!****
26	4	18	50	501	!****
27	1	19	53	527	!*
28	8	27	75	552	!********
29	3	30	83	577	!***
30	3	33	92	602	!***
31	2	35	97	626	!**
32	1	36	100	653	!*

ENGLISH RASTRID IEOP READING EXAM(RESEARCH) SPRING 1996 (CASH *** ITEM ANALYSIS *** 16-Apr-96

OMIT=NUMBER OMITTING THE ITEM DIFF=ITEM DIFFICULTY=% RIGHT # INDICATES CORRECT ANSWER
NA=NUMBER ATTEMPTING THE ITEM DISC=ITEM DISCRIMINATION
NR=NUMBER ANSWERING CORRECTLY =ITEM-SCORE CORRELATION % INDICATES <10% CORRECT

ITEM	OPTION 1/A	2/B	3/C	4/D	5/E	6/F	7/G	8/H	9/I	10/J	OMIT	NA	NR	DIFF	DSCR	ITEM
1	1	35#	0	0	0	0	0	0	0	0	0	36	35	97	0.34	1
2	0	7	0	29#	0	0	0	0	0	0	0	36	29	81	-0.01	2
3	32#	0	3	1	0	0	0	0	0	0	0	36	32	89	0.40	3
4	2	0	5	29#	0	0	0	0	0	0	0	36	29	81	0.03	4
5	0	35#	1	0	0	0	0	0	0	0	0	36	35	97	0.34	5
6	35#	0	1	0	0	0	0	0	0	0	0	36	35	97	-0.00	6
7	11	23#	0	2	0	0	0	0	0	0	0	36	23	64	0.22	7
8	0	0	12	24#	0	0	0	0	0	0	0	36	24	67	0.27	8
9	0	0	1	35#	0	0	0	0	0	0	0	36	35	97	0.13	9
10	18#	17	1	0	0	0	0	0	0	0	0	36	18	50	0.45	10
11	0	0	36#	0	0	0	0	0	0	0	0	36	36	100	0.00	11
12	0	0	0	36#	0	0	0	0	0	0	0	36	36	100	0.00	12
13	0	9	27#	0	0	0	0	0	0	0	0	36	27	75	-0.20	13
14	13#	20	0	2	0	0	0	0	0	0	1	35	13	37	0.47	14
15	3	31#	2	0	0	0	0	0	0	0	0	36	31	86	0.10	15
16	32#	2	1	1	0	0	0	0	0	0	0	36	32	89	0.37	16
17	1	0	28#	7	0	0	0	0	0	0	0	36	28	78	0.45	17
18	29#	4	1	2	0	0	0	0	0	0	0	36	29	81	0.54	18
19	15#	19	2	0	0	0	0	0	0	0	0	36	15	42	0.35	19
20	25#	5	6	0	0	0	0	0	0	0	0	36	25	69	0.45	20
21	15	0	3	18#	0	0	0	0	0	0	0	36	18	50	0.24	21
22	4	31#	0	1	0	0	0	0	0	0	0	36	31	86	0.60	22
23	35#	0	0	1	0	0	0	0	0	0	0	36	35	97	-0.13	23
24	4	31#	0	0	0	0	0	0	0	0	1	35	31	89	0.40	24
25	2	5	28#	0	0	0	0	0	0	0	1	35	28	80	0.53	25
26	0	1	4	30#	0	0	0	0	0	0	1	35	30	86	0.32	26
27	2	12	21#	0	0	0	0	0	0	0	1	35	21	60	0.53	27
28	7	3	23#	2	0	0	0	0	0	0	1	35	23	66	0.55	28
29	22#	3	1	9	0	0	0	0	0	0	1	35	22	63	0.62	29
30	1	32#	2	0	0	0	0	0	0	0	1	35	32	91	0.72	30
31	34#	0	0	1	0	0	0	0	0	0	1	35	34	97	0.69	31
32	1	2	3	29#	0	0	0	0	0	0	1	35	29	83	0.57	32
33	0	2	33#	0	0	0	0	0	0	0	1	35	33	94	0.46	33

APPENDIX D. NEW READING COMPREHENSION SUBTEST

Instructions

This test is designed to measure your reading ability. There are five passages in the test. Read each of the passages and answer the questions which follow it. Mark your answers on the answer sheet given. Please do not make any marks in this test booklet.

Example:

You read the following:

Scientists have often described organic processes by means of analogy. Some analogies are useful and accurate as far as **they** go. For example, the comparison of the heart with a pump or of the kidney with a filter has helped illustrate the nature and function of these organs.

Then, you answer the question.

"**they**" in this passage refers to

a. Scientists
b. Processes
c. Analogies
d. Organs

The correct answer is c, so you should mark c on your answer sheet.

The Language of Gestures

We do a lot of talking, asking, answering, telling, saying. But we do much of our talking without words. We often use a kind of "body language" to show what we think or feel.

This body language is the language of gestures. We point a finger, raise an eyebrow, wave an arm or move another part of the body to show what we want to say. In other words, we "talk" with these gestures.

People all over the world use gestures. In every country, there are gestures that say "Hello" and "Good-bye." However, this does not mean that everyone in the world uses exactly the same body language. We may have some of the same gestures, but different countries have different customs and different gestures. Sometimes the same gesture can mean different thing in different countries. For example, Saudi Arabians and some other speakers of Arabic say "Come here" with a gesture that most Europeans use to say "Good-bye".

Because gestures can say different things in different countries, body language can be a problem for travelers. Learning words in a new language is not enough. If you want to talk to people who speak a different language, you might have to learn some new gestures too.

1. "talk" in line 4 means

   a. move
   b. communicate
   c. observe
   d. think

2. The word "gestures" in this passage means

   a. words used to convey feelings
   b. different ways of learning
   c. several different languages
   d. expressive body movements

3. According to the passage, people express their thoughts and feelings through

   a. speech and body language
   b. speech and picture language
   c. body and picture language
   d. spoken and written language

4. From this passage, we can conclude that

   a. gestures are not important for travelers
   b. gestures are the same everywhere
   c. gestures do not communicate thoughts
   d. gestures can contribute to effective communication

The Bottle

In December 1979 Dottie and John Peckham, a Los Angeles couple, went to Hawaii on vacation. They traveled by ship.

Some people on the ship threw bottles into the ocean. Each bottle had a piece of paper in it. On each piece of paper were a name, an address, and a message: "If you find this bottle, write to us."

Mrs Peckham wanted to throw a bottle into the ocean, too. She put the piece of paper and one dollar into a bottle. She put a cap on the bottle and threw the bottle into the water.

Three years later and 24,139 miles away, Hoa Van Nguyen was on a boat, too. He, his brother, and 30 other people were going to Thailand in a small boat. The boat was in the Gulf of Thailand There wasn't any drinking water in the boat, and Hoa was thirsty. He saw a bottle in the sea. The bottle was floating near the boat. "What's in the bottle? Maybe I can drink something," he thought. Hoa took the bottle out of the sea and opened it. There wasn't any drinking water in the bottle. But there was a dollar bill. There was also a piece of paper. There was a name and an address on the paper. The name was Peckham. The address was in Los Angeles, California.

6. The word "cap" in line 7 means

   a. hat
   b. top
   c. cup
   d. bag

7. In line 12, "was floating" means

   a. was lying on the bottom of the ocean
   b. was slowly sinking below the water
   c. was slowly rising to the surface of the water
   d. was lying on the top of the water

8. Where did Dottie and John Peckham spend their vacation?

   a. Vietnam
   b. Los Angeles
   c. Thailand
   d. Hawaii

9. Why did Dottie Peckham put some money into her bottle?

   a. to pay for postage
   b. to bring her good luck
   c. to make the bottle stand up in the water
   d. to buy a ferry ticket

10. When did Hoa Van Nguyen take the bottle out of the ocean?

    a. 1976
    b. 1979
    c. 1982
    d. 1985

11. Why did Hoa Van Nguyen take the bottle out of the water? He wanted

    a. to get the money out
    b. to read the message
    c. to get the piece of paper
    d. to find something to drink

12. What do you expect will happen next in this story?

    a. The Peckhams will go on another vacation
    b. Hoa Van Nguyen will throw the paper away
    c. Hoa Van Nguyen will write to the Peckhams
    d. The Peckhams will visit Hoa Van Nguyen

An Urban Community College

An urban community college generally offers its students a variety of two-year vocational curricula, ranging from nursing to photography to secretarial science. Each two-year curriculum requires approximately sixty hours of credit, or four full semesters of college work. These two-year programs lead to an A.A. degree or certificate from the college. Many students find the vocational programs helpful in broadening their education, since all require some courses such as English and sociology; moreover, such programs are designed to lead to specific jobs. Other students attend community colleges to enter transfer programs, earning credits that may be transferred to baccalaureate institutions for application toward a B.A. degree.

A student who has selected a transfer institution should discuss his plans with his curriculum adviser or counselor. It may also be advisable to visit the transfer institution, study its catalogue, and consult its admissions office. The acceptance of transfer credits is the prerogative of the receiving institution. Because the urban community college offers both two-year vocational curricula and transfer programs, it attracts a variety of students.

13. "prerogative" in line 11 means

    a. choice
    b. requirement
    c. curriculum
    d. purpose

14. What degree is offered by urban community colleges?

    a. B.A.
    b. A.A.
    c. Nursing
    d. English

15. What are "baccalaureate institutions" (in line 8)?

    a. colleges that offer a B.A. degree
    b. colleges that offer a two-year degree
    c. secondary schools
    d. vocational schools

16. What is an example of a vocational curriculum?

    a. English
    b. sociology
    c. photography
    d. B.A. degree

17. Who is the most likely audience for this passage?

    a. high school students
    b. university graduates
    c. college professors
    d. college employers

The Montessori Method

Dr. Maria Montessori was an Italian educator who was active during the first half of the twentieth century. Today there are Montessori schools in a number of countries, including the United States. Most Montessori students are preschoolers, typically three or four years old.

Children who enter a Montessori class for the first time begin work on a wide range of activities related to real life. They discover how to manipulate shoelaces, buckles, and snaps by practicing with these objects mounted on small wooden frames. They learn to serve juice, scrub their hands, clean their work area when they are finished, and move their chairs quietly when sitting or rising. These jobs are not intended solely to teach a youngster domestic chores. "Children experience joy at each fresh discovery," said Dr. Montessori. "Their satisfaction encourages them to seek new sensations and discoveries." Preparation for such tasks is in the spirit of Dr. Montessori's edict: "Teach the importance of doing even the smallest task well." Through expanding abilities gained in these early assignments, children begin to see order in apparent confusion. They begin to acquire the independence that comes with working for oneself. They begin to learn how to start and finish a job. Perhaps most important, they begin to understand what they can do.

18. "manipulate" in line 5 means

    a. to fasten
    b. to make
    c. to pull
    d. to copy

19. Another expression for "domestic chores" in line 8 is

    a. ideas
    b. places
    c. object
    d. feelings

20. "sensations" in line 10 means

    a. ideas
    b. places
    c. object
    d. feelings

21. Dr. Montessori did her most important work between the years

    a. 1950 and 1990
    b. 1900 and 1950
    c. 1850 and 1900
    d. 1800 and 1950

22. Beginning students in Montessori schools learn to perform

    a. woodworking
    b. simple tasks
    c. reading tasks
    d. mathematics

23. The purpose of doing activities such as serving juice, cleaning, and moving furniture is to teach

    a. job skills
    b. the boredom of housework
    c. the joy of discovery
    d. honesty

24. Satisfaction at performing small tasks will encourage the students to

    a. depend on others
    b. avoid difficult tasks
    c. know their limitations
    d. make new discoveries

25. The main idea of the passage is that the Montessori method

    a. was started in Italy
    b. teaches children to do chores
    c. is based on discovery
    d. encourages dependence on teachers

Back to the Basics

Many Americans want their public education system to go back to teaching young people the basic subjects of English literature and composition, mathematics, and science. For two decades, the U.S. educational system has been offering courses in different subjects such as music, art, cooking, personal money management, and filmmaking. Many people believe that this has led to a lack of attention to the basic skills of writing, calculating, and thinking.

During the same twenty years, students' scores on college entrance examinations have been falling, and colleges have been complaining that the students coming to colleges and universities are not as well prepared as they were before. Universities have begun to teach more remedial courses in the basic skills to bring incoming students up to a level where they can do regular university work. University administrators, employers, and many parents are asking the public schools to return to a system with more required courses and fewer optional courses.

There are others, of course, who argue that today's students are just as well-prepared for college as the students of twenty years ago were, and that tests do not measure the kind of learning that students have been getting. They argue that today's students are better prepared to live and work in the modern world than students were before. This second view is not very common, however, and there is now a popular, back-to-basics movement in education in the United States. Many educators hope that this renewed emphasis on basic skills will both prepare students to succeed in universities and prepare them to live in the modern world.

26. "They in line 14 refers to

    a. university administrators and employers
    b. students of twenty years ago
    c. people who argue today's students are just as well prepared
    d. parents who are asking public schools to return to basics

27. "movement" in line 17 means

    a. trend
    b. motion
    c. gesture
    d. improvement

28. Many people in the U.S. think public schools should

    a. offer more optional courses
    b. emphasize the basic skills
    c. teach more remedial courses
    d. include courses in art and filmmaking

29. According to the passage, college entrance exam scores are falling because

    a. students are not being taught basic skills
    b. tests have become more difficult
    c. more students are taking the tests
    d. students are taking too many required courses

30. Although many people believe students are not well prepared for college today as they were twenty years ago, other people argue that

    a. literature is not necessary for success in the modern world
    b. tests measure the kind of learning students are receiving
    c. students do not study hard enough
    d. students are better prepared for the modern world

31. Which of the following is <u>not</u> an example of an assignment in a basic subject?

    a. solving a calculus problem
    b. performing an experiment in Chemistry
    c. singing a song
    d. writing an essay

APPENDIX E. STANDARDIZED ITEM ANALYSIS OF THE FINAL VERSION OF THE NEW READING SUBTEST

ENGL ABDUL-RASHID RESEARCH IEOP READING EXAM SPRING 1996 (CASH

KR-20 RELIABILITY ESTIMATE = 0.76 *** SCORE DISTRIBUTION ***
AVERAGE TEST SCORE = 73%

ERROR VARIANCE = 4.36
STANDARD ERROR OF MEASUREMENT IN RAW SCORES = 2.09
STANDARD ERROR OF MEASUREMENT IN TSCORES = 48.82

NUMBER TAKING TEST = 48
MEAN = 22.75

VARIANCE = 18.31
STANDARD DEVIATION = 4.28
NUMBER OF SCORED ITEMS = 31

SCORE	N	CUM	%ILE	TSCORE	(NUMBER OF ASTERISKS=N)
11	2	2	4	225	|**
12	0	2	4	249		
13	0	2	4	272		
14	0	2	4	296		
15	2	4	8	319	|**
16	0	4	8	342		
17	1	5	10	366	|*
18	1	6	13	389	|*
19	5	11	23	412	|*****
20	2	13	27	436	|**
21	2	15	31	459	|**
22	5	20	42	482	|*****
23	4	24	50	506	|****
24	4	28	58	529	|****
25	6	34	71	553	|******
26	6	40	83	576	|******
27	4	44	92	599	|****
28	2	46	96	623	|**
29	0	46	96	646		
30	2	48	100	669	|**

ENGL ABDUL-RASHID RESEARCH IEOF READING EXAM SPRING 1996 (CASH *** ITEM ANALYSIS *** 2-May-96

OMIT=NUMBER OMITTING THE ITEM DIFF=ITEM DIFFICULTY=% RIGHT # INDICATES CORRECT ANSWER
NA=NUMBER ATTEMPTING THE ITEM DISC=ITEM DISCRIMINATION
NR=NUMBER ANSWERING CORRECTLY =ITEM-SCORE CORRELATION X INDICATES <10% CORRECT

ITEM	OPTION 1/A	2/B	3/C	4/D	5/E	6/F	7/G	8/H	9/I	10/J	OMIT	NA	NR	DIFF	DSCR	ITEM
1	1	43#	2	2	0	0	0	0	0	0	0	48	43	90	0.20	1
2	3	2	7	36#	0	0	0	0	0	0	0	48	36	75	0.62	2
3	45#	0	2	1	0	0	0	0	0	0	0	48	45	94	0.13	3
4	0	1	3	44#	0	0	0	0	0	0	0	48	44	92	0.28	4
5	2	45#	1	0	0	0	0	0	0	0	0	48	45	94	0.39	5
6	18	26#	3	1	0	0	0	0	0	0	0	48	26	54	0.38	6
7	1	1	15	30#	0	0	0	0	0	0	1	47	30	64	0.37	7
8	0	0	1	47#	0	0	0	0	0	0	0	48	47	98	0.13	8
9	26#	17	5	0	0	0	0	0	0	0	0	48	26	54	0.20	9
10	1	4	43#	0	0	0	0	0	0	0	0	48	43	90	0.47	10
11	1	0	1	46#	0	0	0	0	0	0	0	48	46	96	0.38	11
12	1	10	36#	1	0	0	0	0	0	0	0	48	36	75	0.27	12
13	14#	23	4	7	0	0	0	0	0	0	0	48	14	29	0.26	13
14	5	42#	1	0	0	0	0	0	0	0	0	48	42	88	0.27	14
15	37#	4	4	3	0	0	0	0	0	0	0	48	37	77	0.29	15
16	3	4	31#	10	0	0	0	0	0	0	0	48	31	65	0.44	16
17	37#	4	1	5	0	0	0	0	0	0	1	47	37	79	0.50	17
18	8#	36	2	2	0	0	0	0	0	0	0	48	8	17	0.39	18
19	35#	2	10	1	0	0	0	0	0	0	0	48	35	73	0.33	19
20	22	1	2	23#	0	0	0	0	0	0	0	48	23	48	0.42	20
21	4	42#	1	1	0	0	0	0	0	0	0	48	42	88	0.42	21
22	11	35#	0	2	0	0	0	0	0	0	0	48	35	73	0.50	22
23	5	7	34#	1	0	0	0	0	0	0	1	47	34	72	0.46	23
24	1	1	3	43#	0	0	0	0	0	0	0	48	43	90	0.27	24
25	1	20	25#	2	0	0	0	0	0	0	0	48	25	52	0.50	25
26	5	5	35#	3	0	0	0	0	0	0	0	48	35	73	0.40	26
27	22#	7	3	16	0	0	0	0	0	0	0	48	22	46	0.32	27
28	1	40#	6	1	0	0	0	0	0	0	0	48	40	83	0.08	28
29	42#	1	0	5	0	0	0	0	0	0	0	48	42	88	0.24	29
30	1	4	1	42#	0	0	0	0	0	0	0	48	42	88	0.33	30
31	3	4	38#	1	0	0	0	0	0	0	2	46	38	83	0.66	31

APPENDIX F. RELIABILITY ESTIMATES FOR THE ENGLISH PLACEMENT TEST (EPT)

RELIABILITY ESTIMATES

Tables X and Y give internal consistency and parallel forms reliability estimates for Forms A, B, and C. The information in both tables is based on administrations of the test forms to college-aged students of differing proficiency levels who were enrolled in intensive English courses. The tests were administered in October and November of 1977. Each student took two forms of the test. There were either nine or twelve days (including five days of classroom instruction) between the times each student was tested.

Table X: Internal Consistency Reliability Estimates (KR-21)

	SAMPLE			
FORM	N	MEAN	SD	r_{tt}
A	58	53.83	14.72	.89
A	55	58.94	16.65	.92
B	30	48.77	22.36	.96
B	29	54.45	18.96	.94
C	55	54.94	15.90	.91
C	58	54.96	14.32	.89
C	29	48.96	19.38	.94
C	30	54.77	19.15	.94

Table Y: Parallel Forms Reliability Estimates

	SAMPLES			
FORMS*	N	MEAN	SD	r
A	58	53.83	14.72	.89
C	58	54.96	14.32	
C	55	54.94	15.90	.90
A	55	58.94	16.65	
B	30	48.77	22.36	.92
C	30	54.77	19.15	
C	29	48.96	19.38	.95
B	29	54.45	18.96	

*Order of listing represents order of administration.

Source: University of Michigan English Language Institute. (1978). *Examiner's booklet.* Ann Arbor, MI: Author.

REFERENCES

Alderson, J. C. (1979). The cloze procedure and proficiency in English as a foreign language. *Tesol Quarterly, 13*, 219-227.

Alderson, J. C. (1990). Testing reading comprehension skills (part two): Getting students to talk about taking a reading test (a pilot study). *Reading in a Foreign Language, 7*(1), 465-502.

Alderson, J. C. (1993). Judgment in language testing. In D. Douglas, and C. Chapelle, *A new decade of language testing research*. Alexandria, VA: TESOL.

Alderson, J. C., Clapham, C., and Wall, D. (1995). *Language test construction and evaluation*. Cambridge, UK: University Press.

American Psychological Association. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Angelis, P. J. (1990). English language testing: The view from the English testing program. In D. Douglas. (Ed.), *English language testing in U.S. colleges and universities*. Washington, DC: NAFSA.

Aronson, E., Farr, R. (1988). Issues in assessment. *Journal of Reading, 32*(2), 174-177.

Bachman, L. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly, 19*, 535-556.

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford Press.

Bachman, L. F., & Palmer, A. B. (1981). A multitrait-multimethod investigation into the construct validity of six test of speaking and reading. In A. S. Palmer, P. J. M. Groot, & G. A. Trosper (Eds.), *The construct validation of tests of communicative competence*. Washington, DC: TESOL.

Bachman, L. F., & Palmer, A. B. (In press). *Language testing in practice*.

Bachman, L. F., Anderson, N. J., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test. *Language Testing, 9*(1), 142-160.

Baker, D. (1989). *Language testing: A critical survey and practical guide*. London: Edward Arnold.

Bensoussan, M, Goldenblatt, L, & Kreindler, I. (1984). Changing the difficulty level of multiple-choice EFL reading comprehension questions. *Language Testing, 1*(1), 105-109.

Berk, R. A. (Ed.). (1948a). *A guide to criterion-referenced test construction.* Baltimore: Johns Hopkins University Press.

Berk, R. A. (1948b). Selecting the index of reliability. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction*. Baltimore: John Hopkins University Press.

Berk, R. A. (Ed.). (1980). *Criterion-referenced measurement: The state of the art.* Baltimore: Johns Hopkins University Press.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56*, 137-172.

Brennan, R. L. (1984). Estimating the dependability of the scores. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction.* Baltimore: Johns Hopkins University Press.

Brow, F. G. (1983). *Principles of educational and psychological testing* (3rd ed.). New York: Holt, Rinehart and Winston.

Brown, J. D. (1983). A closer look at cloze: Validity and reliability. In J. C. Fisher, M. A. Clarke, & J. Schachter (Eds.), *On TESOL '80 building bridges: Research and practice in teaching English as a second language.* Washington, DC: TESOL.

Brown, J. D. (1984). Criterion-referenced language tests: What, how and why? *Gulf Area TESOL Bi-annual, 1*, 32-34.

Brown, J. D. (1988). Components of engineering-English reading ability. *System, 16*, 193-200.

Brown, J. D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly*, *23*(1), 65-81.

Brown, J. D. (1990). *Testing in language programs.* Unpublished manuscript, Department of ESL, University of Hawaii at Manoa.

Brown, J. D. (1993). A comprehensive criterion-referenced language testing project. In D. Douglas, & C. Chapelle (Eds.), *A new decade of language testing research.* Washington, DC: TESOL.

Carrell, P., Devine, J., & Eskey, D. (1988). *Interactive approaches to second language reading*. Cambridge, U.K.: Cambridge University Press.

Carroll, J. B. (1972). Fundamental considerations in testing for English language proficiency of foreign students. In H. H. Allen, & R. N. Campbell, *Teaching English as a second language: A book of readings* (2nd ed.). New York: McGraw-Hill.

Chapelle, C. A. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research, 10*(2), 157-187.

Chapelle, C. A., & Abraham, R. G. (1994). Cloze method: What difference does it make? In H. D. Brown, & S. T. Gonzo, *Readings on second language acquisition*. Englewood Cliffs, NJ: Prentice Hall.

Chaudron, C., Crookes, G., & Long, M. H. (1988). *Reliability and validity in second language classroom research*. Technical report #8. Social Science Research Institute, University of Hawaii at Manoa.

Cohen, A. D. (1984). *Assessing language ability in the classroom*. Boston, MA: Heinle & Heinle.

Davies, A. (1984). Validating 3 tests of English language proficiency. *Language Testing, 1*(1), 50-69.

Davey, B., & Macready, G. B. (1985). Prerequisite relations among inference tasks for good and poor readers. *Journal of Educational Psychology, 77*(5), 539-552.

Davis, F. B. (1968). Research in comprehension in reading. *Reading Research Quarterly, 3*, 499-545.

Davis, F. B. (1972). Psychometric research on comprehension in reading. *Reading Research Quarterly, 6*, 628-678.

Douglas, D. (Ed.). (1990). *English language testing in U.S. colleges and universities*. Washington, DC: NAFSA.

Douglas, D., & Chapelle, C. (Eds.). (1993). *A new decade of language testing research: Selected papers from the 1990 Language Testing Research Colloquium*. Washington, DC: TESOL.

Dubin, F., Eskey, E. E., & Grabe, W. (1986). *Teaching second language reading for academic purposes*. Reading, MA: Addison-Wesley.

Lewis, C. M. (1987). Vocabulary, sentences and words: Testing for agreement between two recent measures of reading performance and receptive vocabulary. *Educational Psychology, 7*(2), 129-132.

Madsen, H. S. (1990). Standardized ESL tests used in U.S. colleges and universities. In D. Douglas (Ed.), *English language testing in U.S. colleges and universities.* Washington, DC: NAFSA.

Manatt, R. (Ed.). (1985). *Critiquing criterion-referenced measures, part I: Paper and pencil tests.* Ames, IA: Iowa State University, School Improvement Model.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012-1027.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). New York: Macmillan, 13-103.

Perkins, K., & Grutten, S. (1993). A comparison of indices for the identification of misfitting items. In D. Douglas, & C. Chapelle (Eds.), *A new decade of language testing research.* Washington, DC: TESOL.

Perkins, K., and Miller, L. D. (1984). Comparative analyses of English as a Second Language Reading Comprehension data: Classical test theory and latent trait measurement. *Language Testing, 1*(1), 21-32.

Pierce, B. N. (1994). Demystifying the TOEFL reading test. *TESOL Quarterly, 23*(4), 665-691.

Pierce, B. N. (1994). The Test of English as a Foreign Language: Developing items for reading comprehension. In C. Hill, and K. Perry (Eds.), *From testing to assessment*, London: Longman.

Pettit, N. T., & Cockriel, I. W. (1974). A factor study on the literal reading comprehension test and the inferential reading comprehension test. *Journal of Reading Behavior, 6,* 63-75.

Rost, D. H. (1993). Assessing different components of reading comprehension: Fact or fiction. *Language Testing, 10*(1), 79-92.

Shannon, G. A., & Cliver, B. A. (1987). An application of item response theory in the comparison of four conventional item discrimination indices for criterion-referenced tests. *Journal of Educational Measurement, 24*, 347-356.

Shohamy, E. (1984). Does the testing method make a difference: The case of reading comprehension. *Language Testing, 1*(2), 147-170.

Sternberg, R. J. (1991). Are we reading too much into reading comprehension tests? *Journal of Reading, 34*(7), 540-545.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25*, 47-55.

Stevenson, D. K. (1985). Authenticity, validity and a tea party. *Language Testing, 2*(1): 41-47.

Thorndike, E. L. (1917). Reading as reasoning: A study of mistakes in paragraph reading. *Journal of Educational Psychology, 8*, 323-332.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement.* Washington, DC: American Council on Education, 560-620.

Thorndike, R. L. (1973-1974). Reading as reasoning. *Reading Research Quarterly, 11*, 185-188.

Valdman, A. (Ed.). (1987). *Evaluation of foreign language proficiency.* Lafayette, IN: Indiana University.

Wall, D., Clapham, C. M., & Alderson, J. C. (1994). Evaluating a placement test. *Language Testing, 11*(3): 321-343.

ACKNOWLEDGMENTS

I would like to thank the following people who have helped me to complete my Master of Arts degree program and this research study.

First, I am grateful for the guidance provided by my major professor, Dr. Dan Douglas, who has had confidence in my ability to carry out this research work. Thanks also to my committee members, Dr. Barbara Matthies and Dr. Richard Manatt, for their contributions. I am especially thankful to Dr. Barbara Matthies, the coordinator of the Intensive English and Orientation Program (IEOP), for whom I worked as a research assistant to the program and gained valuable experience.

I would also like to thank the test development committee: Dr. Dan Douglas. Dr. Barbara Matthies, Felicity Douglas, and Dan Harness, who spent hours in creating, discussing, assessing, and re-examining the test items. The quality of the study was enhanced by your expertise and dedication.

Thanks also to the two groups of students who participated in the study: the ISU international students and the IEOP students. Also, I am indebted to the instructors of the students who made the gathering of data possible.

I am grateful to the IEOP reading instructors for their reviews and comments on the test items. You made many valuable suggestions, many of which were implemented in the study.

I am indebted to the Testing and Evaluation Center who set up the evaluation for analysis of the data by the computers. I am also sincerely grateful for the tireless efforts of my thesis editor, Pat Hahn, for her valuable suggestions, patience, and the timely manner in which

she processed my manuscript.

Most importantly, I am deeply grateful to my sponsors, the State of Terrengganu and Kusza College, who made my advanced studies at Iowa State University possible. I am also appreciative of the leadership experiences I had through the Iowa-Terrengganu Sister States. Especially, I want to thank Dr. Gary Aitchison and his wife, Kay, Drs. Janet and Ray Heinicke, Ron and Tony Noah, and Sarah Lande—who have invited me into their hearts. I have many experiences and happy memories of the conferences, meetings, trips, exchange students, etc., that kept me quite occupied in a meaningful way throughout my stay in Iowa and the Midwest. I also enjoyed many weekends at the Heinicke's.

I am most grateful to my mother, Rugayah, who has always believed in me and has remembered me in her prayers. To my eldest sister, Roslina, who has been my confidant away from home, and my eldest brother, Abdul Aziz, thanks for your care and concern about me. I especially appreciated all the phone conversations.

Finally, and most importantly, I am grateful to God for blessing me with the abilities to realize my dreams. I know that all things are possible through Him.