A Comparison of SAT Scores of the LEP Elementary School Students in Duval County

Kathryn Llewellyn Branch
University of North Florida

Suggested Citation
Branch, Kathryn Llewellyn, "A Comparison of SAT Scores of the LEP Elementary School Students in Duval County" (1985). UNF Graduate Theses and Dissertations. 36.
https://digitalcommons.unf.edu/etd/36
A COMPARISON OF SAT SCORES OF THE
LEP ELEMENTARY SCHOOL STUDENTS IN
DUVAL COUNTY, FLORIDA

by

Kathryn Llewellyn Branch

A project submitted to the Division of Curriculum
and Instruction in partial fulfillment of the
requirements for the degree of
Master of Education in Elementary Education

UNIVERSITY OF NORTH FLORIDA
COLLEGE OF EDUCATION AND HUMAN SERVICES

July, 1985

Signature Deleted

Dr. Sandra McDonald, Advisor

Signature Deleted

Dr. Paul Egges, Committee

Signature Deleted

Dr. Jim Mittlestadt, Committee
ABSTRACT

The hypothesis that there is no significant difference between the mean SAT Total Reading Battery percentile scores of LEP students who are mainstreamed and those who are self-contained, was studied. Sixty-eight fourth, fifth, and sixth-grade LEP students were administered the SAT. They were divided into two groups. The self-contained group consisted of 35 students; the mainstreamed group consisted of 33 students. The mean percentile score for the mainstreamed group was 39.57 with a standard deviation of 21.02. For the self-contained group, the mean was 26.0 and the standard deviation was 18.53. The computed t = 2.67, p < .05. The null hypothesis was rejected.
CHAPTER I

Introduction

Bilingual education is not new to the United States. As early as the 1800's several states dealt with the issue of English as the language to be used in classroom instruction and as a subject to be taught. In 1870, California passed a law requiring all schools to use English as the language of instruction. German immigrants established German-English bilingual schools in Pennsylvania; French immigrants did the same in Louisiana, as did Italian immigrants in New York and Spanish immigrants in New Mexico. New Mexico had so few "Anglos" in the state that its laws were originally written in Spanish and later translated into English (Leibowitz cited in Jasonov, 1982). The influx of immigrants and issues associated with the relationship between English and other languages continue even today.

The modern revival of public bilingual education in the United States did not originate in the area of foreign language teaching. Instead, it evolved within the context of minority rights. It was within this context that in 1965 the Elementary and Secondary Education Act (ESEA) was passed. The Act provided funds for the planning and implementation of programs "designed to meet the special needs of children of limited English-speaking ability in
schools having a high concentration of such children from families with incomes below $3,000 per year" (Cordasco, 1981). Then, in 1968, Congress passed the Bilingual Act as Title VII of the amended Elementary and Secondary Education Act. Initially, this Act provided funds for non-English speaking students to learn to speak and to understand English. Later, funds were also provided for the teaching of reading and writing skills to these students.

With the passage of time, the guidelines of the Bilingual Act were changed to allow school districts to use any effective approach to teach English to bilingual students, including total English immersion. Former Education Secretary Terrel Bell noted that previous guidelines leaned too heavily on the one approach of requiring school districts to implement "bilingual programs" (Cordasco, 1981).

Congress moved in this same direction when on October 19, 1984, PL 98-511 was approved. This law extends through fiscal year 1989 the authorization of appropriations for certain education programs. Among these education programs is the bilingual program. PL 98-511 also gives parents the option to decline enrollment of their children in bilingual education programs. The law includes special alternative instruction programs within the definition of bilingual education programs. This interpretation allows school districts to use any effective approach to teach
English to bilingual students. The present bilingual program in Duval County, Florida, would come under this category of "special alternative instruction programs."

Duval County's present bilingual program is an immersion Intensive English Language Instruction Program (IELIP). This program relies on entry and exit procedures. Once the Limited English Proficient (LEP) student is identified according to the Duval County Survey of Primary or Home Language Other Than English form, he/she is serviced in one of two ways: either through integration into a regular classroom with one hour of intensive English instruction per day by an IELIP resource teacher, or through placement in a self-contained classroom situation with only other LEP students with five hours of intensive English per day. The type of instruction received is dependent upon the school attended and the grade level of the student. Elementary students are bussed to five IELIP centers throughout the county. Effort is made to keep elementary students in elementary schools. However, there is one class of sixth-grade students housed in a junior high school, and one class of fifth-and sixth-grade students housed in an elementary school designed for kindergarten-through-fourth grades.

Students placed in IELIP must follow the Duval County Pupil Progression Plan and meet the same promotional criteria as their native-speaking counterparts. That is:
1. they must score at or above the 16th percentile on the Total Reading Subtest and the Total Math Subtest of the Stanford Achievement Test (SAT), or

2. they must score 75% on the criterion-referenced, county-developed Essential Skills Test (EST) in Reading and Math; and

3. they must achieve passing grades in Science, Health, Social Studies, and Language; and

4. they must pass teacher judgment, which means they should master at least 85% of required skills.

In March, 1979, the Office for Civil Rights accepted these promotional criteria for Limited English Proficient students in Duval County.

Many factors affect performance results on standardized tests such as the SAT: amount of sleep the students have had the night before; anxiety about tests; attitude toward tests; general health condition; the nature of the testing environment; and so forth. With bilingual students, these factors are compounded by the fact that the tests are written and administered in English. Therefore, what is intended to be an achievement test often becomes a language proficiency test for LEP students. In addition, the norms for these tests are based on a native English-speaking population. Because English proficiency is an integral part of the SAT, it would seem that the LEP students who are placed in the IELIP self-contained classrooms
would perform better than those who attend the resource room program because they receive longer and more extensive language instruction.

As previously stated, these LEP students either are arbitrarily placed in a self-contained IELIP classroom or are mainstreamed and attend the IELIP resource room program part-time. Since performance results on the SAT are used to determine promotion, and since class placement is either mainstreamed or self-contained, it is the purpose of this study to determine if there is a significant difference between the SAT scores of those LEP students who are mainstreamed and the scores of those who are not.

Delimited Problem Statement

The purpose of this project is to determine if there is a significant difference in the SAT Total Reading Battery percentile scores between the 33 fourth, fifth, and sixth-grade LEP students in Duval County, Florida, School District who are mainstreamed and the 35 fifth and sixth-grade LEP students in the District who are self-contained.
Terms
Achievement tests--tests used to measure knowledge, abilities, understanding, or skills acquired from academic work.
Bilingual--students who can communicate effectively in more than one language.
Bilingual education--the teaching of regular school courses in both the national language and a second language.
Criterion referenced tests--tests in which the items are linked to explicitly stated objectives and where the scores are interpreted in terms of these objectives rather than a group norm.
ESL--English as a Second Language, or English as a foreign or non-native language.
ESOL--English for Speakers of Other Languages.
IELIP--Intensive English Language Instruction Program.
Immersion programs--educational programs in which all curriculum materials are taught in a second language.
Limited English speaking--individuals who know English as a foreign language but without sufficient proficiency to participate fully in an English-speaking society.
Mainstreaming--including and maintaining exceptional students in classes with regular or normal students,
with steps taken to see that special needs are satisfied within this arrangement.

Norms--statistical description of the typical performance, behavior, form, function and so forth of a given population.

Resource room programs--part-time programs in which specially trained teachers assist students who, because of their special needs, have been referred by educational professionals.

SAT--Stanford Achievement Test.

Self-contained classrooms--classes having the same teacher or team of teachers for all or most of the daily session.

Standardized tests--tests for which content has been selected and checked empirically, norms have been established, uniform methods of administering have been developed, and which may be scored with a relatively high degree of objectivity.

Test bias--unfairness in the construction, content, administration, or interpretation of tests, either for or against various groups such as minorities, the disabled, women, or socioeconomic classes.
CHAPTER II

Review of Related Literature

This chapter is composed of several parts. The first part deals with a brief history of testing, standardized testing, and definitions of standardized tests. The next part discusses some uses and abuses of standardized tests, particularly for the LEP student. The last part focuses on two common types of instructional organizations for teaching LEP students.

The major portion of this review deals with the abuses of standardized tests for LEP students. It entails a rather in-depth discussion of the various types of biases found in standardized tests. It also discusses a few of the methods used by test publishers to "debias" their tests. This approach is significant in that standardized tests are used to make important decisions about LEP students and their academic careers. A thorough understanding of these abuses then is essential to the purpose of this project.

Standardized Tests

The tradition of testing is centuries old, dating back to at least 2200 BC when the Chinese emperor examined his officials every third year to determine their fitness for continuing in office (Waldrop, 1976). Dubois (cited in Waldrop, 1976) states that the use of examinations in university settings dates back to at least 1219, when formal examinations in law were being conducted at the University of Bologna. DuBois goes on
to say that written examinations had been recognized in England, Europe, and the United States as an appropriate basis for important decisions by the middle of the nineteenth century. With such a long tradition of testing for competence, it is easy to see that the use of testing to make judgments about an individual's competence receives widespread acceptance.

The turn of the 19th century brought the first standardized tests. In France, Alfred Binet developed his intelligence scale which is essentially still used today. Joseph M. Rice, an American, is probably better known for his standardized spelling tests; however, he did develop a standardized arithmetic test and a language test. Between 1908 and 1914, E.L. Thorndike and his students developed tests to measure skills in arithmetic, handwriting, English composition, spelling, language, and reading (Waldrop, 1976). Prior to 1923, all achievement tests were single, subject-matter tests. It was in 1923 that the first standardized survey battery was published. This was the Stanford Achievement Test (SAT) designed primarily for elementary school students. A short time later the Iowa High School Content Examination, the first standardized survey battery for high school students, was published. Since that time hundreds of different standardized achievement tests have been developed (Mehrens & Lehmann, 1975).
While discussing the history of standardized tests, it is important to know exactly what they are. "Standardized tests are characterized by several features: (a) they are designed by specialists; (b) they are administered under standard conditions; (c) they are scored objectively; and (d) they are interpreted with reference to some norm group" (Waldrop, 1976, p. 2). Waldrop further explains that the content-area specialists who design standardized tests have been trained in measurement theory and principles of test development. The conditions under which the test is to be administered are specifically outlined in the instructions for administering the test. These instructions include a description of desired testing conditions; directions to examinees; procedures for answering common questions of the examinee; and time limits for the various parts of the test. Standardized tests are scored objectively which means that equally trained scorers will obtain the same results when scoring the same set of responses. The test norms are based on a sample or norm group. This norm group is chosen to represent the population for whom the test is designed. Geographic regions, socio-economic levels, racial and ethnic minorities, types of community, and any other characteristics seen as relevant are represented in the norm group (Waldrop, 1976).
Mehrens and Lehmann (1975) define standardized tests as "commercially prepared by measurement experts" (p. 3). These tests provide methods of obtaining samples of behavior under uniform procedures. The scoring is objective although essay questions may appear on a standardized test. Usually the test has been normed so that the examinee's performance can be compared with others'. Similarly, Thorndike (cited in DeBlassie, 1974) describes a standardized test in the following terms:

The word "standardized" in a test title means only that all students answer the same questions and a large number of questions under uniform directions and uniform time limits, and that there is a uniform or standard reference group to the performance of which a student's performance can be compared. The term "standardized" does not mean that the test measures what should or could be taught at a particular grade level, or that the test provides "standards of achievement" that students should or could reach at a particular grade level. All that a standardized test does is describe present performance on a uniform set of tasks administered, presumably, under uniform conditions, either for an individual student or the students in a school system. The description is basically in relative terms, that is, in
relation to the performance of a sample carefully chosen to represent some more delimited norm group. But a somewhat more absolute interpretation can be arrived at by examining the specific tasks that pupils are and are not able to handle. (p. 119)

Lastly, Williams and Perrone (cited in Peters, 1979) summarize the definition of standardized tests as "published, norm-referenced measures which are administered under standard conditions to a sample of examinees who are supposedly representative of the population for whom the test is intended" (p. 2). So, although there are many definitions of standardized tests, they vary only slightly, not substantially.

Standardized tests can be classified into three major groups: (a) intelligence tests (mental ability); (b) achievement tests; and (c) special aptitude, interest, and personality tests. While each of these is worthy of discussion, only achievement tests fall within the scope of this project.

Typically, standardized-achievement-test batteries include measures of reading or language abilities, mathematics abilities, social studies abilities, and science abilities. They may also include study skills or "basic skills." Achievement tests are almost always group tests. Another common feature is the use of converted scores such as percentile ranks,
stanines, or grade-equivalent scores. According to Waldrop (1976), "achievement tests are intended to measure the outcomes of exposure to instruction, the skills or knowledge an individual possesses as a result (presumably) of instruction" (p. 44). DeBlassie (1974) describes achievement tests as those that:

(a) serve as a yardstick for pupil and teacher in measuring progress toward proposed goals; (b) point out to the pupil and teacher the degree of efficiency of tasks performed in the various subject matter areas as a result of specific instruction; and (c) indicate, in a diagnostic way, assets and liabilities in the pupil's academic life as they relate to the various subject matter areas. (p. 121)

Similarly, Mehrens and Lehmann (1975) say that any test with a representative sampling of course content and designed to measure the extent of present knowledge is an achievement test. Therefore, achievement tests are basically designed to measure learning. Standardized achievement tests, then, are those that meet the criteria as outlined previously and measure learning.

For the remainder of this paper, unless otherwise stated, the use of the phrase "standardized test" shall mean standardized achievement test.

Uses and Abuses of Standardized Tests

DeBlassie's (1974) definition of achievement tests also encompasses some of the uses or purposes of
standardized achievement tests. More specifically these tests are used for one or more of the following purposes: (a) to monitor pupil progress, (b) to provide information for grouping students, (c) to diagnose a student's strengths and weaknesses, (d) to provide a source of information for curriculum evaluation and planning, and (e) to provide an indication of the effectiveness of instruction (Waldrop, 1976). In addition, Horrocks and Schoonover (cited in DeBlassie, 1974) state these uses for achievement tests: "to predict future success as well as present readiness . . . [and] to provide a basis for selection, promotion, and termination" (p. 126-7). In essence, achievement tests are used to help in decision making. Decisions are made not only about curriculum and instruction but also about the student involved.

Standardized tests are used to identify learning disabilities, eligibility for gifted programs, and a variety of other special programs. Some school districts use standardized achievement tests to determine promotion. For the LEP student, standardized tests are used to determine eligibility for a Title VII program, and eligibility to exit that program. Some school districts use standardized achievement tests to determine if a LEP student will be mainstreamed.
(Bulkin & Sica, 1984; DiMartino et al., 1983; Keyes & Shulman, 1984).

Because standardized test results are frequently used in making decisions about a student, there is a great opportunity for misuse. The most frequent abuses involve using test scores to label or categorize individuals, to make overgeneralizations about individuals or groups, or simply to misinterpret test results (Mehrens & Lehmann, 1975). In recent years there has been great concern over the misuse of the standardized test scores of minority students. The concern arises over the fairness of standardized tests to minority groups. Mehrens and Lehman (1975) found that when achievement tests are used as measures of outcomes of education, few people question their applicability to minority groups . . . but when either achievement or aptitude tests are used as predictors of future success (and therefore as screening devices), the applicability of the tests is often questioned. (p. 341)

Most standardized tests are verbal and presuppose that the test taker has a knowledge of the language, culture, and values of the testmaker. However, tests are based on a monocultural model in that they represent essentially the language and values of only one segment of the American population. That one segment is the white, middle class (Taylor, cited in Peters, 1979).
To the extent that a standardized test is unfair to identifiable groups of the general population, it is said to be biased (Green & Draper, 1972). If the test discriminates against some individuals because of their differences in cultural backgrounds or unique individual attributes, it also is said to be biased (DeAvila & Havassy, 1977). Stated another way, if the test measures different things for different sets of individuals, it is a biased test (Green & Draper, 1972).

However, if tests did not help to make distinctions within and among individuals and within and among groups or classes of people, they would be worthless (Mehrens & Lehmann, 1975). The whole purpose of testing is to discriminate between individuals to show how individuals are different. But this must be done fairly and not at the expense of any one group or groups.

At this point it is necessary to refine the term "test bias". Test bias could be a catch-all term to refer to unfairness of any kind in testing. For the purposes of this review, bias refers to bias in the test itself and not to biases in how the test is used (Shepard, 1982) or bias as a consequence to administration (Green & Draper, 1972).

Test fairness or bias is difficult to define precisely. Even experts in the field cannot agree on a definition that is reasonable without being
contradictory (Cleary, Cole, Darlington, Linn, Thorndike cited in Mehrens & Lehmann, 1975). One of the reasons bias is so difficult to define is that its study is so new and so little is actually known about the nature of bias in tests (Green & Draper, 1972; Lenke, 1982).

In 1975, Green (cited in Peters, 1979) defined a biased test as

. . . generally understood to be [one] that produces results that are systematically unfair to some group. For this to happen, the test must ordinarily measure variables for that group at least partly distinct from those it measures for other people in the population. (p. 5)

Mehrens and Lehmann (1975) state that a test is unfair if it discriminates between races or subcultures and if the differences in scores are not related to what is predicted. In 1953, Eels (cited in Peters, 1979) listed three ways to determine if a test is fair:

(a) it must contain materials common or familiar to all groups to be tested, (b) the language and symbols of the items must be equally familiar to all groups, and (c) the test must be designed to stimulate equal interest and motivation for the various groups. (p. 5)

Another aspect of bias is in the norming. Ideally, the sample group represents a cross section of all groups to be tested. All geographic regions, socio-
economic levels, racial and ethnic minorities, types of communities and so forth are represented. Bernal (1981) charges that minority groups are victims of test abuse because they have not been adequately represented in the sample group. However, no matter how carefully and accurately the sample group is selected, norms in and of themselves require 50% of the examinees to fall below the average no matter how well they perform (McKenna, 1977), or no matter what districts or teachers do to help to improve scores (Shuy, 1979).

There is a difference between test bias as discussed previously and item bias. A test item can be biased but not affect the overall score significantly, thus not causing the test to be biased (Green & Draper, 1972). Test publishers strive to rid their tests of item biases because if too many items are biased the test then can be said to be biased (Green, 1982). There are several kinds of item biases. This next portion of the review will discuss language item bias, culture item bias, and content item bias.

Standardized tests other than math computation are verbal instruments which require reading and symbol manipulation. The language of these tests is frozen, idealized, and artificial. It is a prescriptive, grammatical model which does not represent language as it is spoken in every-day social contexts (Peters, 1979). Wolfram (cited in Peters, 1979) hypothesized
that the more distant the speaking style of natural conversation known to the examinee is from the language of testing, the more the potential linguistic interference there will be for the examinee. Wolfram went on to state that the interference caused for the non-standard English speaker is more serious than for the standard English speaker. There is greater distance between formal standard English and informal non-standard English than formal and informal standard English. Consequently, Rosier (1979) says that these standardized tests are not testing skills and concepts but how well a student manipulates standard English.

Furthermore, according to Kennedy (1972) the examinee's test performance is adversely affected when the language of the test does not match the language development and dialect of the examinee. This is also true when the speech of the tester does not match that familiar to the examinee. The examinee must comprehend to produce. Peters (1979) lists these areas of comprehension as: (a) the literal comprehension of test questions and the comprehension of the task which is demanded; (b) the comprehension of phonological sequences; (c) the comprehension of syntactic (sentence) structures; (d) the comprehension of lexical meanings; and (e) the comprehension of sentence meanings. Test writers assume 100% comprehension of the language of the test (Kennedy, 1972). Yet, Kennedy further states that test instructions, test items, and task requirements are
complex and are sometimes beyond the developmental stages of the examinees. It follows, then, that without 100% comprehension of the language of the test, examinees cannot perform adequately. Therefore, language bias can be defined as the extent to which the examinee's interpretation of test items and instructions does not match that intended by the test writer (Peters, 1979).

A test can be phonologically biased if the pronunciation of the examinee does not match the examiner's. There is a section on the Stanford Early School Achievement Test (SESAT) Level II: Word Reading in which the examinee must identify a dictated word from several words. Minimal pairs such as had/hat, this/these, fur/far are used (Peters, 1979). Unless enunciated very carefully by the tester, these words would be difficult to discriminate especially for a speaker of non-standard English. The Stanford Achievement Test (SAT) has a listening comprehension subtest which consists of a vocabulary meaning section and a paragraph meaning section. If the tester's dialect does not match the testee's, confusion results and performance is adversely affected (J.L. Branch, personal communication, April, 1984). Other phonological biases include the speed of speaking and the quality of enunciation. These affect comprehension of not only test instructions but also test items.
Syntactic bias involves sentence structure used in test directions and instructions. Basic linguistic processes are still being acquired at ages 11 or 12 (Kennedy, 1972). Kennedy further asserts that it is not reasonable to assume that children understand syntax at the same levels as adults, yet these directions are written on a complex, adult level. The instructions must be read exactly as written, however, to maintain the integrity of the test.

The oral directions on standardized tests frequently say "Do not turn this page until told to do so" (Kennedy, 1972, p. 164). This sentence contains three linguistic devices which are difficult for young children to comprehend. These devices are (a) the deletion of "you", (b) the use of the passive voice, and (c) the use of a negative with a temporal conjunction (Kennedy, 1972; Peters, 1979).

Slobin (cited in Kennedy, 1972) studied a group of children from 6 to 12 years old. The study shows that the passive voice took significantly longer to respond to than the active voice sentences. Another study by Beilin and Spontak (cited in Kennedy, 1972) reports similar findings. Ninety-three percent of the subjects correctly responded to active voice sentences, while 73% responded correctly to the equivalent passive voice sentences. When an indirect object was added the scores dropped dramatically: 93% responded correctly
to active voice sentences, 23% responded correctly to passive voice sentences.

The temporal conjunctions such as before, when, and after become particularly confusing when used with a negative. For young children it has almost the same effect as a double negative. Kennedy (1972) found that students respond more rapidly when the order of tasks is in the same order as the action required. In a study by Olds (cited in Kennedy, 1972) unless is more frequently interpreted as if rather than if not. Olds saw this difficulty with interpretation even among nine-year-old children.

Ambiguous test items and instructions are additional examples of syntactic bias. Often multiple choice tests will contain the following instructions: "None of the following are true except." Statements such as this tend to confuse even the best students. Consequently transformational grammarians suggest that the ability to detect ambiguity indicates a person's competence in recognizing grammar rules (Kennedy, 1972).

This discussion of syntactic bias has focused on problems encountered by native-speakers of English. It could be argued then that syntactic bias presents an even greater problem to limited English speakers.

Vocabulary subtests are generally part of standardized achievement tests. Item bias can be seen
in vocabulary subtests also. Vocabulary subtests assume that all children have had exposure to the same content. Peters (1979) points out that vocabulary differences reflect varied needs, experiences, backgrounds, and interests of people. If a person never talks about a particular thing he/she would not need a word for it. The converse is also true as evidenced by the many synonyms for "snow" in the Eskimo language. According to DeAvila and Havassy (1977), it is impossible to determine if minority students have been exposed to a word or if they lack the capacity to understand it.

Standardized reading achievement tests are also subject to biases. Reading is not an isolated function (Shuy, 1977). It reflects language, and language is what is used to describe culture and environment. Therefore, reading tests reflect not only reading skills, but also language manipulation skills, and environment and cultural knowledge (Rosier, 1979). Because of this, it can be argued that LEP examinees are at a considerable disadvantage.

Standardized tests in reading are concerned with sight words in isolation and with phonic and word analysis skills. Second language learners are concerned with developing skills which are more pertinent to reading itself, such as syntactic rules and gaining vocabulary through context (Murphy, 1980).
Murphy further states that second language learners read for different purposes and with different objectives than do native language readers.

The language of standardized tests exerts a powerful effect on the testee. In test situations, differences of interpretation become serious difficulties (Mohon, 1979). Linguistic variables can influence children's test performances (Kennedy, 1972). The multiple choice format penalizes those who think creatively and innovatively because the distractors must be plausible to minimize guessing (McKenna, 1977). Troike (1983) summarized paradoxically when he stated that "... language knowledge and skills may in fact be better assessed by tests not overtly [italics added] designed to test language" (p. 209). Standardized achievement tests are not overtly designed to test language yet that is what they appear to do (Troike, 1983).

Culture bias is still another aspect of item bias. Differences in culture can lead to differences in interpretation on tests (Mohon, 1979). Learning occurs in context, and the learner brings his/her past experiences to the learning situation (Shuy, 1977). Thus, a child who does not have the same socio-cultural experiences as the sample group is affected adversely on a standardized test.

There are obvious examples of culture bias which require a range of cultural knowledge to answer. Test
items may include nursery rhymes or childhood jingles with which minority children are not familiar because of having a different culture. Other items such as questions about customs, holidays, or traditions are particularly discriminatory to ESL students (Mohon, 1979).

Test items which measure the family value system are also culturally biased. Hypothetical situation questions such as "What would you do if you saw someone forget a package in his/her seat on the bus?" require the child to be socialized under the particular ethical situation implied by the question (DeAvila & Havassy, 1977, p. 43). True, LEP students should be tested on cultural knowledge if they are going to live and function in the United States or any other country. But, culture knowledge has no place on a subtest of a standardized achievement test. Culture should be tested separately (Mohon, 1979).

Additional culture bias can be seen in vocabulary subtests of achievement tests. Peters (1979) gives this example: "If a person does something against the law he or she is an: ambassador, offender, official, or officer" (p. 12). According to Peters, all choices may be considered correct depending on the examinee's political and social awareness. Peters gives yet another example when the examinee is asked to pick the best synonym for "inequality--absence, foreign,
difference, similarity or poor." From the minority examinee's point of view "foreign", "difference", and "poor" are clearly associated with "inequality" (p. 13). Often pictures are used on vocabulary subtests for children. Even these pictures require interpretation by the child. The correct interpretation often depends on acceptance of particular cultural conventions (Peters, 1979).

The previous examples of culture bias are rather obvious; however, there are some examples of rather subtle culture bias on standardized tests. Standardized tests assume that all examinees will produce as many responses as they are able in a quick, efficient way. However, in some cultures, students are not motivated in this way. Thus, the level of aspiration is not the same for all examinees (DeAvila & Havassy, 1977). Items on standardized tests are sequenced in order of increasing difficulty. As examinees encounter more difficult questions the level of frustration increases. For examinees from some cultures, at the first indication of failure and frustration, they become discouraged and "give up" (DeAvila & Havassy, 1977).

In summary, there is "no test [that] is language free and no language [that] is culture free" (Rosier, 1979, p. 55). To a large degree, then, achievement tests reflect how well minority children manipulate standard English and middle class cultural concepts in relation
to middle class students. Obviously, those who cannot manipulate English well and have different environmental and cultural backgrounds generally will not do well on standardized tests (Rosier, 1979).

The last type of bias to be discussed in this review is content bias. There must be a match between instructional material and test items to avoid content bias. Content bias and content validity are directly related (Green, 1982). Test publishers strive for high content validity on their tests. Therefore, the degree of content bias should be inversely related to the content validity. A test with high content validity would be low in content bias. In spite of this, Williams and Rivers (cited in Peters, 1979) state that tests have been "...standardized by white researchers on white children and do not involve ethnically relevant content" (p. 3).

Frequently, standardized tests are used to determine the language proficiency of LEP students. Decisions regarding the student's academic career are made on the basis of these tests (Briere, 1969; DeBlassie, 1980; Gross, 1983; Lumallas, 1983). It is not uncommon for school districts to use some type of standardized testing to evaluate students for promotion. Calderon, Cummins, and Larsen-Pusey (1982) charge that the reason for assessing language proficiency is not to determine students' needs, but to
meet standards for graduation and to meet federal and state requirements. It is difficult to assess language proficiency because so few tests are available except in English (Calderon et al., 1982). Regardless of the rationale for assessing language proficiency, it occurs every day. Language skills pervade every area of school (Oller cited in Cummins, 1982).

And language saturates every area of instruction from curriculum guides to final exams (Oller cited in Cummins, 1982; Troike, 1983). Performance on most subject-matter exams requires language competence not only on the actual test but also in the recall of stored information (Troike, 1983). A study by Oller and Streiff (cited in Cummins, 1982) shows that academic and cognitive variables are strongly related to at least some measures of all four language skills—listening, speaking, reading, and writing. Cummins (1982) admits that more research is needed in this area to determine to what extent language proficiency overlaps the constructs of intelligence and academic achievement. Language proficiency seems to account for the greatest variance in educational tests such as verbal and nonverbal IQ measures, achievement batteries, and even personality inventories and affective measures (Cummins, 1982).

According to research by Cummins (1980), it takes immigrant children five years of residence in the host country to approach native norms in conceptual and
literary skills. Calderon et al. (1982) further state that it takes five to seven years to approach age-appropriate academic skills. And it generally takes one and one-half to two years to achieve age-appropriate face-to-face communication skills.

Other studies show that as LEP students advance through the grades, they become more academically retarded (Boyce cited in Morris, 1972; Briere, 1969; Coombs cited in Morris, 1972; Gaarder cited in Morris, 1972; Morris, 1972; Smith cited in Morris, 1972). Morris explains that the emphasis in the primary grades is on decoding skills, and vocabulary and concepts are rigidly controlled. In the intermediate grades, there is a tremendous increase in the difficulty of vocabulary, content and concepts. Academic achievement is stressed, and less time is spent on developing concepts. There is another jump in the level of difficulty when the student reaches secondary school. Only here students do not receive instruction in reading and are expected to read to learn. Therefore, the LEP student is once again penalized. Because of a decreasing emphasis on concept development, the LEP student is left to learn for himself/herself by reading in a second language. Due to the complex nature of the curriculum at the secondary level and the LEP student's language weaknesses, learning becomes more difficult.
How then can language proficiency be assessed? According to Troike (1983) "...in an absolute sense it is clear that language cannot be tested, and an examination of many language tests raises strong doubts that what they are testing is language in a meaningful sense" (p. 215).

In spite of the controversy over standardized testing, it is not going to disappear. If anything it is on the increase. However, if standardized testing is used properly it can be a valuable tool. According to Rosier (1979), an effective and comprehensive evaluation program combines standardized achievement testing and criterion-referenced testing. It must be remembered that no test predicts future behavior perfectly. Test scores are not fixed measures but estimates of attributes (Holmen & Docter, 1977).

Faced with this overwhelming evidence of bias, and the need for standardized testing, test publishers work diligently to rid their tests of bias. In the process they try many approaches and achieve a degree of success in their endeavors. Local norms for the SAT have been substituted for national norms. However, this was found not to have improved competency but rather, to have made mediocrity acceptable (Rosier, 1979). Translating the tests into the native language of the LEP student also creates problems: (a) words are used differently (due to dialects and regional differences); (b) frequently the LEP student cannot read his/her
spoken language; and (c) the spoken language is frequently a combination of dialects and languages (DeAvila & Havassy, 1977).

Other ways to "debias" tests have to be found. The idea of test fairness as a question of validity seems to be an oversimplification, but it is a place to begin (Mehrens & Lehmann, 1975). If test validity were maximized, test discrimination would be minimized. Another way to "debias" tests is to judge the test's content validity through application of linguistic semantics (Mohon, 1979). Care in interpreting test results provides still another method for removing biases from tests. Test results should be cross-checked using supplemental measures of the construct in question (Bernal, 1981). Bernal further states that the valid application of a test assumes that the examinees are similar to the sample group used to develop the norms. So if important psychological differences exist, test results must be cautiously interpreted. However, test publishers felt a need to do something more constructive. Prior to 1966, it was thought that careful attention to content validity was enough to minimize bias in achievement tests (Green, 1982). Publishers have since decided that content validity is not enough to ensure a bias-free test.
The publishers at California Test Bureau (CTB)/McGraw-Hill, publishers of the California Achievement Test (CAT) and the California Test of Basic Skills (CTBS), have devised four steps to eliminate bias in their tests:

1. Careful attention to content validity.
2. The inclusion of bias considerations and the application of various McGraw-Hill guidelines in the test specifications used by the writers and editors.
3. Bias reviews by both CTB editors and by external experts.
4. Analyses of item tryout data separately by ethnic group in order to find and delete items that appear to be undesirable for one or more groups. (Green, 1982, p. 233)

The McGraw-Hill guidelines in the test specifications used by the writers and editors are guidelines for multiethnic publishing, equal treatment of the sexes, and fair representation of disabled people. The outside reviewers represent ethnic minorities and are composed of at least one black and one Hispanic. These reviewers are usually teachers, curriculum experts, or specialists in the field of education for minority students. Sometimes even the experts incorrectly predict how students will react to an item. Therefore, empirical data is gathered on each item. The test is "tried out" on a black sample group,
a Hispanic sample group, and a regular sample group. If an item proves to be undesirable, it is discarded or reworded. This last step Green (1982) calls the latent trait method for debiasing a test. Green can say that "... every approach CTB has used has improved the tests for minorities somewhat by the criteria used; and since these criteria have involved traditional indices of item quality, that improvement must be worth something" (p. 240).

The Psychological Corporation which publishes the SAT is also involved in removing biases from its tests. Tests published prior to 1976 were reviewed only for facial bias, that is, an item appears to be biased, because so little was known about bias at that time (Lenke, 1982). Lenke further explains that the items on the 1973 edition of the SAT were subjected to the National Item Analysis program. Members of this program were minority-group educators. The items were reviewed for curriculum appropriateness, ethnic and racial bias, quality of items for the greatest number of students, adequacy of content coverage, and clarity of item presentation. More than twice the number of items that are usually needed for the final forms of the test were reviewed. Nearly all items which appeared to have some facial bias were eliminated; others were altered slightly to eliminate judgmental criticisms. This system seems archaic compared to the
SAT revision which is presently underway. There are three steps to the item selection process:

1. The items are reviewed by two members of a ten-member advisory panel which represents various ethnic groups.

2. The comments by these panel members are weighed the same as the tryout data. That is, if the panel deems an item biased, it is given the same attention as if the data indicated bias.

3. The application of the latent trait theory is being further explored (Lenke, 1982).

The Psychological Corporation will continue to use subjective and objective ratings of item bias (Lenke, 1982).

In conclusion, as the body of knowledge regarding biased tests grows, test publishers pursue ways of developing tests that will be "fair" to all (Lenke, 1982). It is reasonable to say that current standardized tests are more fair and more appropriate for diverse groups than they used to be. However, there is still room for improvement (Green, 1982).

**Instructional Organizations for Teaching LEP Students**

The grouping of students for instruction has changed dramatically from the old one-room schoolhouse. Today's schools are organized as to: (a) a graded or non-graded organization, (b) a homogeneous or heterogeneous organization, (c) a team or platoon
teaching organization, and (d) a departmentalized or self-contained organization (Shepherd & Ragan, 1982).

The relationship between student learning, instructional practice, and the organization of the school is statistically unstable (Shepherd & Ragan, 1982). Proponents of each plan claim theirs is "best." However, it must be remembered that "no plan or organization can, of itself, improve instruction; the most it can accomplish is to provide a framework" (Shepherd & Ragan, 1982, p. 50) for instruction. Since the focus of this project is on bilingual students who are placed in self-contained IELIP classrooms and those who are mainstreamed, only these two organizations will be discussed in this review.

A self-contained classroom can be described as one in which a group of students is placed with one teacher for the majority of the school day (Shepherd & Ragan, 1982). In a nationwide survey of 2,318 elementary school principals reported in The self-contained classroom (cited in Shepherd & Ragan, 1982), more than 95% used the self-contained approach in the primary grades. The figures dropped in the intermediate grades: 88% used self-contained in grade four; 80% in grade five; and 71% in grade six.

As teachers have become more aware of the individual needs of their students, programs have been developed to help meet these needs. Students identified as gifted,
learning disabled, emotionally handicapped, or speech impaired, receive additional educational services in these programs. The students remain in their regular classroom for most of the day but are "pulled out" to receive this extra instruction. This organization is called a "pull-out" program (Bernal, 1981) or a resource room program. It actually is a type of mainstreaming since these students remain in the regular classroom, the least restrictive environment, and are removed for only short periods of time each day for special instruction.

These two organizations are also used for LEP students. Frequently, the initial phase of a bilingual program is characterized by placement of LEP students in a self-contained classroom with other LEP students only. Sometimes LEP students are mainstreamed for nonacademic subjects such as art, music, or physical education. In a truly bilingual program, the LEP students receive academic content instruction in their native language and in English. Even so, Troike (1978) states that often English is spoken as much as 67% of the time.

The immersion approach is another type of self-contained classroom for the LEP student. It is characterized by "a carefully sequenced system of content instruction which helps children intuit the language as classes progress" (Cohen; Lambert & Tucker cited in Bernal, 1981, p. 24). The students
remain in these self-contained situations until their English proficiency enables them to be fully mainstreamed (Bulkin & Sica, 1984; DiMartino et al., 1983; Keyes & Shulman, 1984).

While some may argue that self-containing LEP students amounts to segregation, Finocchiaro (1972) states that temporary homogeneous grouping which accelerates admission to a regular program should not be viewed as segregation. A worse segregation is to have LEP students uninvolved in a regular classroom. Finocchiaro goes on to state that homogeneous, self-contained classroom grouping is the only viable organizational pattern at the present time.

LEP students who are mainstreamed are placed in a regular classroom as if they were native speakers. They are then "pulled out" to receive English language instruction or tutoring by the ESL resource teacher (Bernal, 1981). The size of the group varies depending on the number of students enrolled in the program. The groups usually have no more than 10 to 12 students (Bonn, 1979). They meet for 40-60 minutes per day; those with the least command of English often meet twice a day (Bonn, 1979). Finocchiaro (1972) is critical of the "pull out" or resource room programs. She states that unless they are carefully planned they: 
"(a) do not provide long enough periods of intensive help, (b) do not insure continuity of instruction, and
(c) generally do not make integration of what English they [LEP students] have learned into what is needed in the content areas" (p. 133).

The bilingual program in Duval County, Florida, utilizes both the self-contained IELIP classrooms and the resource room program. There are only three self-contained classrooms in the County, and these were set up so as to limit the bussing of LEP students. There has been no research conducted in Duval County to determine which organization provides a more positive learning environment.

This lack of research is a weakness in bilingual programs across the nation (Troike, 1978). Over $500 million were spent on bilingual education between 1968 and 1978, with less than one-half of one percent being spent on research (Troike, 1978). The information that is available comes from doctoral dissertations, small-scale studies, or program evaluations. Troike (1978) states that the vast majority of program evaluations are worthless because they contain useless information. He lists their shortcomings as:

1. no control for socio-economic status;
2. inadequate sample size, improper techniques, or excessive attrition rate;
3. no baseline comparison data, no control group, no non-relevant comparison;
4. no control for initial language dominance;
5. significant differences in teacher qualifications or characteristics, or other confounding variables;
6. insufficient statistical information or improper statistical applications; and
7. for research reports, lack of immediate relevance, new data, or accessibility. (p. 4)

Despite the lack of research and the inadequate evaluation reports, Troike (1978) states that enough evidence has accumulated to show that bilingual education can be effective. He goes even so far as to say that if the program is not producing positive results, something is wrong and changes need to be made.

There are factors which influence the progress LEP students make in bilingual programs. According to Gross (1983) these factors are (a) diversity of cultural and educational backgrounds; (b) variation in amount and quality of prior formal education; (c) emphasis in the home on schooling and good behavior; (d) diversity of socio-economic backgrounds; and (e) family stability. Troike (1978) also states that children who succeed in the immersion programs are for the most part middle-class, from supportive homes whose language and culture are in no way threatened nor demeaned by the children being taught another language.
Summary

Tests have been used for a long time to determine the competence of individuals and to make decisions about them. Standardized tests were developed to help make distinctions between individuals so more that accurate decisions could be made. However, the very nature of standardized tests has caused some distinctions to be made unfairly. As testers have become more aware of these biases, steps have been taken to eliminate the biases. There are so many factors which affect performance on a standardized achievement test that it is extremely difficult to devise the perfect test that is fair to all.
CHAPTER III

Procedures

The purpose of this project is to determine if there is a significant difference in the SAT Total Reading Battery percentile scores between the LEP students in Duval County, Florida, School District who are mainstreamed and those in the District who are self-contained. SAT scores are part of the promotion criteria for elementary school students in Duval County. They are used also to determine eligibility of a LEP student to exit the IELI Program. Some LEP students are placed in self-contained IELIP classrooms while some are mainstreamed. Therefore, it is important to know if there is a significant difference between the SAT percentile scores of these two groups because important decisions are made based on these scores.

The population was those LEP elementary school students in grades four, five, and six who were enrolled in the IELI Program in Duval County. The LEP students who were participating in a pilot program which used computer assisted instruction were not included. The population was 68 students. There were 35 students who were self-contained: 13 sixth graders at John Gorrie Junior High School, and 22 fifth and sixth graders at San Jose Elementary School. Of the 33
students who were mainstreamed, there were three sixth graders, seven fifth graders, and 23 fourth graders.

Although a study which would have included the total population of LEP elementary school students would have carried some strength, for the purposes of statistical analysis, only fourth, fifth, and sixth grade LEP students were studied. Fifth and sixth grades are the only elementary grades that are self-contained. In addition, a study of fourth, fifth, and sixth grades provides a baseline for future studies of mainstreamed and self-contained LEP students.

The students were divided into four ethnic groups: Cambodian, Vietnamese, Spanish, and Others. The self-contained population included 17 Cambodians, 4 Vietnamese, 5 Spanish, and 9 Others. The mainstreamed population included 11 Cambodians, 6 Vietnamese, 6 Spanish, and 10 Others. The socio-economic level of both groups was low as evidenced by over 80% participation by both groups in the federally funded lunch program.

The instrument used was the Stanford Achievement Test (SAT). LEP students appear to have less difficulty on the math subtests, especially Math Computation, than on the reading subtests. Their Total Math Battery percentile scores are closer to the national average than their Total Reading Battery percentile scores. The Reading Battery requires extensive language manipulation while the Math Battery contains subtests which require very little language
manipulation. Since the focus of this project is on second language learning, only the Total Reading Battery was be considered when figuring the mean percentile scores for each group.

The mean Total Reading Battery percentile score for each group was subjected to a t test analysis. The null hypothesis was that there is no significant difference between mean SAT Total Reading Battery percentile scores of LEP students who are mainstreamed and those who are self-contained. The null hypothesis was rejected if the computed t was significant at the .05 level.
CHAPTER IV

Results

The mean and standard deviation for each group were computed. For the mainstreamed group, $\bar{X} = 39.57$ and $S = 21.02$. For the self-contained group, $\bar{X} = 26.0$ and $S = 18.53$. For the mainstreamed group, $S^2 = 441.88$, while for the self-contained group, $S^2 = 343.24$.

Further analysis with a two-tailed t test indicated the null hypothesis must be rejected with the computed $t = 2.67$, $p < .05$ with $t .025, 66 = 1.96$ as the point of rejection. Therefore the two groups' means are significantly different.

Table 1. Mean Scores and Standard Deviations for LEP Groups

<table>
<thead>
<tr>
<th>GROUPS</th>
<th>N</th>
<th>$\bar{X}$</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainstreamed</td>
<td>33</td>
<td>39.57</td>
<td>21.02</td>
</tr>
<tr>
<td>Self-contained</td>
<td>35</td>
<td>26.0</td>
<td>18.53</td>
</tr>
</tbody>
</table>
Figure 1. A Normal Distribution of Stanines, Percentile Ranks, Normal Curve Equivalents, and Performance Classifications. From Technical Manual by The Psychological Corporation, 1983, p. 12.
CHAPTER V
Discussion and
Recommendations For Further Study

Discussion

The results of the t test indicate rejection of the null hypothesis. Therefore, it can be said that there is a significant difference between the mean SAT percentile scores of those LEP students who are mainstreamed and those LEP students who are self-contained. Care must be taken to interpret the t as only rejection of the null hypothesis, and not rejection of the self-contained instructional organization. Any decisions to alter the IELI Program, based on this study, without consideration of other factors, or further study, would be inappropriate.

As discussed in Chapter I, placement of students in self-contained or mainstreamed classrooms is dependent upon the school attended and the grade-level of the student. As cited in Chapter II, self-contained classrooms were set up primarily to limit the bussing of LEP students. To date, no research has been conducted in Duval County to determine which organization provides a more positive learning environment. There are advantages to both types of organizations (Shepherd & Ragan, 1982)
as cited in Chapter II. Those for mainstreaming cite exposure to native-speaking classmates as being an advantage. Conversely, those for self-contained cite the more intensive English instruction as an advantage. Arguments for self-contained organizations include enhancement of the affective domain, more rapid acculturation, more flexibility in scheduling, and more time to spend meeting individual needs. The argument for both organizations will continue to rage with proponents claiming theirs is "best." It must be remembered that "No plan or organization can, of itself, improve instruction; the most it can accomplish is to provide a framework" for instruction (Shepherd & Ragan, 1982, p. 50) as cited in Chapter II. However, Finocchiaro's (1972) work and criticisms of the resource room program which are cited in Chapter II, must not be overlooked.

As stated in Chapter III, it would seem that those LEP students in a self-contained classroom would score higher on the SAT than those who are mainstreamed because they receive more extensive English instruction. However, it can be argued that those LEP students who are self-contained have more opportunity to use their native language with classmates than those who are mainstreamed. In a LEP self-contained classroom, all students are LEP
and they usually have at least one other classmate who speaks their language. Conversely, a mainstreamed LEP student may be one of two, or at the most, three LEP students in the class and has less opportunity to speak his/her native language. It follows then that a mainstreamed LEP student may have an advantage in learning English because he/she is forced to use English more frequently during the school day.

In addition to use of native-language, there are other factors which may contribute to the significant variance between the groups. To name but a few: (a) length of time in the country; (b) length of time in the IELI Program; and (c) previous academic experience (Gross, 1983) cited in Chapter II. Of further significance is the fact that of the 13 self-contained sixth-graders at John Gorrie Junior High school, nine are Cambodian and one is Vietnamese. Often the Cambodians and Vietnamese are refugees and have been deprived of formal schooling due to the political unrest in their countries.

Prior to their arrival in the United States, they have been in detention camps for sometime and their formal education had been neglected. Also, for some reason, the newly-arrived Southeast Asian refugees seem to settle in the John Gorrie area. Therefore, it can be argued that
these students have had less formal education and have acquired less language facility than their counterparts at San Jose School. This, of course, would affect the SAT scores.

In addition to all of these reasons which could explain the significant variance of the two groups, it is important to take a closer look at the statistical analyses of the two groups. In Chapter IV, the mean for the mainstreamed group was reported as 39.57, and the mean for the self-contained group as 26.0. The difference between the two means is so great that it indicates significant variance. However, before this conclusion can be reached, it is necessary to study Figure 1. (Psychological Corporation, 1983, p. 12) It is interesting to note that while the means for the two groups appear to be significantly different, according to Figure 1., both means fall within the range of the fourth stanine. This would indicate that perhaps the difference is not so significant after all. True, the fourth stanine is at the low range of average, but considering the language deficiencies of these LEP students, it is within realistic expectations. So, while the means are significantly different, they both fall within an acceptable stanine range, which would indicate that both organizations are effective.
To summarize, the null hypothesis was rejected. Care must be taken not to interpret this rejection as rejection of the self-contained instructional organization. Many factors contribute to the significant variance of the two groups. In spite of the significantly different means, both fall within the range of the fourth stanine. This would indicate that perhaps the difference is not so significant after all.

**Recommendations for further study**

It is apparent that further study is called for. The lack of research nationwide (Troike, 1978) as cited in Chapter II, is sufficient cause to justify further study. The results of this present study are inconclusive and would indicate further study, also.

As discussed in Chapter III, a study of the total population of the LEP students in Duval County would carry some weight. However, the great majority of LEP students are mainstreamed, with only three classes throughout the County being self-contained. This would make the numbers of the two groups disproportionate, and would again prohibit decisive conclusions from being drawn.

Perhaps a more accurate study would be to monitor individual progress for several years. That is, records
would be kept on each student as he/she progressed through the Program. Specific attention would be given to improvement in SAT scores each year. As the body of data was gathered, comparisons could be made between the gains of students in self-contained or mainstreamed organizations. This would also allow for comparison of scores for those students who are new to the country and Program who happen to be placed in a mainstreamed or self-contained organization. A study such as this would yield more accurate data and permit more decisive conclusions to be drawn.

A matched-pair study might also be conducted. The samples would be smaller, but the variables would be more tightly controlled, so the results would be more conclusive.

In conclusion, the results of this study further support Troike's (1978) statements as cited in Chapter II that there is enough evidence to show that bilingual education is effective, however the lack of conclusive research continues to be a weakness of the program.
References


Peters, C.A. (Aug., 1979). Some common problems related to testing minority populations with


