DOCUMENT RESUME

ED 341 266 FL 020 036

AUTHOR De Avila, Ed
TITLE Assessment of Language Minority Students Political, Technical, Practical and Moral Imperatives.
SPONS AGENCY Office of Bilingual Education and Minority Languages Affairs (ED), Washington, DC.
PUB DATE Sep 90
NOTE 54p.; In: Proceedings of the Research Symposium on Limited English Proficient Students' Issues (1st, Washington, DC, September 10-12, 1990); see FL 020 030.
PUB TYPE Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Elementary Secondary Education; Eligibility; English (Second Language); Evaluation Methods; *Language Proficiency; *Language Tests; *Limited English Speaking; Moral Issues; Second Language Learning; *Student Evaluation; *Student Placement
IDENTIFIERS *Language Minorities

ABSTRACT

This paper reviews several issues and problems associated with the creation and application of tests and decision models for determining entry/exit, eligibility, placement, treatment, and reclassification procedures used to remedy the limited English proficiency of students from homes where English is not the primary language. Two concepts critical to the assessment process are discussed: language dominance and language proficiency. It is argued that the concept of language proficiency is not only more linguistically and scientifically sound, but also more amenable to mathematical or statistical manipulation because of the known properties of the test score distributions. A number of the ideas used in this argument are used to review some of the problems with current testing practices related to eligibility, placement, and reclassification. It is suggested that the failure to work from a common set of definitions and principles has compromised not only the process of entry/exit but also both the evaluation of Title VII programs and research on the effects of bilingualism. The moral issue of assessing students is discussed. An operational definition of limited English proficiency is given that includes the following components: limited English proficiency student, comparable students, language proficiency, and probability of success. Contains approximately 75 references. (LB)

ASSESSMENT OF LANGUAGE MINORITY STUDENTS POLITICAL, TECHNICAL, PRACTICAL AND MORAL IMPERATIVES

Ed De Avila

INTRODUCTION

Language proficiency testing of students from non-English speaking backgrounds has, in the past fifteen years since the 1974 Lau v. Nichols Supreme Court decision, undergone considerable change, sparked enormous debate and created the need to re-examine our approach to language testing in general.

Identification, Placement, Treatment and Reclassification

The process of identifying children as eligible for special language services, placing them in programs, providing services and finally reassigning them to mainstream classrooms derives from several different sources including federal and state law and has been referred to in various ways. The term "entry/exit criteria" has been used to describe the assessment process used in identification, classification, and reclassification. The term "reclassification" refers to the process of relabeling, with the possibility of return to a "mainstream" or "all-English" classroom. Thus, for example, a student initially identified as "Limited English Proficient" (LEP) may be "reclassified" as "Fluent English Proficient" (FEP). However controversial or diverse the approach, the fact remains that assessment (both formal and informal) of language proficiency is found at every step.

Some Common Underlying Problems

Several aspects underlying the process of identification, placement, treatment and reassignment have been particularly problematic. First, definitions driving the process have not been very clear or consistent. Second, there has been a shortage of appropriate assessment devices which are both psychometrically and linguistically sound. Third, perhaps because of these two problems, testable or viable decision making models consistent with the need to serve children who do not understand the language of the schools have not been forthcoming.

A major difficulty in the development of entry/exit assessment models and instruments stems from the fact that the lay theory underlying the concept

of language proficiency has not been particularly well defined. While a number of recent shifts have occurred in our views regarding language proficiency testing (O'Malley,1989), our understanding of language proficiency as a scientific construct has not been fully operationalized.

It is well known in scientific investigation that if the elements of a given phenomenon are not separately defined, it becomes difficult to operationalize the phenomenon as a whole, to understand the relative importance of its elements and to use the results of scientific investigation in a constructive fashion.

In this regard, current legislative definitions of language proficiency based on an oral proficiency/academic achievement sequence may confuse effect with cause and have contributed to the current state of affairs. Oral proficiency results from the continuous interaction of child and linguistic environment. It is the result of both nonformal and formal instruction. Academic achievement, on the other hand, results from an organized presentation of material normally imparted, often orally, by somebody trained to do so. Howard (1983, p. 257) describes this distinction in terms of "explicit and implicit" acquisition and knowledge of rules.

Oral language proficiency may be seen as providing the necessary (although not sufficient) conditions for the development of literacy skills. The notion that oral skills are an integral part of the development of literacy was a guiding force in the lay theory used in the Lau vs. Nichols decision, which provided the impetus for testing students from homes where English is not the primary language. In this connection, however, De Avila and others (1978) pointed out that "many second and third generation language minority groups demonstrate 'survival English' — that is, they score as English fluent on many language screening tests but perform poorly on achievement tests." More directly, De Avila and others stated, "Not all children fluent in English achieve at the norm." Finally, in a large study involving nine different language minority groups, De Avila and Duncan (1982) concluded that while oral "language (proficiency) in and of itself is not a sufficient condition for thinking (as defined by intellectual development) it does seem to be critical for school achievement."

De Avila (1987) has argued that "while oral language proficiency seems to be a necessary condition for success in the mainstream, it is no guarantee." Similarly, Cummins (1984) argued that assessment of simple communicative skill on the assumption of educational sufficiency is misguided. De George (1988) has even more recently commented that "the notion that language skills (oral) alone may not be sufficient for a student to acquire content-area knowledge has caused considerable rethinking among administrators and teachers about assessment procedures as they exist today." Finally,

the need to reconsider our approach to the problem has been echoed by O'Malley (1989), who called for an "attendant shift from identifying isolated language skills to gaining a broader understanding of a student's ability to convey meaningful utterances through speech and writing." Davies (1959) implied much the same thing over thirty years ago when he outlined the advantages and disadvantages of "integrative" and "discreet-point" testing.

Unfortunately, however, legislative interpretations of lay theory have often failed to include the distinction between necessary and sufficient conditions. In efforts to comply with federal, state and local requirements for identification, placement and reclassification, districts have been left to their own devices in the selection of instruments and procedures for combining reading and writing and listening and speaking scores to make placement decisions. This has led to a wide diversity of approaches some of which unknowingly create more problems than they solve. Many districts, for example, use currently available, standardized norm-referenced tests of school achievement to fulfill federal and state requirements for assessment of reading and writing. Some have gone so far as to use an oral language proficiency tests to assess academic achievement (See EAC-WEST Newsletter, vol. 3, no. 1). There are also a number of districts or states that exempt LEP students from any form of achievement testing. On the one hand, the use of nationally norm-referenced tests of academic performance carries a number of problems. On the other hand, failure to conduct any assessment at all is also problematic. More will be said of this later.

Both informal and formal techniques are used for the assessment of oral proficiency. Informal procedures consist primarily of interviews and/or rating scales based on observations by the teacher or some other person. Formal techniques include more direct testing of specific language skills, abilities and behaviors within a standardized environment. My own view is that the distinctions between informal and formal procedures have more to do with issues and differences in practice than with differences in substance. In other words, observational techniques, rating scales and the like are subject (or should be) to the same rigors (psychometric considerations) as any formal test. The fact that a process is informal in no way relieves it of its scientific burdens. Without establishing both validity and reliability, rating scales are little more than self-serving arguments for face validity.

Problems with Informal Approaches

A number of writers have argued, perhaps wisely, that standardized approaches to the assessment of language proficiency should (must) "be accompanied by teachers' own judgments and observation data" (Canales, 1990). Unfortunately, however, we are left to our own devices on how to combine these data, and arguments for the use of informal procedures and multidimensional indicators have not been systematically implemented. The

question becomes how do you make both informal and formal procedures work together systematically to everybody's satisfaction?

Hige and Coladarci (1989) reviewed sixteen published articles on the relation between teacher ratings of student academic performance and performance on standardized tests. According to these writers, the correlations between the student ratings and test scores ranged from .22 to .92 with a median of .66. The authors interpret these coefficients as moderate to strong and contend that there is sufficient reason to re-evaluate their use in research and instructional decision making.

The danger in this type of study is twofold. First, authors tend to overinterpret correlational data. A correlation coefficient does not a finding make. In other words, there is a good deal more to the establishment of validity than a single correlation coefficient. A median correlation of .66 is not particularly strong. It suggests that, on the average, we can expect to find that teacher judgment accounts for only forty-three percent (the square of .66) of the total achievement test variance. Given the importance of placement decisions, it would be dangerous to conclude that teacher ratings should totally replace formal assessment. Second, the article by Hige and Coldarci leaves the impression, perhaps unintended, that their findings can be applied to the rating of language proficiency. It is important to recognize that the studies reviewed in the article refer to academic achievement and not to language proficiency. The same criticism has been applied to studies by Ulibarri et al. (1981), Mace-Matluck, Dominquez and Turner (1979) and Jackson (1980). De Avila (1984) found that teacher ratings of language proficiency were influenced by a number of factors such as the teacher's language background, attitudes toward bilingual and language minority programs and so on.

Given the present context in which test scores (or teacher observation ratings/scores) are used for placement, the use of correlations to establish "validity" is probably inappropriate. Correlations are measures of association not measures of agreement. The issue is one of agreement between two methods of identification, not one of association or pattern as is represented in a correlation coefficient. A more appropriate statistic would be something like Kendall's Coefficient of Concordance or a simple Chi-Square test (see Kerlinger, 1973, pp. 290-293). Although correlations are useful and interesting, they can sometimes be misleading.

Finally, there is nothing inherently right or wrong with using a teacher's judgment of language proficiency. Aside from costs and other logistic issues, the problems are in demonstrating standardization, reliability and validity and the possible placing of teachers into a conflict of interest should test results in any way affect their employment status. See Gruein and Maier (1983) for more detailed discussion on the uses of nonformal assessment techniques.

Problems with Formal Approaches

When De Avila and Duncan (1976) reviewed the available oral language tests in 1976, they found great diversity. Tests reflected a wide variety of purpose, content, approach, method, standardization, validation and utility. In fact, De Avila and Duncan had originally intended to review the psychometric properties of available tests. However, they found such diversity as to make comparisons almost impossible. A number of issues that surfaced in 1976 are still with us and apparently require further discussion. As will be seen, this discussion necessarily leads to a general consideration of the concept of measurement as applied to language testing.

The most pervasive problem discussed by De Avila and Duncan had to do with definitions resulting from the differences in perspective between legislative and scientific points of view. The problem of how to define "limited English proficiency" is as much an issue today as it was then. In an attempt to provide policy direction for schools faced with the education of large numbers of language minority students, the U.S. Office of Civil Rights prepared a set of administrative guidelines known as the "Lau Remedies," which were circulated throughout the United States to assist districts in determining the linguistic proficiency of language minority students (Crawford, 1989).

The problem was, as we argued in 1976, that the linguistic categories referred to in the OCR document bore "no resemblance to operational definitions found in the sciences. . . "What this means, unfortunately, from the point of view of a researcher, is that there is no clear way of deciding how these categories apply to actual behavior. . . ." (De Avila & Duncan, 1976). Furthermore, the Lau Remedies offered no discussion on how to combine reading and writing assessment results with oral results in order to assess language proficiency fully, and for a number of years their assessment has been largely ignored. In other words, the Lau Remedies offered no clear way to operationalize the construct of language proficiency through tests. We maintained then, as we do now, that attempts to develop tests on the basis of judicial and/or legislative recommendations, without consideration of technical issues, were bound for controversy at best and outright failure at worst. History seems to have borne us out as many of the tests developed specifically from the Lau categories have dropped from use.

De Avila & Duncan's review concluded that:

1. Different tests measured different aspects of language (i.e., phonological, lexical, syntactic or pragmatic).

2. Tests were inconsistent in the way they were developed, validated, normed or used.

3. Few tests were based on clear explicit theory regarding either purpose or method.

4. Many tests were based on default definitions or theories of language development that rendered measurement virtually meaningless or impossible.

In 1976 little more could be done than to list available tests according to what the test developers claimed to be measuring. However, this led to the use of a set of nominal (descriptive) categories that was itself problematic. Listing tests according to what developers and publishers claimed to be measuring uncovered a second major definitional problem. Of the then available tests (forty-six were reviewed), fifteen were classified as (or claimed by the publisher to be) "language dominance" tests; twenty-three were "language proficiency" tests; and seven claimed to measure both "dominance" and "proficiency."

In searching for clarification of the terms "dominance" and "proficiency," De Avila and Duncan found little in the literature to make the distinctions between them clear. On the one hand, they found a great deal of discussion on the concept of language proficiency. On the other hand, they reviewed over forty major texts and research references and found virtually no mention of the concept of language dominance. In other words, there was little discussion to be found in the research or theoretical literatures to defend the term as it was being used in practice.

It is noteworthy that the vagueness of the categories listed by the Lau Remedies and the confusion over dominance versus proficiency led to attempts to define and to place students in or out of programs without regard for their actual ability to speak English or the home language (Dulay and Burt, 1980). Dulay and Burt contended that identification and placement could be made primarily on the basis of "language dominance" or the stronger language. Barnes (1979) went even further and used a Language Dominance Index to show that the number of language minority students in need of help was less than a third of that reported by O'Malley (1978).

In more recent discussions of the process of identifying students one continues to find reference to the idea of "language dominance." One state department of education refers directly to a determination of language dominance as the second step in the process of student classification and placement. These points regarding the problems that surround the identification process

could not be made more clearly than in the recent Chapter I Rules and Regulations, in which Chapter 1 applicants are admonished to identify eligible students by using factors such as teacher evaluation of student performance, language dominance tests in combination with other indicators that may be used separately, as a composite score or as a composite with weighing to select children on a basis other than English language deficiency (Rayford, et al., 1990).

Attempts to clarify the Chapter 1 Regulations (see Rayford et al.) seem, unfortunately, equally misdirected in that the recommended models continue to be based on the presumption that a test of dominance produces a clearcut dichotomy between the home language and English. This has the effect of leaving the second and third generation child (e.g., the "pocho" or "Spanglish" speaker) most at risk since test scores show many of these students to be limited in both the home language and English. According to De Avila & Duncan, while a test of language dominance may be a convenient way to satisfy the legal demands of Lau (or a quick way to meet Chapter 1 rules), it tells us nothing about the specific needs of an individual child. A student who ranks in the seventy-ninth percentile in English and the sixty-fifth percentile in Spanish is easily classified as English dominance. The real truth is that the child may have problems in both languages. The failure to distinguish between dominance and proficiency has been pervasive, particularly in research on both the effectiveness of bilingual education (Baker & De Kanter, 1973; Danoff, 1977; Rosell, 1989; Willig, 1985) as well as on the cognitive effects of bilingualism (De Avila, 1987).

The problems with both the original Lau categories and the concept of language dominance should be obvious. The former lacked adequate definition/operationalization whereas the latter lacks a normative base on which to make comparisons. Unfortunately, much of the original thinking that went into the formulation of both the categories and the idea that languages were at war and that one would dominate the other, is still with us. This point will become even clearer in the following discussion.

There is certainly a need to appreciate the relative strength of the child's home language relative to English (De Avila & Duncan, 1980). The primary weakness in a "dominance" interpretation, however, is its lack of a normative base. In other words, it is important to consider the child's proficiency in the home language relative to English as well as the strength of both in relation to proficient speakers of each language. In more technical terms, there is a need to consider language proficiency both within (the individual student) and between students (the individual student and other students) as discussed in classical analyses of variance designs (see Kerlinger, 1973).

What De Avila and Duncan concluded in their 1976 review was that the concept of "language dominance" was driven more by socio-political education concerns and a desire for a quick way to categorize students as either English or home language dominate than by either linguistic or understanding of student education needs.

In his recent review of current trends in language proficiency testing, O'Malley has listed four major problems with most tests. O'Malley's list bears great similarity to the concerns voiced by De Avila and Duncan over ten years previously. It would appear that change is slow. O'Malleys criticism of current tests falls into two broad categories. The first has to do with the question of test content. O'Malley argues, like many others, that there is a need to test more than simple oral/aural proficiency. He goes somewhat beyond others in arguing that, in addition to the assessment of the four language skills (listening, speaking, reading and writing), testing should include information on the content areas as they affect reclassification decisions. On one hand, I'm not sure about including content area assessment as it may confound antecedent (limited proficiency) with consequent (academic achievement). On the other hand, along with the Supreme Court and most state departments of education, I certainly agree with the need to assess the four skill areas. The need to assess all four skill areas has always been with us. We just haven't attended to it. The failure to include measures of reading and writing stems from two problems. First, there has been a dearth of appropriate tests, and second, we have had few viable (empirically defensible) decision making models for combining the results of the four seemingly different tests or measures. More discussion of this issue follows in a later section.

O'Malley's second level of concern has to do with more technical aspects of test construction and general issues of measurement than with the concept of language proficiency per se.

He notes the following:

> Having a single form of the instrument fails to reflect the need for the repeated pre-post assessment of students in order to determine growth, as is necessary for program evaluation or for monitoring student progress.

O'Malley's issue here is one of parallel forms. In order to avoid the learning effects resulting from repeated administrations of the same test, publishers produce several versions of the same test. It is not the case, as O'Malley implies, that none of the currently available tests has parallel versions. Nonetheless, the sense of his argument rings true. There is certainly a need for parallel versions of any test used repeatedly over time. Others, however, argue by default (based on practice rather explicit argument or theory) that language

proficiency tests are used for identification and entry only and to use them across time or in evaluation would be to confound purposes. Thus, according to this position, we would identify with one set of tests and exit with another. The argument is nonsense. Not only is it based on faulty assumptions regarding the purposes for testing, but it leads to a proliferation of testing. Finally, the process of using different measures across time (e.g., one for entry and a different one for exit) leads to inconsistent criteria and presents an impossible task for program evaluation and tracking of student data.

O'Malley's final two points are more complex than they appear and merit a good deal more consideration than afforded in the original discussion. "Using scores on a five point or similarly defined scale does not reflect the full underlying continuum of Language Proficiency." The comment seems almost glib. No single number or set of numbers ever totally captures any naturally occurring phenomenon. The world is far too complex to be reduced to a finite set of numbers. However, we have few alternatives; the use of numbers as on a one-to-five scale is nothing other than an approximation, a pragmatic attempt to apply the fundamental properties of numbers, which are known, in an attempt to reflect reality, which is unknown. To expect more or less from science is sophomoric.

Two other aspects in O'Malley's statement are worth elaborating. The first deals with an implicit or unstated set of notions regarding what should be assessed: that we know and can agree on the exact nature of the "full underlying continuum of Language Proficiency." I'm not so sure we can. The second deals with the more general issue of test scores and their scale values. Both of these concerns are addressed more directly below.

Finally, O'Malley alludes to the problem underlying the concept of language dominance:

> Using scores without normative information leaves projects with no basis for making comparisons relative to other populations or to determine growth.

The real issue is one of comparison and the establishment of base rates or norms. With whom should language minority students be compared? Should it be with students like themselves or should it be with mainstream nonminority students? This is a terribly important issue and tends to be glossed over by test user and publisher alike.

Issues raised by O'Malley are at the heart of test development and the theory of measurement. In the following we have tried to step back and to consider some the above problems in a more abstract manner. Perhaps consideration of some of the underlying issues will lead to a clarification of

current limitations and confusions regarding present testing practices. The following is not intended to be a review of tests. Tests will not be listed or commented on by name. I will, however, refer to several reviews and compilations that address various strengths and weaknesses. My purpose is more abstract. Issues and problems will be discussed as they are encountered in conducting research and developing tests. In many instances it was not possible to do more than to point out some of the problems with different approaches. Answers to the question, "What do we do about the problem?" are not always quick in coming.

Testing and the Theory of Measurement

The invariant feature in any approach to language testing is the assignment of numbers to language phenomena. Test scores are based on either the mathematics of classical test theory (see Gullikson, 1950) or on more recent developments in test theory and construction (see Embretson, 1985; Mac Arthur, 1987) or on advances in technology (see Stansfield, 1985; Freedle, 1990). For the present we will focus on tests currently being used by school districts to identify, place and exit. We will further restrict ourselves to tests derived, or claiming to be derived, from classical theory rather than from experimental approaches based on Latent Trait or Rasch models (De Long, 1985) or applications borrowed from artificial intelligence models (Freedle, 1990). It is not our purpose here to review tests per se. There are numerous such reviews, several of which will be mentioned.

Before discussing some of these compilations or lists, I would, however, like to make the point that one of the main reasons the field is in such disarray is that there are few tests that will stand up to a thoroughgoing psychometric analysis. This was true in 1976, and it is true today.

To illustrate: A good number of tests are little more than a haphazard collection of items. Often the items in these tests are selected or taken directly from curricular materials, lists of education objectives or even other tests purporting to measure entirely different constructs. All of this is, of course, done with little or no regard for scale or test properties, let alone linguistic or psychological theory. That these tests should run into psychometric trouble is not surprising. Moreover, developers not trained in either classical or more recent test theory are led to make nonsensical claims.

For example, according to the review of tests listed by ETS, one test manual reports a reliability coefficient (Alpha) of .99. Whoever reported such a figure obviously knows little or nothing about test construction or theory. A reliability of .99 is virtually unheard of, if not impossible. An Alpha of .99 means that interitem correlations are so high as to imply that all items in the test are measuring exactly the same thing. In other words, there is little need

for more than one item. A one item test is ridiculous. How the developers of such a test can then claim to be able to categorize students according to six linguistic proficiency types (categories) is incomprehensible. Test development is ostensibly a technical exercise. It is a long and often tedious task involving the integration of a number of different disciplines. But more of this later.

Test Compilations/Reviews

Over the past ten years there have been numerous compilations of tests used for language minority students. For example, in 1974 Ehrlich and Ehrlich; in 1976 the Texas Education Agency; and in 1976 Silverman, Noa and Russell compiled a list of several hundred tests. A similar list was compiled by Pletcher and others in 1978. Since then, similar lists have been developed that focus more on the tests most commonly used in the field. The Test Information Center at the Educational Testing Service, for example, has developed a table describing the characteristics of oral English language proficiency including thirteen of the most popular tests. The ETS document includes technical information on the tests, usage, skills assessed, types of scores and scales produced by them.

Similarly, there has been no shortage of discussion of the strengths and weaknesses of tests. A good many states prepare and distribute test evaluations. California and Texas are among the leaders in this effort. The reviews by Anderson and others (1989) and RMC (1989) are probably the most comprehensive of the current instruments. Unfortunately, however, most compilations are written from a single point of view and do not include alternative discussion or debate as found in Buros. See also Sweetland and Keyser (1986) for nonevaluative descriptions of a wide variety of tests. It should be pointed out that of all the test compilations, only Buros applies any criteria before tests are reviewed. It is doubtful that very many of the current tests would pass Buros' criteria for inclusion or review.

In addition to test compilations and reviews, various states have developed formal criteria for identification, placement and reclassification as well as for the evaluation and use of specific language tests. De George (1988) provides an excellent summary of most of the currently used procedures. For the most part they tend to be quite similar.

Virtually all states that have formalized processes use the same basic procedure. Language minority students are first screened though the use of a home language survey. Those indicating the use of a home language other than English are then tested with one of the state approved tests to determine the extent to which the student is a non, limited or fluent speaker of English. Depending on the level of English proficiency, some students are then tested to determine the extent of proficiency in the home language. These data are

then combined and students are placed in various programs depending on need. Exit or reclassification is triggered in various ways from teacher judgment to performance on one or another test to time in the program to parental decision. It is curious that in virtually none of these compilations were we able to find any direct technical information about teacher observations or other informal assessment procedures. While there are a good many references to the advisability of multiple criteria that include informal techniques, the lack of empirical research is astonishing. It would appear that the developers and users of these informal techniques have been forgiven the burden of demonstrating validity and reliability.

It may be of interest to recall the prior comment regarding the need to determine the validity of informal assessment in the same way that formal assessment is determined. In this connection, consider the home language surveys used by most states to determine the initial pool of possible eligible students. A search for references to studies that have examined the validity of such surveys failed to produce any findings. One is left wondering about the validity of asking parents, who are often unclear about their own linguistic habits, to make detailed retrospective judgements about their children's linguistic patterns. While parents are certainly in the best position to know, the use of technically unevaluated questions could well be problematic. Some parents may simply not know how well the child speaks a language that the parents do not speak. Some parents may well be loath to admit the use of a language other than English. Some may be fearful about placing their child in a special program and are confused over the purpose of the survey or fear reprisal from immigration authorities and so on.

Reviews and compilations of current tests for language minority students include only a few references to tests of reading and writing. According to O'Malley (1989) there were only two tests, at the time of his writing, that addressed reading and writing in conjunction with listening and speaking. Actually there are three such tests.

The lack of appropriately developed reading/writing tests has led many districts to employ standardized norm referenced tests (NRTs) such as the CTBS. Others have employed Criterion - Referenced tests. Duncan and De Avila (1988) list four basic problems with using NRTs originally developed for mainstream use:

1. NRTs assess ability across a broad of academic subject matter. Not all of the content assessed by NRTs necessarily fits within the concept of language proficiency. For example, math computation would not fit, whereas word problems would.

2. NRTs are designed to assess academic performance across a broad range of abilities. NRTs sample from the lowest to the highest ability levels. As a result not all items are of equal probability across all levels of ability. There are a few very difficult items intended for the high achiever and a few that are quite easy for the low achieving student. The difficulty (p-value) of most items under ideal circumstances is about .50. From the point of view of classical test theory a p-value of .50 maximizes the information contained in the item.

3. NRTs do not assess oral language proficiency. Moreover they provide no guidelines for how to combine NRT results with oral proficiency test results. Differences in scale, content, format and standardization make such cross-referencing difficult at best.

4. Standardization of test instructions can be a problem. NRTs often make it difficult for language minority students to understand what they are being asked. If students' oral skills are below a certain level, it is doubtful that test instructions will be understood. As a result, it becomes difficult to determine that a child has failed because he or she failed to master the material being assessed or because he or she did not comprehend orally administered test instructions.

In summary, NRTs were not designed with language minority students in mind. As a result their construction is not compatible with the purposes underlying the testing of language proficiency. The fact that they continue to be used in this way is unfortunate. Further, lack of stable assessment results has been used to argue against language minority programs in general. Some have taken a position against the need for special services for language minority students because, "a good many students from mainstream backgrounds score in the Limited English Proficient range." In fact, Baker and Rosell have gone so far as to say that low achievement on the part of language minority students has little to do with language. Oller and Perkins (1978) take the opposite position and argue that language and achievement tests assess the same thing regardless of student characteristics. Finally, still others have taken the position that the NRTs are biased and should not be used with language minority students because of the potentially negative effects of tracking and labeling (see Ulibarri, 1990).

Ulibarri (1990) has discussed the use of NRT's, commenting on the issue of test bias. His review concludes that the research on test bias as opposed to cultural bias has been misdirected and that studies designed to show the bias in testing have shown just the opposite, that "bias against minorities does not exist, or that where bias does exist it is in favor of minority students" (see also Jensen, 1980). Ulibarri's conclusion regarding problems of bias is worth noting. According to him, the reason test bias is often not found is that standard definitions of test bias focus on "over and underestimates" of test scores rather

than whether the test are measuring the same construct for all students. The issue for Ulibarri — and I would agree — is one of the validity and fairness of achievement tests, a discussion of which could take us far afield from the present topic (see Fairtest for more discussion regarding the controversy over the use of multiple choice test formats).

Another reaction has been to recognize the potential for unfairness and to avoid assessing achievement at all. Thus, language minority students are not tested; a host of other problems result. For example, if language minority students are participating in a program, failure to test at the onset of the program (pre-test) will tend to elevate pretest scores and make it difficult to show programmatic gain since the baseline data will not have included the students who would probably have stood the greatest chance to show growth. See Quesada (1979) for still other problems created by this seemingly benign strategy.

Finally, test information derived from NRTs often engenders invidious comparisons between the individual student and his or her peers. Worse yet, these group comparisons are made irrespective of what has been learned over the course of the year. Thus, reporting reading level as a national percentile tells the parent (or teacher) precious little about how well the child actually reads, only that she or he reads better of worse that his or her peers. This tends to make parents passive agents in the education enterprise. A more constructive approach would be to design the reporting system in such a way as to suggest what a student needs to work on and how the teacher and/or parent can help.

A number of recent and not so recent surveys suggest that discussion should focus on a limited number of tests and not on testing in general. Numerous surveys show that, over fifteen years, districts have tended to use only a few tests. For example, the Evaluation Assistance Center at the University of New Mexico reports the results of a recent survey in which 416 Title VII projects were asked to provide information regarding district testing procedures. Information from 145 programs found that twenty-eight different tests were being used. More importantly, seventy-six percent of those responding to the survey used one of four tests. Data from other similar surveys would tend to support this finding; approximately eighty percent of the students are tested with one of four tests.

In another survey commissioned by the U.S. Department of Education (OBEMLA), the Special Issues Analyses Center(1989) found that a fairly limited number of tests were being used both to identify and to assess achievement. In fact, many were being used for both purposes. The most striking finding reported, however, was that the most frequently cited instrument used to assess achievement was an "Unspecified Standardized Test."

More than likely these have been criterion referenced tests. It is doubtful that any of these tests would meet even the most modest psychometric standards.

Criterion-referenced tests have been attractive because they represent an attempt to tie assessment to instruction. Hoffmiester (1975) describes the link as follows: "Criterion-Referenced testing can reach its full potential when it is so integrated with the day-by-day functioning of the classroom that it cannot be easily separated out as a testing activity" (p. 77). Test items written for a criterion-referenced test are thus taken almost directly from classroom instruction or a particular instructional program. In fact, in one instance we have been able to document, test items have been directly extracted from the curriculum, artwork and all. The danger is, of course, that instruction may have little to do with achievement in the general sense. Thus, for example, a student may be instructed on a simple vocabulary list consisting of twenty words. Based on a criterion test the student passes fifteen items, which meets criteria set by a panel of "experts." Can it really be said that the student is "English proficient?"

In this example, rather than a test-driven curriculum, we have a curriculum-driven test. Both are problematic. While the former leads to teaching to the test, the latter leads to assessing the curriculum. Both strategies are misplaced. The former is what happens when the importance of NRTs is overemphasized. The latter places an overemphasis on instructional specificity with the result that the student is able to pass the particular criterion reading test but unable to read text different from that used in instruction. In other words, there is a lack of generalizability from the context of instruction to other contexts.

Another related problem, which many seem to be unaware of, has to do with setting criterion levels for deciding that a student has mastered a given topic or domain. There are only two alternatives. Criterion levels can be set according to "expert judgment" or empirically. Ultimately, both approaches are data based. In the case of expert judgment, judges base levels according to their own experience about what can be expected. Thus, "eight out of ten" becomes a reasonable expectation or criterion level. The alternative to setting criterion levels is to collect data or to set levels based on "what the average student can be expected to complete successfully." Both approaches are ultimately data based; the former is based on prior informal experience whereas the latter is based on formal analyses of student performance.

Finally, to the extent that criterion-referenced tests are tied to specific instruction, there are problems of comparability, particularly when it comes to evaluating program effects. Consider, for example, a situation in which there are six different approaches to instruction, each emphasizing a different aspect of language. Each test would be different and, therefore, not directly compa-

rable. In the same way as a curriculum should be generalizable, a test should be sufficiently broad to accommodate differences in instructional approach. Criterion referenced tests, contrary to what seems to be the prevailing belief, are subject to the same constraints and requirements as any other assessment procedure or device. Again, as with informal approaches, many seem to feel that they have been relieved of their scientific burdens.

Research on the Tests

There has been very little research on the general issue of the use of tests with language minorities. Criticism, however, abounds and has been often applied in a somewhat indiscriminate fashion, used politically to argue for the elimination of all testing or for the use of alternatives.

One of the studies most often cited was conducted by Spencer and a number of her colleagues (Ulibarri, Spencer and Rivas, 1981), who compared the results of four different oral language proficiency tests. What these researchers found was that the four tests varied widely as to the numbers of students identified or categorized as non, limited or fluent speakers. Further, based on another set of findings, Merino and Spencer (1983) concluded that "there are substantial reasons to doubt the comparability of these oral language proficiency instruments across languages."

A number of other studies have examined the equivalency of different tests used across different populations. Cabello (1983) states one position rather succinctly: "creating a Spanish language test which is equally comprehensible to Mexican, Puerto Rican and other Hispanics of varying educational and social backgrounds is as difficult as creating a test in English to serve American, English and other English speaking students as well." I would think that there would have to be sufficient overlap between different dialect versions of the same language to justify assessment; how else could they be referred to by the same name? The trick is to design the items to represent the overlap and not the differences. Sharon Duncan and I have spent a number of years trying to deal with exactly this problem (De Avila & Duncan, 1982, p. 125).

A number of writers have taken equivocal findings to conclude that the entire "objective" assessment approach should be abandoned in favor of a more "flexible" approach, including teacher judgment, socio-economic indicators and the like (see Baker, 1982; Rosansky, 1980; Sanchez, 1979) See Jensen (1980) for a thoroughgoing discussion of the logical extension of this argument.

That the four tests used in the Spencer studies failed to produce the same results is not surprising. However, the conclusion that all testing should be eliminated in favor of even more elusive and politically dangerous processes

seems foolhardy at best, particularly in light of the absence of any agreed upon criteria (formal or otherwise) for judging the accuracy (validity) of oral language tests. The Spencer studies were studies of convergent validity, which is only one form of validity. Willig's (1985) conclusions are typical of those many have drawn from these studies:

> It is a known fact, however, that language tests in general and tests in particular that are used to determine entry and exit into bilingual programs, have low reliability and low convergent validity. (p. 301)

The tendency to rectify equivocal findings is common. Consider, for example, that two of the three papers cited by Willig in support of her contention were published in 1974 before either the Lau regulations or several of the current tests were published. A fourth citation is an unpublished manuscript and two others were reviews of the same material covered by Willig. In fact, of the seven papers cited, only two actually involved any data collection or analysis (Gilmore & Dickerson, 1985; Ulibarri, Spencer, Rivas, 1981).

There are other types of validity (and reliability) that would seem worth investigating given the questionable validity of some of the tests studied. They include "face," "construct," "convergent," and "predicative." All four are necessary. Some critics seem to be operating on the theory that one bad apple can spoil the entire batch. While it is not my purpose to defend these tests, I would suggest that those critical of them stick a little closer to the data. The lack of validity or reliability of one or more tests has nothing, either logically or technically, to do with other tests although it would certainly lower the convergence or agreement between them. My own bias is to be far more concerned about the predictive validity of a given test and its ability to fit into a systematic procedure that takes advantage of both quantitative and qualitative data rather than whether or not it agrees with another, perhaps poorly constructed test.

There are three fundamental conclusions that can be drawn from these studies. First, the establishment of the validity of oral language proficiency tests will remain equivocal until such time as there is an agreed upon operational definition(s) of language proficiency. We seem to be searching still for agreed upon or acceptable criteria against which to validate the tests. In 1976 there were no agreed upon criteria whatsoever. Second, these and other studies have addressed the question of validity from as widely divergent perspectives as test developers have used to develop them. Politically motivated evaluations of tests are as common as belief or disbelief in the programs that spawned them. Third, there is a good deal of work to be done to develop both informal and formal assessment procedures. In other words, judicial and legislative fiat has created

assessment issues that may be insoluble, leaving us with no alternative but to continue muddling through. That we have come this far is quite amazing.

Test Results and the Problem of Scale

The selection of a test involves, at a minimum, a consideration of the content of the test as well as the information produced by it. Test content has been discussed by numerous writers (Oller & Perkins, 1978; De Avila, 1983; Fradd & Tikunoff, 1987; Berko, 1985). Present discussion will focus on the information produced by most currently and widely used tests of language proficiency and not on what should or should not be assessed. As will be seen, the information derived from a test is largely determined by the test developer's approach to test scaling.

With respect to test content it would seem that everybody is right. There seem to be as many important aspects in language as there are people to write tests for them. For example, one criterion-referenced approach employs over 1800 test items to assess the development of proficiency within a particular curriculum. There are other similar examples of criterion-referenced approaches. From my point of view the real question is to determine which constellation of elements (subtests) makes linguistic/educational sense, provides the greatest information (accounts for the greatest variance), and agrees or converges toward common results that are predictive of their criteria. In other words, the choice of test content should be guided by the four types of validity. Tests or subtests failing any of the criteria for validity should be eliminated. For those who have difficulty with this position, I strongly recommend a review of Campbell and Fiske (1959).

The creation or selection of scale type becomes all-important insofar as it determines and limits what can be done with the information produced by the test. There are four kinds of scales in classical measurement theory. Thus, there are only four possible units on which we can base our measurement of things as diverse as the osmotic movement between permeable membranes, the speed of light or phonemic control. The four kinds of scales include nominal, ordinal, ratio and interval. Each has different properties and utility depending on purpose.

Strictly speaking, most language categorizations found in either legislative guidelines or school designations are nominal categories. A nominal category refers to a class of objects, ideas or individuals that have something in common, but have no particular mathematical relationship. Thus, a Ford and a Lexus are both automobiles just as females and males are both human beings. It would be ridiculous to refer to an automobile's having more or less "Fordness" or "automobileness" or a person's having more "human-beingness." A number of currently used language tests employ nominal categories almost

exclusively. Data from these measures are useful in counting the numbers of LEP or FEP students but provide little else. Nominal scales are little more than classification systems and are not the products of measurement in the mathematical sense. They are valuable because they are close to the real data; as such, they are high in face validity.

Fundamental to the nominal scale is its lack of a mathematical base. In fact, a nominal scale is actually the product of a classification system more than a system of measurement (see Mc Arthur, 1987). This means that the establishment of formal criteria (numerically based) or cutoff scores is virtually impossible. See Baker (1988) for a detailed discussion of the inadvisability of using language dominance categories in establishing test norms. Categorical distinctions are normally set by test developers, by whatever means they might use, by teacher judgment, best guess or time of year. One limitation of nominal categories is that it is difficult, if not impossible, to track the growth. For example, movement from the low end of one category to the upper end would not show up in an evaluation based on counting the numbers of limited or non speakers. About all that can be done with categorical data is to count the numbers of students in one or another category over time.

The numerical properties of an ordinal scale, on the other hand, enable us to track progress over time in a more precise manner rather than simply counting students in different categories. The failure to track student growth over time has certainly been a shortcoming in language minority education. An ordinal scale is a set of numbers, such as height or weight, ordered by increasing value. Language proficiency, to the extent that it is based on a test made up of increasing values (e.g., 1 to 100), would be an ordinal scale. The usual process creates ordinal scales by simply summing correct responses without regard for item difficulty.

Some approaches to the creation of language categories employ ordinal scales. For example, one test sets cutoff scores or levels based on standard deviation units drawn from the norm group frequency distribution. This is totally acceptable from a psychometric point of view as long as the number of levels is held to a minimum and the standard error of measure is taken into account as placement decisions are made.

Ratio and interval scales refer to the creation of numbers that are derived from other numbers. These derived units of measure are the result of various mathematical transformations. For example, percentile scores are an example of a ratio scale, whereas Normal Curve Equivalents (NCE) are an example of an interval scale. The importance of the interval over the ratio scale in this context refers to the equipotentiality (equal value) of test item scores. Thus, for example, Normal Curve Equivalents are created from percentiles and enable us to perform the mathematical operations necessary for program

evaluation because the difference (interval) between any two adjacent scores is the same, whereas adjacent percentile or grade equivalency scores are not necessarily of the same value. "The interval and ratio scales are by far the most useful measurement scales employed in science" (Torgeson, 1958). They enable us to go beyond simple summing of correct answers. We can introduce greater complexity to scores by means of standard mathematical transformations. Dividing subscale scores by the total number of items in the subscale in order to make subscales with different numbers of items comparable is an example of a simple transformation. Creation of most standardized scores is based on similar ratio scales. For example, the creation of Normal Curve Equivalents from percentiles produces an equal interval scale. The advantage of NCEs over other scales is that their mathematical properties are thoroughly known and far more amenable to sophisticated analyses than other scales. This means that data can be subjected to rigorous analyses without violation of mathematical assumptions. The question becomes what kind of transformations can be made (which add information or under-standing) without loss of the original empirical information.

The way in which data are reported is critical to fulfilling informational need at different levels within the administrative/instructional hierarchy. Not all of metrics are universally useful. In some cases they are even misleading. For example, a number of currently popular tests report results in percentile scores. Percentile scores are derived from the frequency distribution of scores generated from norm group data. The rank of each score is computed as a function of the percentage of students who received a particular score. Thus, percentile rank scores represent the value position of each score relative to any other score in the obtained frequency distribution of scores from the norm group. Percentile scores have long been criticized for not having equal intervals. Tallmadge and Wood (n.d.) argue that, although percentiles satisfy the need for a common index, "they should not be used in arithmetic computation" as would be necessary for pre/post comparisons. Instead, Tallmadge and Wood argue that the "NCE metric is an equal interval scale" and, therefore, can be legitimately used in arithmetic computations such as those needed for determining pre/post gains.

NCEs divide the area under a normal curve into ninety-nine equal parts (eleven points per stanine). NCEs also have the feature of a common mean of 50 and a standard deviation of 21.06. Their values match percentile values at the 1st, 50th and 99th percentiles. The process for converting test scores to percentiles has been described by Tatsuoka (1970) and others. Conversion of percentiles to NCEs can accomplished by means of table conversions prepared by Tallmadge and Wood for Chapter 1 evaluations.

NCEs are obtained from percentiles by means of a "Standard Linear Transformation" using the binomial theorem where

$$Y = aX + b.$$

To the extent that total scores can be themselves standardized at equal intervals, they can be used for statistical analyses. Consider, however, just how many of tests are in fact standardized.

In so far as percentiles and NCEs are based on frequency distributions obtained from the norm group, sample selection becomes all the more important. Consider, for example, if the norm group is made up of entirely of native speakers, then it is highly likely that there will be very little variance, and scores will be skewed toward the higher end of the scale. This means that a good many of the limited and, more than likely, all of the non speakers will fall off of the scale since their scores can be expected to be at the lower end of the scale. In other words, there will be a bimodal distribution. (See Figure 1).

In reviewing various approaches to understanding the relationship between linguistic and academic performance, we have concluded that there are only two approaches to the creation of a viable entry/exit model that are practically viable and theoretically defensible. The two approaches, discussed below, are not necessarily competing or conflicting. They may, in fact, compliment each other.

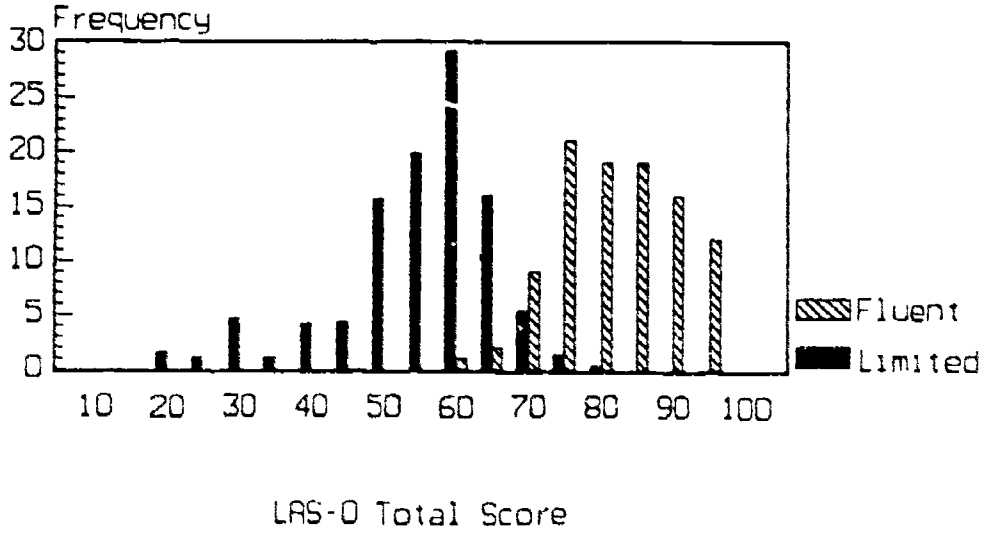
Multiple-Criteria Approaches

Most workers in the field argue that multiple indicators are needed in order to make sound decisions. Virtually everybody in the field agrees that it is not a good idea to make program determinations on the basis of a single test score. Thus, in addition to using reading, writing, listening and speaking, it is recommended that practitioners make decisions on the integrated information. Projects should rely on multiple sources of information obtained through varied types of data collection and thereby increase the accuracy of selection, placement, diagnosis and evaluative functions.

The use of multiple indicators to make decisions requires a consistent decision making process. Under ideal circumstances decision rules set cutoff or criterion levels for each variable according to its relative importance. Thus, for example, we can decide that six important factors will be considered in classifying students as non, limited or fluent speakers of English. The problem, of course, comes in deciding how to weight each factor included in the process. Should reading be more heavily weighted than listening or vice versa?

As discussed in the section on criterion-referenced tests, weights, like cutoff scores, can be set in two ways — through expert judgment or empirically. In the following, we will briefly describe these two approaches, the former based on judgment and the latter based on an empirical relationship between predictor and criterion variables.

Figure 1
Frequency Distribution
LAS-O Total Score Form I C & D
Limited Vrs. Fluent English Speakers



Let us say that we have agreed that language proficiency is made up of measures of reading, writing, listening, speaking, teacher judgments and proficiency in the home language. In effect, this constitutes a working operational definition of language proficiency.

> Lang. Prf. = Listening + Speaking + Reading + Writing + Teacher Judgments + Home Lang. Prf.

Scores for each variable in the equation can, according to this model, be obtained and totals calculated. This leaves five problems. First, there is nothing to say the above six variable are indeed the critical variables. There could be others. Second, there is nothing inherent in the variables themselves to suggest specific cutoff scores. We don't have any information about how much of each is needed. Third, the model assumes an additive relationship among variables. They may not all be linearly related. Fourth, our definition of language proficiency is as limited or as broad as the variables used to describe it. Fifth, there are no ready-made criterion measures against which to validate the predictors.

One way to examine the problem of determining which variables are more important would be to collect data on each predictor variable (right side of the equation) and the criterion (left side of the equation). Then, through the use of multiple regression techniques, we could empirically establish the relative importance of each factor. This would lead to an equation in which, based on the data entered, each factor is weighted (beta) according to its relative importance or contribution to the total predictor/criterion variance. The resultant equation is thus an empirically derived description of language proficiency that includes each factor in order of importance. The process is seductively simple, and several writers have been led to the conclusion that all we have to do is "find the right set of predictors" (see Baker, 1983; Winter, 1984).

The problem of statistical colinearity is not readily apparent to the lay public. The issue is that it is extremely difficult to identify the relative importance of correlated variables, and importance can be obscured by a simple altering of the order in which they are placed or entered into the analyses process. While there are processes for minimizing these effects (see Wonnacott & Wonnacott, 1987), one has to be extremely careful in attempting to generalize results to policy decisions. The point here is that, regardless of statistical sophistication, there are no ready made techniques for dealing with all of the complexities that must be addressed in creating a viable model. Moreover, a purely empirical approach is apt to oversimplify both the problem and its solution.

The Socio-Linguistic Approach of Cummins

Cummins (1984) borrowed Donaldson's (1978) "context reduced/ embedded" distinctions to explain difficulties in linguistic communication faced by language minority students. The approach has received widespread attention over the past several years as an explanation of how language minority students seem to be English proficient yet perform poorly in school content areas. Cummins explains this phenomenon by suggesting that two sets of skills define language proficiency. The first involves what Cummins refers to as "basic interpersonal communication skills" (BICS) and the second involves "cognitive academic language proficiency skills" (CALPS). The primary distinction between the two rests in the extent to which the communicative act is context reduced or embedded. BICS refers to context embedded speech whereas CALPS are acts that take place in a context-reduced environment. A context-reduced environment is one in which situational cues, such as those provided by verbal or other feedback, have been reduced. Context-embedded communication is more like what takes place in everyday communication between individuals. The former thus relies on external interpersonal cues whereas the latter relies on internal knowledge of appropriate responses.

Cummins argues that most language proficiency testing is actually little more than an assessment of interpersonal communication skills (BICS). As a result students are often exited prematurely or before their chances for success are realized. According to Cummins, many tests fail to include sufficient assessment of more cognitive/academic content. In this connection talking to a friend about another friend would involve BICS whereas writing an essay would involve CALPS. A thoroughgoing assessment of language proficiency in Cummins' view would, therefore, consist of both BICS and CALPS items.

Unfortunately, however, currently available tests were not constructed with Cummins' distinctions in mind although a number of them claim to measure BICS and CALPS. Moreover, with the possible exception of Gottlieb (1990), there have been no attempts, to our knowledge, to develop a set of measures designed specifically to assess BICS and CALPS. According to Chamot and O'Malley (1986), instruments for measuring cognitive academic language proficiency are not available. De George (1988) maintains that "most English Language proficiency tests do not measure academic language proficiency . . . and standardized tests in English confound content knowledge with language proficiency."

Irrespective of some of the theoretical difficulties, the approach has a good deal of intuitive appeal. The operationalization of BICS and CALPS within a testable framework, however, remains elusive. For the most, part the

distinctions between BICS and CALPS have been used to explain existing data (a posteriori) and not to predict (priori) or actually to create a model.

On the other hand, Cummins has offered valuable and early insight regarding the fallacy of many approaches to the entry-exit process. For Cummins (1980) the entry-exit fallacy is in the belief that students can be placed into and out of programs on grounds of basic interpersonal skills and little else.

While it is not within the scope of this report to present a detailed analysis of Cummins' position or to develop a research agenda to validate the theory, it is important to recognize Cummins' admonitions about the importance of considering language proficiency in the broadest sense and not to restrict it to simple everyday linguistic interactions.

My impression is that Cummins' ideas are holistic, far more on the instructional and qualitative side of the equation as opposed to the quantitative or assessment side. The same can be said for the input hypothesis offered by Krashen (1982). These are useful metaphors, however difficult to quantify.

De Avila's Probabilistic Approach

De Avila, Cervantes, and Duncan (1978) reviewed various state and federal requirements for entry/exit criteria and concluded that the establishment of a model required the simultaneous yet independent consideration of both academic and linguistic skills. In an attempt to develop an empirically testable model consistent with federal and state requirements (i.e., that children must be provided a means to participate in the educational system), it was reasoned that children should be considered eligible for program entry whenever their English proficiency is significantly below that of their English monolingual peers. By extension, they argued that children should remain in programs until their expected level of academic achievement or probability of success is indistinguishable from that of mainstream children or, conversely, until expected failure cannot be attributed to limitations in language proficiency. The logic of the argument followed from the Lau decision that reasoned that children were failing because they did not understand what was taking place in the classroom.

Finally, De Avila and his colleagues argued that the point of intersection or crossover between school achievement and oral proficiency was the most defensible point at which to establish an exit cutoff because it could be reduced to an empirical definition. In order to test the proposition, however, they argued that several other issues had to be addressed. Their first concern again dealt with definition, particularly with the distinction between dominance and proficiency tests.

After reviewing different approaches, they concluded that data provided by dominance tests (nominal categories) were of little use in the development of entry and/or exit criteria because dominance tests do not provide continuous data and ~ e, therefore, difficult to relate to achievement in a way that would facilitate the establishment of cutoff scores or models that could be empirically tested. Dulay and Burt (1978) provide additional discussion on the topic from a somewhat different perspective.

De Avila et al. presented their model in two parts. First, in operational terms, a probability based model assumes a linear relationship between linguistic proficiency and academic achievement. The model may be better understood by referring to Figure 2, which presents the linear relationship between academic performance and linguistic proficiency.

The second aspect of the model involves the application or inclusion of a cutoff criterion or exit score based on academic performance. This component may be added to the model by including the average academic performance of the population with whom the language minority children are to be compared. Figure 3 shows the average academic performance of the majority comparison group as a straight line running parallel to the line indicating oral language proficiency. Note that the figure assumes that linguistic proficiency for the comparison group (language majority) is held constant or unchanging across different levels of language proficiency for the language minority student. In other words, the model assumes that, while individual variation in English language proficiency exists for the monolingual comparison group, it is insignificant in comparison to that of language minority students. This variation is accommodated within the De Avila and Duncan model by setting or defining cutoff or criterion levels as a bandwidth that allows for individual variations.

Under contract with the California State Department of Education, three small scale studies (De Avila, Cervantes, and Duncan, 1978) were conducted to test the model. Data were collected on approximately 500 children at a number of schools throughout California and several other states in grades one through twelve. A commonly used test of language proficiency and standardized tests of achievement in reading and mathematics were administered to all children participating in the study. A total of eighteen separate analyses was conducted. The analyses included (1) an analysis of variance (ANOVA) to test the hypothesized achievement difference across five language proficiency levels; (2) tests of linearity to examine the hypothesized straight-line (linear) relationship between the two sets of scores; and (3) correlational analyses to examine similarity of pattern. Of the eighteen analyses conducted, fifteen were found to be statistically significant in the predicted direction.

Figure 2
Hypothesized Relationship Between
Language Proficiency &
Academic Performance

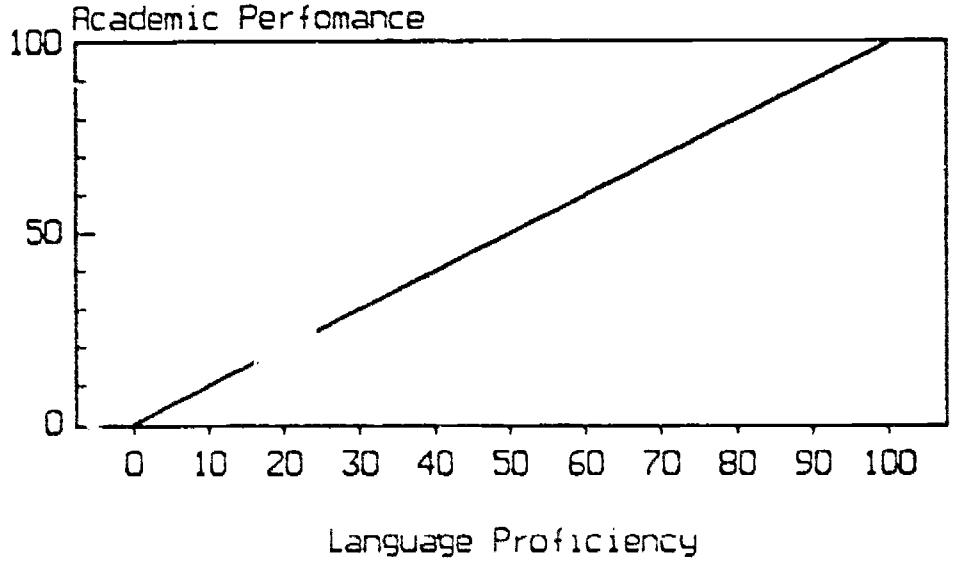
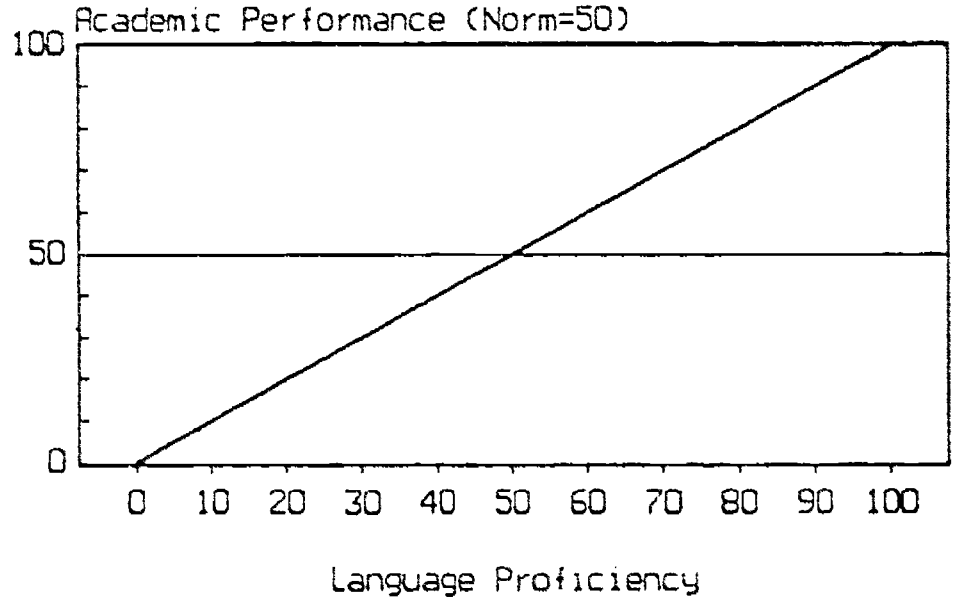


Figure 3
Hypothesized Relationship Between
Language Proficiency & Academic Norm



Data compiled from several other studies which were examined by De Avila, Cervantes, and Duncan found nonsignificant correlations between oral language proficiency and academic achievement among language majority or mainstream students. Other studies report similar results among proficient speakers. Scores for mainstream students were, with rare exception, well within the proficient range. The lack of variance, as predicted, resulted in low-order correlations (see De Avila and Duncan, 1981).

In a study carried out by the Houston Independent School District involving several thousand students, similar support for the approach was found. In this study, researchers plotted frequency of chance performance on an achievement test against oral language proficiency levels. They found an inverse relationship between proficiency and chance performance. As proficiency level went up, the frequency of students performing at a chance level of achievement went down. These data provide direct support for the approach from another analytic perspective.

Assessment of English language proficiency as a predictor of school achievement in monolingual English speaking settings, however, provides no information about the probability of success in particular programs (i.e., bilingual, ESL, sheltered English, etc.). Moreover, level of proficiency in one language (the home language or English) cannot be taken as indicative of proficiency in a second language or of the probability of success of instruction in that language. Therefore, both languages should be assessed in order to maintain full understanding of the student's capabilities.

The incorporation of home language proficiency into the current model (a form of linguistic parity) is accomplished by simply repeating the same process described for English. Unfortunately, however, few states seem to require assessment of home language proficiency. Fradd and Tikunoff (1987, p. 25) cite a recent survey by Development Associates in which districts indicated that only about two percent of the reporting districts actually use home language proficiency data to determine special language services eligibility.

While there seems to be good support for the probabilistic approach, there are potential problems in using achievement and oral proficiency tests that have not been validated against the model. For example, there is some difficulty in equating different (e.g., English and Spanish) achievement tests. However, De Avila and Duncan (1984) have pointed out that as long as both tests cover the same academic content and are reported in standard units, it would be possible to plot achievement scores for English and the home language on the same curve.

On the other hand, there is a lack of comparability between different oral proficiency tests which arises from the fact that the item P-values (difficulty levels) are different for the different tests. In practical terms this means that some of the tests are very difficult (i.e., there is a low statistical probability of a correct response); others are very easy, and so on (see Ulibarri, Spencer & Rivas, 1981). Thus, students could be kept out of programs through the use of easy tests to identify eligibility or kept in through the use of a very difficult test to reclassify.

De Avila and Duncan (1988) recently replicated various features of the above studies in which they compared the data obtained for the 1978 study and to data collected in 1987. Results revealed a strong degree of consistency between the two sets of data.

Figure 3 provides a graphic representation of the model in which normative expectations (arbitrarily set at 50 for both) are provided for both minority and majority populations. In a recent memo prepared by the State Department of Education/The University of New York (Walton, 1989), cutoff levels were compared across different states. There was a range from a low of the twenty-third percentile (Texas) to a high of the forty-ninth (Kansas). In no case was it indicated that any of the levels was empirically established. Most seem to be guided by other factors such as Chapter 1 criteria, availability of resources, and so on. The use of a single score or cutoff is dangerous in that it sometimes stretches the accuracy beyond its ability to discriminate between adjacent scores (De Avila & Duncan, 1982). Thus, the idea of using a bandwidth was introduced as long ago as 1976. Unfortunately, few tests offer bandwidths as viable solutions to the problem of standard error of measurement (Kerlinger, 1973).

In a limited sense, one might also conclude that the above results support Cummins' notion of concept embeddedness, in which oral data are taken as embedded and achievement data are thought of as disembedded or context reduced. See Tannen (1982) for a critique of the assertion that writing, for example, is necessarily more decontextualized than speech. The major difference between the probalilistic model and the Cummins position rests on the distinction between BICS and CALPS. A probabilistic model bears no burden with respect to specifying content.

The importance of sound assessment at every step of any model should be obvious. If the assessment of language proficiency, for example, fails to produce a linear relationship, the entire system can be questioned. I strongly question whether all of the tests listed by various test compilations can meet this requirement. This is why studies of convergent validity such as described above produce disappointing results. Failure to attend to other aspects of validity only exacerbates the problem.

Finally, Spolsky's (1968) comment on this issue is refreshing:

> Interpreting test scores calls for experience, flexibility, and willingness to remember that one's statements are probabilities rather than certainties. To expect more of a test is, at best, foolish. To claim more is, at best, naive.

Informational Needs and More Measurement Issues

O'Malley (1989) lists five major uses of test information (selection, diagnosis, placement, reclassification, and evaluation) that fall into three administrative levels within the educational system. Gonzalez (1984) offers similar insights from a governance point of view.

Information gained from tests is typically used at three administrative levels within the educational establishment. First, test information is used at a policy level by state, federal and legislative offices. The information needed at this level includes general group statistics regarding educational attainment level and numbers of students of different types. Impetus for testing at the federal level comes from four different sources, including Public Law 94-142, Native American BIA (see Milne, 1987).

There are no fewer than ten states that have specific policy regrading assessment of the educational progress of language minority students (see De George, 1988). Such data are used to establish programs (e.g., drop-out prevention) and set performance objectives. The recent report entitled "A Summary of State Reports on the Limited English Student Population" is an example of the use of information generated largely from test results. At the local supervisory level, district administrators, program directors and evaluators use test data to satisfy state and federal requirements in addition to school board demands.

At a second level, local district administrators, program directors, and evaluators need assessment information. In one sense, they require the same information as the superordinate agencies in order to establish, design, monitor and evaluate programs. Moreover, test information is used to assess local need, to determine the number of teachers required, and to allocate other resources. Finally, building personnel need to track student progress at the classroom and individual student levels in order to place and reclassify students.

Unfortunately, classroom needs are often the last considered. Testing requirements, including selection of tests and testing schedules, are usually set at the supervisory or superordinate levels. Teachers are seldom involved in the process of test selection or the design of methods for the use of data. Moreover,

teachers are seldom trained on how to use test information. Involving teachers who have not been trained on what tests can and cannot do sometimes only adds to the confusion. Teachers need to know how to interpret test results in order to diagnose specific strengths and weaknesses of individual students. Teachers also need to be able to track progress on a daily, weekly or monthly basis in order to adjust the content and method of instruction. Seldom are teachers provided with either reports or assistance in using test data and, as a result, tend to view testing as little more than an intrusion or interruption of instructional time. There is a plethora of teacher made tests that, to a greater or lessor extent, satisfy teacher needs. They serve little purpose beyond the classroom insofar as they are not comparable from one classroom to the next.

In this connection it is worth commenting that the preparation of test items is

> essentially creative — it is an art. Just as there are no set formulas for producing a good story or a good painting, so there can be no set of rules that guarantees the production of good test items. Principles can be established and suggestions offered, but it is the writer's judgment in the application — and occasional disregard — of these principles that determines whether good items or mediocre ones are produced. Each item, as it is being written, presents new problems and new opportunities. Thus item writing requires an uncommon combination of special abilities and is mastered through extensive practice. (Wesman, 1971, p. 8)

With respect to test design, it should be borne in mind that, in addition to concerns over item design, final item selections are governed by both practical and theoretical concerns. For example, since measures of oral proficiency are often administered to large numbers of language minority students (who may not comprehend standardized test instructions), there are such constraints as logistics, training of examiners and test utility that must be considered. While, in the most ethnologically ideal of all possible worlds, such measurements would involve the collection of natural language samples in various sociolinguistic settings (except for the school, which may not reflect the child's home culture), the harsh reality about language assessment is that:

1. Children are tested in school settings (often in less than ideal circumstances). Concern over other language settings is not of particular importance to the schools.

2. Testing adds greatly to school burdens (often with no immediate benefit to the individual school). Results tend to be used primarily for administrative purposes.

3. Lengthy transcriptions and linguistic analyses are often beyond the budgets and/or training of most school district personnel.

4. Data analyses and tracking systems needed to take full advantage of the data provided by testing are often not found in local school settings. Moreover, data processing centers are reluctant to use mainframe computers for additional purposes.

5. The people who administer and score language samples and tabulate results are not linguists, psychologists, sociologists, anthropologists or educational administrators. They tend to be community people, who, unfortunately, in many instances may have the same language problems as the students.

6. Testing time is limited to only a few minutes per child.

7. State regulations often require that testing take place in the beginning of the school year when things are in greatest flux and students, particularly the younger ones, are not used to the school routines.

8. Follow up testing for students who score on the "cutoff" is seldom conducted.

9. Interjudge reliability is seldom checked.

10. Tests are selected on the basis of cost, administration time and real or imagined connection to the curriculum as much as on the basis of psychometric considerations.

Evaluation: More Problems

Test scores are used to identify students who are eligible to receive special services, place them in particular programs and reclassify (or exit) them into mainstream classrooms. By virtue of state and federal regulations, test scores are also used to evaluate the effectiveness of programs. The purpose of program evaluation is to determine the effect of program participation. That is, to what extent can gains or losses in academic achievement (is this the sole criterion?), as measured by an acceptable test, be explained by program participation? Districts are required to conduct annual assessment of educational status or progress while controlling for extraneous or unrelated factors that might weaken the validity of evaluation outcomes and, to report their findings according to specific regulations.

On the federal level, both Chapter 1 and Title VII regulations require districts to report student outcome data as part of their evaluation of educational progress. Since both Chapter I and Title VII serve LEP students, it would seem worthwhile to examine how some of the above issues and problems impact the evaluation processes. As will be seen, a juxtaposition of rules and regulations reveals a number of inconsistencies, paradoxes and downright confusion over how best to define, assess and serve children who come from non-English speaking backgrounds and who are experiencing difficulty in the schools.

A cursory examination of Chapter 1 and Title VII rules and regulations underscores some of the confusion particularly in relation to identification and selection of students. Chapter 1 defines eligibility on the basis of whether or not a student is "educationally disadvantaged." Title VII defines eligibility even more loosely on the basis of English language proficiency. Chapter 1 is designed to serve students whose "educational attainment is below the level that is appropriate for children of their age" (Fed. Reg., 1989, p. 21758). Rules and regulations for Chapter 1 are quite explicit in stating that "lack of English Language Proficiency in and of itself is not regarded as sufficient reason to declare a student educationally disadvantaged."

It would seem that Chapter 1 and Title VII are designed to serve two entirely different populations and that Title VII students constitute a subset of the Chapter 1 population, that is, students who are low achieving and who, in addition, are limited English proficient. Closer examination of eligibility standards reveals a far more confusing picture, particularly when the ever present issues of definition and assessment are taken into consideration.

Lack of English language proficiency is supposed to distinguish Title VII students from Chapter 1 students. In order to determine whether or not a student is eligible for services, Chapter 1 applicants are required to use "systematic and objective" measures, implying the need to test language proficiency and academic achievement.

Some of this confusion arises out of the Chapter 1 definition which is, from a practical point of view, indistinguishable from the definition used to identify students for bilingual or ESL programs. Language proficiency is defined as consisting of two aspects, much as Cummins has suggested. The first aspect concerns normal everyday communication. The second concerns academic communication. The first is measured by tests of conversational English whereas the second is measured by norm-referenced or criterion-referenced tests of language, reading, writing and mathematics (Rayford, et al., 1990). These latter tests are the same as those used to identify the Chapter 1 population as a whole. In other words, it would appear that LEP students may well qualify simply on the basis of low achievement without regard for how well they may or may not speak English. Moreover, there is some evidence to suggest

that for elementary school age students, the two communication skills (BICs and CALPs) are closely related.

In a recent set of studies Sharon Duncan and I tested the listening, speaking, reading and writing skills of several thousand language minority and mainstream students. We found that there were almost no students in the elementary grades who were able to read and/or write (as defined by the fortieth percentile on an NRT) who could not speak English. Table 1 shows the relative frequencies for non, limited and proficient speakers across three levels of reading/writing.

Upon further investigation of the population we found as much as 98 percent overlap between Chapter 1 non LEP and LEP students. In other words, oral language skills seem to be critical for the development of reading and writing skills, at least at the elementary level. If we were to examine the oral proficiency of mainstream Chapter 1 students along with reading and writing skills, we would find many students with both oral and literacy problems.

At the secondary level the picture was more complex because of developmental differences. These data showed that there was a significant number of junior and high school students who were able to read and write in English but were not able to speak English. Analyses revealed that students of this type tended to be more affluent recent arrivals who had studied English in the homeland. In this instance, they tended to be older Chinese students from Hong Kong. The phenomenon is also common in the United States, where college students study a foreign language, master its grammar, read and write in a satisfactory manner but are still unable to speak the language.

Nonetheless, it seems somewhat disingenuous to say that Chapter 1 funds cannot be used to overcome limited English proficiency when part of the definition of language proficiency includes the very same elements that define the Chapter 1 population as a whole. This overlap in definition can produce only confusion and, to the extent Chapter 1 programs are prevented from addressing limited English proficiency, programs that fail to meet the needs of language minority students who are LEP. Current practices do not recognize the type of subtleties needed to make these distinctions. The point is that there are many kinds of LEP students.

TABLE 1

READING/WRITING LEVEL by ORAL LEVEL
Cell contents: Frequency/Row percent/Column percent

	Non Reader/ Writer	Limited Reader/ Writer	Competent Reader/ Writer	Total
Non Speaker	18 75.0 2.02	6 25.0 0.48	0 0.0 0.0	24 0.7 1
Limit Speaker	620 56.93 64.97	419 38.48 33.82	50 04.59 03.95	18 9 32.08
Proficient Speaker	251 11.00 28.23	814 35.67 65.70	1217 53.33 96.05	228 2 67.22
Total	889 26.19	1239 36.19	1267 37.32	339 5

The confusion and failure to meet student need has been recently documented in a survey of Chapter 1 programs. According to the findings of the Council of Chief State Officers (1990), "it appears that there is no differentiation in the instructional services provided to Chapter 1-eligible students and the instruction provided to Non-LEP Chapter 1 students." In other words, LEP students are treated in the same way as other lower achieving students.

The lack of clarity surrounding language proficiency and the social circumstances defining it has also led to problems in both the study of bilingualism and the evaluation of Title IV programs. De Avila and Duncan (1980) reviewed over one hundred studies on the effects of bilingualism (proficiency in two languages) conducted in the United States over the past fifty years and found that in only a few cases (four) the actual extent of bilingualism was assessed. With rare exceptions, subjects were grouped on the basis of ethnicity; proficiency was assumed without distinguishing it from dominance. They found that the failure to control for the absolute language proficiency of comparison groups had resulted in a confounding of language with intellectual development and cognitive style.

Confusion over language proficiency has also led to confounding with social class and other variables. The paper by Dunn (1989) on the intelligence of language minorities is a good example of confounding of language proficiency and language minority group membership with social class variables.

In the area of program evaluation there is an even better example of how the failure to distinguish proficiency from dominance and to operationalize the former has resulted in fifteen years of equivocal results. The point is well illustrated in the metaanalyses of bilingual education evaluations conducted by Baker and De Kanter (1983) and in the reanalysis by Willig (1985). A major difference between the two sets of analyses, however, is the fact that while Baker and De Kanter (and a good many others) made no attempt to account for proficiency differences, Willig points out that

> it is apparent that the equating of experimental and comparison groups on ABSOLUTE proficiency in both languages is imperative if one is to make fair comparisons for purposes of educational evaluation.

Willig's point is underscored by a survey by Development Associates (cited above), which found that fewer than two percent of programs include assessment of home language proficiency in program placement decisions.

While Baker and others (see Rosell, 1989) interpret the lack of clear cut findings to conclude that bilingual education does not work, Willig argues for random assignment in future studies. The technical and moral intricacies of random assignment have been discussed at length by Campbell (1969). A review of Campbell's comments is strongly recommended.

My own conclusion is that we are confronted with the same problems as confronted us in the discussion of the entry-exit process. Until language proficiency is defined in operational terms and tests have been successfully validated against this definition (or other definitions), evaluation of Title VII programs and research on bilingualism will be compromised and equivocal.

Although Sharon Duncan and I have been working on the problem from our own point of view for some time and are encouraged by our progress, I am not very optimistic about the near future, particularly when I review the most recent Chapter 1 models for selecting LEP students and Title VII research designs. Neither seems to recognize the importance of the distinctions discussed above. Moreover, to leave these complexities up to the schools in the name of flexibility seems a cruel hoax, designed for failure. Until recently, with the funding of the Evaluation Centers for Title VII programs, it seemed that whatever good came out of local attempts has been in spite of the confusion exhibited by the leadership.

Taking Language Proficiency Apart and Putting it Back Together In A New Way

In the following section I describe the process whereby a set of test items (created, I hope, in an "artful" manner) is used to develop a set of subscales (that mirror educational/linguistic values) and then, from these subscales, several types of total or summary scores. One such set of summary scores comes from the combination of listening, speaking, reading and writing. The juxtaposition of literacy and oral skills taken from two different tests leads to the Language Proficiency Index, a nominal scale, intended to describe different types of students in a qualitative manner.

The creation of a score on an interval scale is the next step in the process. By recombining listening, speaking, reading and writing according to information processing principles of input and output, we were able to move from the nominal categories of the LPI to an ordinal scale. Finally, we were able to move from a simple ordinal scale to a ratio scale and then on to an interval scale.

The Language Proficiency Score (LPS) is possibly a useful metric upon which to conduct a wide variety of statistical and mathematical operations. The LPS is based on a good deal of information reduction. From over 200 test items and exercises a single number is created. The question becomes whether or not that single number reflects reality or the original empirical information represented by the LPI described below. As will be seen, the process yields some rather interesting results.

Combining Apples and Oranges: Creating a Language Proficiency Score

Language proficiency has been described by various state and federal regulations as consisting of listening, writing, speaking and listening. In previous work, Duncan and De Avila discussed the merits of considering literacy and oral skills simultaneously when making language proficiency determinations. Toward this end, they introduced the concept of the Language Proficiency Index (LPI; see below), which they defined as the student's level of oral proficiency relative to his or her reading/writing level.

Although the LPI offers a face-valid nominal categorization scheme, in that it clearly illustrates that not all LEPs are the same, it suffers from some of the same problems as other nominal categories in that direct comparisons are not possible. Moreover, since the scheme is based on categorical distinctions and not on ordinal scores, it is difficult to plot growth other than by counting the numbers of students in each category. The LPS represents an attempt to combine all four into a single score in order to create an ordinal scale that, in turn, can be transformed into an equal interval scale.

LANGUAGE PROFICIENCY INDEX

LPI	Category	Description
1/2 1/3	LEPa	low level R and W skills; mid level (limited) L and S skills
1/4 1/5	LEPb	low level R and W skills; high level (proficient) L and S skills
2/1 2/2 2/3	LEPc	mid level R and W skills; mid level (limited) L and S skills
2/4	LEPd	mid level R and W skills; high level (proficient) L and S skills
3/1		high level R and W skills; low level (limited) L and S skills
3/2 3/3	LEPe	high level R and W skills; mid level (limited) L and S skills
3/4 3/5	FEP	high level R and W skills; high level (proficient) L and S skills

(RW/LAS-O)* R = reading; W = writing; L = listening; S = speaking

A seemingly straightforward approach to the creation of a single score would be to add scores for each test component together on the assumption that their respective scales are compatible. Several examples shown below were taken from actual scores on one of the three tests that measure all four skill areas:

	Reading/Writing				Oral				
	Reading		Writing		Listening		Speaking		LPI
Exp1	60	+	50	+	90	+	70	= 270	2/4
Exp2	40	+	30	+	95	+	95	= 270	1/5
Exp3	70	+	65	+	70	+	65	= 270	2/3

The problem with this approach should be obvious. While the three examples show rather different combinations of scores, they all received the same total of 270. Similar problems occur with other approaches that fail to distinguish between different configurations. The examples show that an additive model would be unable to distinguish between a proficient speaker who was a limited reader/writer and a proficient speaker with no literacy skills at all. Similarly, neither could be distinguished from a student who was limited in both areas.

The Language Proficiency Score represents an attempt to create a single metric that produces a unique score for different configurations of skills and is ordinarily distributed. To accomplish this requires that we look at language proficiency from a slightly more abstract point of view.

Virtually any communicative act can be described as consisting of three elements, including the input, processing and ultimate output. Affect, within this model, serves as a modifier variable, as a filter or perceptual/ motivational set (see Haber, 1966). It drives the system by focusing the perceptual apparatus on the communicative act. The input/output continuum can be seen in a variety of corresponding ways depending on discipline. For example, linguists and educators alike have referred to receptive and expressive skills. Others have described the process as reception and production and applied the concepts to human and machine communication alike. Probably the strongest influence in this direction has come from cognitive psychology particularly from information processing theories (see Neisser, 1968).

The present approach largely represents an integration of linguistic structures and psychological processes. It also bears some limited resemblance to the ill fated "language skills framework" developed by SWRL as part of a large effort to construct entry/exit criteria and associated assessment procedures for bilin-

gual programs. The SWRL project, however, caved in under its own weight, leaving districts to develop their own tests and procedures. The full system was complex and ponderous; it is doubtful that it was ever used in a real setting.

The following is a fundamentally operational definition of "limited English proficiency." It takes the four aspects of language proficiency as they apply to a school context, and defines them accordingly.

Defining Limited English Proficiency (LEP)

Definition of terms

"LEP student"

A student from a home language other than English whose language proficiency is such that the probability of his or her success in a mainstream/regular classroom is less than that of comparable students.

"Comparable Students"

Students whose scores on standardized tests of academic achievement are at or about the national average or some other agreed upon cutoff score.

"Language Proficiency"

The set of combined skills in four linguistic domains including reading, writing, listening and speaking. The combination of scores from these domains shall be reflective of the continuous natural variation in skill levels for the four domains defining language proficiency.

"Probability of Success"

The relative proportion of language minority students passing a standardized test of academic achievement as a function their language proficiency.

Thus, given scores on a student's language proficiency in each of the four domains, it should be possible to state a student's "chances for success" in mainstream academic subject matter. For example, given X score on a test of language proficiency, a student may be said to have Y chances for a passing score at or above an agreed upon value.

The value of the approach is that all terms included in the definition can be operationally stated. Precise agreement with respect to cutoffs, proportions and all other numerical values can be empirically stated or established. Finally, additional values (e.g., teacher judgments) can be added to the equations in order to adjust borderline scores at or about the cutoffs. The question becomes one of deciding how to put things together.

Regardless of discipline, the constant or invariant process underlying all forms of communication seems to be that the transmission of information involves, at a minimum, a "sender" of the information and a "receiver" of the information. How the information is received, processed and ultimately transmitted is at the very heart of the study of mental processes. It is not our purpose here to review the wealth of literature in this area but rather to borrow the metaphor in a relaxed sense. We are not interested in hectic academic controversies surrounding the fine grain detail of the theories. Rather we have employed the distinction as a shorthand, a way in which to group phenomena hitherto grouped by default test usage.

Given this approach, traditional approaches to assessment by large publishers would seem misdirected. Rather than grouping oral language skills separately from literacy skills, it would seem more appropriate to group them according to the above distinction. In this way reading and listening would be grouped under the rubric of "input," "reception" and so on. Speaking and writing would be grouped under "Output," "Production" and so on.

The Language Proficiency Score distinguishes input (receptive skills) from output (productive skills) as shown in the following:

LPS = (Listening + Reading) + 2*(Speaking + Writing)

Speaking and writing are weighted by a factor of two in order to reflect the relative importance of production over reception. Language Proficiency Scores for the three examples would be recalculated as shown below:

	Receptive (Input)				Productive (Output)				
	Reading		Listening		Speaking		Writing		
Exp1.	60	+	90	+	2 x (50	+	70)	=	390
Exp2.	45	+	95	+	2 x (30	+	95)	=	380
Exp3.	70	+	70	+	2 x (65	+	65)	=	400

The reorganization of reading, writing, listening and speaking according to the above discussion has the effect of spreading out the different test configurations shown in the three examples. A major question to be addressed in the future will be the extent to which the spreading out of scores faithfully reflects the qualitative information contained in the LPI.

Sharon Duncan and I are currently involved in a number of studies to examine more closely the empirical side of the approach. For example, in one set of analyses we examined the relationship of Relative Language Proficiency to the Language Proficiency Score in an effort to see to what extent qualitative (LPI, nominal) and quantitative (LPI, ratio/interval) scales mesh. To test the proposition that the LPS reflects the LPI we ran a series of ANOVA as discussed in the preceding section. Results were supportive of the approach.

Probably one of the most immediate uses of the LPS is as an eligibility score in the sense described by Campbell (1969). "When several criteria are available they can be combined statistically into a single eligibility score."

In the near future we plan to examine further the nature/value of the topology. Toward this end we plan to generate profile analyses of the most frequently encountered LPIs selected from the frequency distributions of LPIs such as shown on Table 2 below.

TABLE 2

Frequency Distribution: Language Proficiency Index or
STUDENT TOPOLOGY

Reading & Writing Level / Oral Proficiency Level

LP I	N	%	LP I	N	%	LP I	N	%
1/1	66	.024	2/1	31	.014	3/1	0.0	0.0
1/2	11	.050	2/2	67	.030	3/2	36	0.16
1/3	46	.02	2/3	157	.157	3/3	99	.045
1/4	7	.001	2/4	491	.222	3/4	366	.211
1/5	0	0.0	2/5	248	1.112	3/5	269	1.222

Table 2 can be viewed in a variety of ways. For example, it is worth noting that the most frequent LEP type, (the 3/4 and 3/5 were comprised exclusively of monolingual English speakers) was the 2/4 or students with proficient oral skills but limited reading and writing skills. Given the above discussion on the relationship between Title VII and Chapter 1, these would be eligible for Chapter 1 services. On the other hand, the data seem to indicate that roughly 28 percent of those included in the study would have been identified as eligible for Title VII services. In summary, there were three groups or classes of students identified by the above approach. They include students who are limited in both reading/writing and oral skills, students who are limited in reading/writing skills but orally proficient and students who are proficient in both oral and reading/writing skills. The small percentage of students who were competent readers/writers but limited in oral skills (3/2 and 3/3) tended to be secondary level recent arrivals who had studied English in the homeland. Finally, notice that there were virtually no 1/5 or 3/1 students.

Future analyses will be directed toward a more detailed (more qualitative/quantitative) analysis of the profiles of different student types based on the Language Proficiency Index.

The Moral Imperative

The problems discussed above arise out of the fact that a great many students need special help, and there are limited funds to go around. Hence, a selection process is required to identify the most needy. The process of selecting who gets served affects not only who goes in or out of the program but how evaluations of program effectiveness are conducted. Random assignment, which is the preferred procedure from the point of view of the evaluator/researcher, is morally unacceptable from the point of view of the program provider. Thus, there is an underlying moral issue.

The moral issue itself, however, offers a unique opportunity for rigorous evaluation. Campbell (1969) has described a number of points in this connection. Campbell's comments regarding random assignment and evaluation/experimental design are particularly relevant to the present discussion in which identification and assignment to experimental or control group become one and the same.

OBEMLA has outlined five approaches (models) to assess program effectiveness (see Rayford, et al.). They include pre/post, gap reduction, nonequivalent comparison group, grade cohort and regression discontinuity. In addition, eleven threats to validity have been identified as having various effects on the five evaluation models listed above. Threats to the internal validity of program evaluation range from confounding the effects with those

of another to ill feelings on the part of students not receiving program benefits. Many of these threats have been alluded to in the above discussion, particularly issues having to do with instrumentation, comparability of pre/post measures, parallel forms and so on.

Of principal concern, according to OBEMLA, is that evaluations use a nonproject comparison group as similar as possible to the project group (i.e., the Title VII students) except that they do not participate in the program. The identification and selection of control group students is fraught with difficulty. Moreover, the most rigorous designs may be impossible to implement because of problems over and beyond the resources of all but a few projects. In fact, I would argue that, as design specifications become more sophisticated and able to accommodate greater control over threats to internal validity, the ability of projects to implement such models is actually lessened. Not only are the demands on the instrumentation and tests greater but so are demands on selection procedures, project personnel and financial resources.

In fact, given the above discussion on entry/exit issues and tests, implementation of the more powerful models (e.g., gap reduction and regression discontinuity) is impossible at the local level in all but a very few districts. Moreover, the funds necessary to conduct full scale evaluation are not readily available. Finally, project personnel are often reluctant to participate in evaluation, which appear to have effects on future funding and jobs. To do so, some may feel, is a form of self-incrimination. OBEMLA would do far better to expand the role of the evaluation centers in this area, leaving the more mechanical aspects of evaluation, such as data collection and testing functions, at the local levels.

While some districts have been able to take advantage of the Evaluation Assistance Centers, most have not. The same can be said of the now defunct Lau Centers. There are many reasons for this situation, a discussion of which would lead the present discussion somewhat off the present purpose or topic. Suffice it to say that without the Lau Centers things would have been even worse. Moreover, the EACs offer a ray of hope in that their mandate is directly related to assisting districts solve the problems under discussion. Hopefully, their role will be strengthened in the future.

Summary, Conclusions and Recommendations

We have reviewed a number of issues and problems associated with the creation and application of tests and decision models for determining entry/eligibility, placement/treatment, anu reclassification/exit processes used to remedy the limited English language proficiency of students from homes where English is not the primary language. Our review focused first on the problem of definition. Two concepts critical to the assessment process were discussed,

language dominance and language proficiency. It was argued that the concept of language proficiency is not only more linguistically sound and scientifically robust but more amenable to mathematical/statistical manipulation because of the known properties of the test score distributions. A number of the ideas used in this argument were then used to review some of the problems with current testing practices related to eligibility, placement and reclassification. It was also argued that the failure to work from a common set of definitions and principles has compromised not only the process of entry/exit but, in addition, both the evaluation of Title VII programs and research on the effects of bilingualism.

One of the major purposes of the discussion is to point out the need to disentangle (operationalize) cause (limited language proficiency) and effect (achievement in mainstream school settings) so that they can be reformulated into a meaningful calculus. One attempt was outlined. In the creation of the LPS we have attempted to maintain the sense of the concepts underlying the assessment of language proficiency while, at the same time, reducing the process to empirically testable steps. The need for other similar approaches should be self-evident.

A principal problem with currently used tests stems from a lack of conformity with standard psychometric practices. It is important to bear in mind that the responsibility for fair testing practices resides with test developer and test user alike. Both have their responsibilities. In this regard the American Psychological Association, working in collaboration with the American Educational Research Association, the National Council of Measurement in Education and the Canadian Psychological Association, has recently prepared a document on the "development of a code of fair testing practices." The practices outlined in the publication of the proceedings of the Joint Committee on Testing Practices (Fremer, Diamond & Camara, 1989) are as applicable to language proficiency testing as to the testing of the general population. I strongly recommend the review of these proceedings to test developer, user and reviewer alike.

In many respects the guidelines outlined by the JCTP may be useful as a way to standardize not only the development of tests but their use and evaluation. Developers, users and reviewers would be working from a common base. Developers would know what was expected of them. Users would know what to expect, and reviewers would have a common ground on which to base their reviews.

The JCTP has outlined elements that should be borne in mind by test developers and test users. The first addresses the processes of developing and selecting tests. The sixteen practices concern such test development measures as "explaining relevant measurement concepts as necessary for clarity" and, from the point of view of the test user, becoming familiar with how the test was

developed and tried out. Similarly, provision of interpretable test scores requires care on the part of the developer as well as vigilance on the part of the user.

Both test developer and user also have responsibilities to the test takers, to inform them of the purpose of testing and of the uses to which the information will be put. The major issue for the JCTP concerns the fairness of test use. There are practices that can be exercised by both developer and user to make tests fairer to students. In fact, the issue for many is not so much the validity of tests but rather the fairness with which they are used. It appears, however, that being fair is not as easy as one would think.

REFERENCES

Anderson, C., Krahnke, K., & Stansfield, C. (1987). A review of oral language tests. Rosslyn, VA: TESOL.

Baker, C. (1988). Normative testing and bilingual populations. Journal of Multilingual and Multicultural Development, 9(5).

Baker, K., & de Kanter, A. (1983). Federal policy and the effectiveness of bilingual education. In K. Baker & A. de Kanter (Eds.), Bilingual education: A reappraisal of federal policy. Lexington, MA: Lexington Books.

Barnes, R. Estimates of LEP students in the US, a reanalysis of O'Malley's data. As cited in Secada, W. (in press). Race, ethnicity, social class, language and achievement in mathematics. In D. Grouws (Ed.), Handbook of research on the teaching and learning of mathematics. New York: Macmillan.

Berko, J. (1989). The development of language. New York: Merrill Publishing Company.

Cabello, B. (1983). A description of analyses for the identification of potential sources of bias in dual language achievement tests. NABE Journal, 7(2).

Campbell, D. (1969). Applied social experiments: Reforms as experiments. American Psychologist, 24(4).

Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multicultural-multimethod matrix. Psychological Bulletin, 56.

Canales, J. (1990). Assessment of language proficiency: Informing policy and practice. Mimeo, Southwest Educational Development Laboratory.

Chamot, A., & O'Malley, M. (1986). A cognitive academic approach: A content-based ESL curriculum. Wheaton, MD: National Clearinghouse for Bilingual Education.

Chavez, H., Dulay, H., & Burt, M. (1978). Language dominance and proficiency testing: Some general considerations. NABE Journal, 3(1).

Council of Chief School State Officers. (1989). A concern about . . . educating limited English proficient students: A CCSSO survey of State Education Agency activity. Concerns, 26.

Crawford, J. (1989). Bilingual education: History, politics, theory and practice. Trenton, NJ: Crane Publishing.

Cummins, J. (1984). Bilingualism, language proficiency and metalinguistic development. In P. Homel, et al. (Eds.), Child bilingualism: Aspects of linguistic, cognitive and social development. Hillsdale, NJ: Lawrence Earlbaum.

Cummins, J. (1980). The entry and exit fallacy in bilingual education. NABE Journal, 4.

Danoff, M. (1977). Evaluation of the impact of ESEA Title VII Spanish/ English programs. Palo Alto, CA: American Institute for Research.

De Avila, E. (1987). Bilingualism, cognitive function, and language minority group membership. In P. Homel, M. Pail, & D. Aaronson (Eds.), Childhood Bilingualism: Aspects of linguistic, cognitive and social development. Hillsdale, NJ: Lawrence Earlbaum Assoc., Publishers.

De Avila, E. (1984). Language proficiency: confusions, paradoxes, and a few admonitions to psychologists, linguists and others developing assessment procedures for language minority students. In C. Rivera (Ed.), Placement procedures in bilingual education: Education and policy issues. Avon, England: Multilingual Matters Ltd.

De Avila, E., & Duncan, S. (198?). A convergent approach to language assessment: Theoretical and techn specifications on the language assessment scales (LAS). San Rafael, CA: Linguametrics Group.

De Avila, E., & Duncan, S. (1980). Definition and measurement of bilingual students. In Bilingual program, policy and assessment issues. Sacramento, CA: State Department of Education.

De Avila, E., & Duncan, S. (1976). A few thoughts about language assessment: The LAU decision reconsidered. Paper presented at the Conference on Research and Policy Implications of the Task Force Report of the U.S. Office of Civil Rights. Southwest Educational Development Laboratory, Austin, TX.

De Avila, E., Cervantes, R., & Duncan, S. (1978). Bilingual exit criteria. CABE Research Journal, 1(2).

De George, G. (1988). Assessment and placement of language minority students: Procedures for mainstreaming. Focus, Publication of the National Clearinghouse for Bilingual Education, Occasional Papers, 3.

de Jong, J. (1987). Item selection from pretests in mixed ability groups. In C. Stansfield (Ed.), Testing and technology. TESOL, Washington, D.C.

Donaldson, M. (1978). Children's minds. Glasgow: Collins.

Dulay, H., & Burt, M. (1980). The relative proficiency of limited English proficient students. In J. Alatis (Ed.), Current issues in bilingual education. Washington, DC: Georgetown University Press.

Duncan, S., & De Avila, E. (1988). Technical report: Validity and reliability of the LAS reading and writing tests. Monterey, CA: CTB, McGraw/Hill.

Dunn, L. (1987). Bilingual Hispanic children on the U.S. mainland: A review of their cognitive, linguistic, and scholastic achievement. Circle Pines, NM: American Guidance Service.

Erlich, A., & Erlich, R. (1975). Tests in Spanish and other languages, English as a second language, and non-verbal tests for bilingual programs. Annotated B.E.A.R.U. Bibliography. Hunter College, NY: Project Best.

Embretson, S. (1985). Test design, developments in psychology and psychometrics. New York: Academic Press.

Fradd, S., & Tikunoff, W. (1987). Bilingual education and bilingual special education: A guide to administrators. Boston, MA: Little Brown and Company.

Freedle, R. (1990). Artificial intelligence and the future of testing. Hillsdale, NJ: Lawrence Earlbaum.

Fremer, T., Diamond, R., & Camera, M (1989). Developing a code of fair testing practices. American Psychologist, 44(7).

Fuerstein, R. (1980). Instrumental enrichment: An intervention program for cognitive modifiability. Baltimore, MD: University Park Press.

Gilmore, G., & Dickerson, A. (1979). The relationship between the instruments used for identifying children of limited English speaking ability in Texas. Unpublished mimeo, Region IV Education Service Center, Houston, TX.

Gonzales, J. (1984). Policy issues in language assessment. In C. Rivera (Ed.), Placement procedures in bilingual education: Education and policy issues. Avon, England: Multilingual Matters #12, Ltd.

Gottlieb, M. (1990). The role of communicative competence in first and second language achievement as demonstrated. Unpublished doctoral dissertation, University of Illinois at Chicago. Ann Arbor, MI: University Microfilms.

Guerin, G., & Maier, A. (1983). Informal assessment in education. Palo Alto, CA: Mayfield Publishing Company.

Gullikson, H. (1950). Theory of Mental Test. New York: John Wiley & Sons.

Haber, R. (1966). Nature of the effect of set on perception. Psychological Bulletin, 74.

Hige, R. D., & Colardarci, T. (1989). Teacher based judgments of academic achievement: A review of the literature. Review of Educational Research, 59(3).

Hoffmiester, A. (1975). Integrating criterion-referenced testing and instruction. In W. Hively & M. Reynolds (Eds.), Domain-referenced testing in special education. Reston, VA: Council for Exceptional Children.

Howard, D. V. (1983). Cognitive psychology: Theory, language, and thought. New York: Macmillan.

Jackson, S. (1980). Analyses procedures and summary statistics of the language data of a longitudinal study of the oral development of Texas bilingual children. Paper presented at the National Conference on Language Arts in the Elementary School, San Antonio, TX.

Jensen, A. (1980). Bias in mental testing. New York: The Free Press.

Kerlinger, F. (1973). Foundations of behavioral research. New York: Holt Rinehart and Winston.

Krashen, S. (1982). Principles and practices in second language acquisition. Oxford: Pergamon Press.

Mace-Matluck, B., Dominguez, D., & Turmer, W. (1979). Assessing Reading in Bilingual Programs. Mimeo, Southwest Educational Development Laboratory.

McArthur, D. (1987). Alternative approaches to the assessment of achievement. Norwell, MA: Kluwer Academic Publishers.

Merino, B., & Spencer, M. (1983). The comparability of English and Spanish versions of oral language proficiency tests. NABE Journal, 7(2).

Niesser, U. (1967). Cognitive Psychology. New York: Appleton-Century-Crofts.

O'Malley, M. (1981). Children's English services study: Language minority children with limited English proficiency in the United States. Rosslyn, VA: National Clearinghouse for Bilingual Education, InterAmerica Research Associates.

O'Malley, M. (1989). Language proficiency testing with limited English-proficient students. Georgetown Roundtable. Washington, DC: Georgetown University Press.

Oller, J., & Perkins, K. (1978). Language testing in education: Testing the tests. Rowley, MA: Newbury House.

Pietcher, B., Locks, N., & Reynolds, D. (1978). A guide to assessment instruments. New York: Santillana.

Quesada, R. (1979). Desegregation: Future issues and trends. Planning and Change, 10(2).

Rayford, L., & Thayer, K. (1990). Models for selecting limited English proficient students to receive Chapter I services. Unpublished mimeo, RMC Research Corporation, Hampton, NH.

Richardson, V. (1990). At risk programs and critical inquiry. Evaluation and Social Justice: Issues in Public Education, 45.

Rosansky, E. (1980). Issues of validity and reliability in language proficiency testing: A consideration of several instruments. Paper delivered at the 9th Annual International Bilingual Bicultural Conference, Anaheim, CA.

Rosell, C. (1989). The effectiveness of educational alternatives for limited English proficient children. In G. Imhoff (Ed.), The social and cultural context of instruction in two languages: From conflict and controversy to cooperative reorganization of schools. New York: Transaction Books.

Rosenbaum, H. (1980). The development and structure of the language skills framework of the student placement system for bilingual programs. In J. Alatis (Ed.), Current Issues in Bilingual Education. Proceedings of the Georgetown Roundtable on Languages and Linguistics.

Sanchez, R. (1976). A critique of oral language instruments. NABE Journal, 1.

Secada, W. (1990). Race, ethnicity, social class, language and achievement in mathematics. In D. Grouws (Ed.), Handbook of research on the teaching and learning of mathematics. New York: M cmillan.

Silverman, R., Noa, J., & Russel, R. (976). Oral language tests for bilingual students: An evaluation of language dominance and proficiency instruments. Portland, OR: Northwest Regional Educational Development Laboratory.

Spolsky, B. (1968). Language testing: The problem of validation. TESOL Quarterly, 2.

Sweetland, R., & Keyser, D. (986). Tests: A comprehensive reference for assessments in psychology, education, and business. Kansas City, KA: The Test Corporation of America.

Tallmadge, K., & Wood, C. (no date). ESEA Chapter 1 evaluation and reporting users guide. Unpublished mimeo, Redwood City, CA: RMC Corporation.

Tannen, D. (1982). The myth of orality and literacy. In W. Frawley (Ed.), Linguistics and literacy. New York: Plenum.

Tatsuka, M. (1969). Standardized scales: Linear and area transformations. Number 2. Champaign, IL: University of Illinois.

Torgeson, W. (1958). Theory and methods of scaling. New York: John Wiley & Sons.

Ulibarri, D. (1990). Use of achievement tests with non-native English speaking language minority students. In A. Barona (Ed.), Children at risk: Poverty, minority status, and other issues in educational equity. Tempe, AZ: University of Arizona Press.

Ulibarri, D., & Costa, M. (1979). Test-retest reliability of the language assessment scales. San Rafael, CA: Linguametrics Group.

Ulibarri, D., Spencer, M., & Rivas, G. (1981). Comparability of three oral language instruments and their relationship to achievement variables. NABE Journal, 5.

Walton, A. (1989). Exit criteria for limited English proficient students. Unpublished memo to New York Education Department Board of Regents, New York.

Wesman, A. (1971). Writing the test item. In R. Thondike (Ed.), Educational Measurement. Washington, DC: National Academy Press.

Willig, A. (1985). A meta-analysis of selected studies on the effectiveness of bilingual education. Review of Educational Research, 55(3).

Wonnacott, T., & Wonnacott, R. (1987). Regression: A second course in statistics. Malabar, FL: Robert E. Krieger Publishing Company.