DOCUMENT RESUME

ED 227 692 FL 013 563

AUTHOR Spencer, Mary
TITLE Testing Instruments, Packet 2. Bilingual Program Planning, Implementation, and Evaluation, Series A. Teacher Edition. Bilingual Education Teacher Training Packets.
INSTITUTION Evaluation, Dissemination and Assessment Center, Dallas.
SPONS AGENCY Department of Education, Washington, DC.
PUB DATE 82
NOTE 78p.; For related documents, see FL 013 562-564.
AVAILABLE FROM Evaluation, Dissemination and Assessment Center, Dallas Independent School District, Dallas, TX 75204 ($1.50).
PUB TYPE Guides - Classroom Use - Guides (For Teachers) (052) -- Reports - Descriptive (141)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Bilingual Education Programs; Instructional Materials; *Language Proficiency; *Language Tests; Program Evaluation; Program Implementation; Teacher Education; Teaching Guides; *Test Reliability; *Test Validity

ABSTRACT

This teacher's edition of training materials on bilingual program planning, implementation, and evaluation focuses on testing instruments. The guide is part of a series directed at bilingual educators and intended for use in institutions of higher education and inservice teacher education programs. Training objectives, a pretest and posttest, and instructional materials and activities are included. The topics covered include (1) the role of testing in bilingual education programs, (2) norm-referenced and domain-referenced language proficiency tests, (3) indirect and direct measures of language proficiency, (4) test validity and reliability, (5) norms, (6) reviews of nine popular language assessment instruments, and (7) examples of direct measures of language. (RW)

Teacher Edition.

Bilingual Education Teacher Training

Packets.

SERIES A: BILINGUAL PROGRAM PLANNING, IMPLEMENTATION, AND EVALUATION

PACKET 2: TESTING INSTRUMENTS

developed by:

DR. MARY SPENCER

The project reported herein was performed pursuant to a Grant from the U.S. Department of Education, Office of Bilingual Education and Minority Languages Affairs. However, the opinions expressed herein do not necessarily reflect the position or policy of the U.S. Department of Education, and no official endorsement of the U.S. Department of Education should be inferred.

This publication was printed with funds provided by Title VII of the Elementary and Secondary Education Act of 1965, as amended by Public Law 95-561.

Published by
Evaluation, Dissemination and Assessment Center—Dallas
Dallas Independent School District
Dallas, Texas 75204
(214) 742-5991

CONTENTS

LIST OF EXHIBITS

BILINGUAL EDUCATION

TEACHER TRAINING MATERIALS

The bilingual education teacher training materials developed by the Center for the Development of Bilingual Curriculum - Dallas address five broad areas of need in the field of bilingual education:

Series A:	Bilingual Program Planning, Implementation, and Evaluation
Series B:	Language Proficiency Acquisition, Assessment, and Communicative Behavior
Series C:	Teaching Mathematics, Science, and Social Studies
Series D:	Teaching Listening, Speaking, Reading, and Writing
Series E:	Actualizing Parental Involvement

These materials are intended for use in institutions of higher education, education service centers, and local school district in-service programs. They were developed by experts in the appropriate fields of bilingual education and teacher training.

Series A addresses the critical issue of the effective planning and implementation of programs of bilingual education as well as efficient program evaluation. Sample evaluation instruments and indications for their use are included. Series B contains state-of-the-art information on theories and research concerning bilingual education, second language acquisition, and communicative competence as well as teaching models and assessment techniques reflecting these theories and research. In Series C, the content, methods, and materials for teaching effectively in the subject matter areas of mathematics, science, and social studies are presented. Technical vocabulary is included as well as information on those

aspects rarely dealt with in the monolingual content area course. Series D presents the content area of language arts, specifically the vital knowledge and skills for teaching listening, speaking, reading, and writing in the bilingual classroom. The content of Series E, Actualizing Parental Involvement, is directed toward involving parents with the school system and developing essential skills and knowledge for the decision-making process.

Each packet of the series contains a Teacher Edition and a Student Edition. In general, the Teacher Edition includes objectives for the learning activity, prerequisites, suggested procedures, vocabulary or a glossary of bilingual terminology, a bibliography, and assessment instruments as well as all of the materials in the Student Edition. The materials for the student may be composed of assignments of readings, case studies, written reports, field work, or other pertinent content. Teaching strategies may include classroom observation, peer teaching, seminars, conferences, or micro-teaching sessions.

The language used in each of the series is closely synchronized with specific objectives and client populations. The following chart illustrates the areas of competencies, languages, and intended clientele.

COMPETENCIES, LANGUAGE OF INSTRUCTION AND INTENDED CLIENTELE

AREAS OF COMPETENCIES	LANGUAGE	CLIENTELE
SERIES A. Bilingual Program Planning, Implementation, and Evaluation	English	Primarily supervisors
SERIES B. Language Proficiency Acquistion, Assessment, and Communicative Behavior	Spanish/ English	Primarily teachers and supervisors
SERIES C. Teaching Mathematics, Science, and Social Studies	Spanish/ English	Primarily teachers and paraprofessionals
SERIES D. Teaching Listening, Speaking, Reading, and Writing	Spanish/ English	Primarily teachers and Paraprofessionals
SERIES E. Actualizing Parental Involvement	Spanish	Primarily teachers, parents, and community liaisons

In addition to the materials described, the Center has developed a Management System to be used in conjunction with the packets in the Series. Also available are four Practicums which include a take-home packet for the teacher trainee.

The design of the materials provides for differing levels of linguistic proficiency in Spanish and for diversified levels of knowledge and academic preparation through the selection of assignments and strategies. A variety of methods of testing the information and skills taught in real or simulated situations is provided along with strategies that will allow the instructor to meet individual needs and learning styles. In general, the materials are adaptable as source materials for a topic or as supplements to other materials, texts, or syllabi. They provide a model that learners can emulate in their own classroom. It is hoped that teacher trainers will find the materials motivational and helpful in preparing better teachers for the bilingual classroom.

OBJECTIVES

Upon the completion of this Packet, the student will be able to:

Part I:

1. Enumerate the different roles of testing in Title VII programs.
2. Understand how different definitions of language proficiency will affect the nature of tests used to determine eligibility for Title VII services, and to determine when a student is ready to be reclassified (exited).
3. Understand the necessary properties of tests used for varying functions and purposes.

Part II:

4. Name and describe the two different ways of constructing tests.
5. Name the common uses of indirect measures of language proficiency.
6. Discuss the possible sources of bias and distortion in indirect measures of language proficiency.
7. Name the four linguistic systems and identify tests which measure each.
8. Explain the concept of relative language proficiency, and discuss the debate about it on the basis of the research.
9. Name and describe the linguistic parameters assessed by language tests.
10. Identify the domains of language used in various language tests.
11. Explain the difference between discrete point and integrative testing techniques, giving the advantages and disadvantages of both.
12. Identify the extent to which a test addresses BICS and CALPS.

Part III:

13. Define the various types of validity and apply these standards to the review of language tests.
14. Define the various types of reliability and apply these standards to the review of language tests.
15. Define the meaning of the term norm and apply this to the review of language tests.

Part IV:

16. Understand the relative psychometric qualities of several popular language tests.

PRE-TEST

Part I:

1. Name the major purposes of testing in a Title VII program.

2. Give three different definitions of language proficiency.

3. Explain what properties a test should have to measure each of the different definitions of language proficiency.

4. Explain what different properties a test should have in order to be appropriate to 1) classification of students for eligibility for Title VII services; 2) for program evaluation; 3) for individual student diagnosis.

Part II:

5. What are the two basic ways of constructing tests? Describe each and contrast.

6. What is the difference between indirect and direct measures of language proficiency?

7. What are some common sources of bias and distortion in indirect measures?

8. What is the conclusion of research findings on the concept of relative language proficiency?

9. What are the major linguistic parameters? Name and describe. What are the pros and cons of using each to assess language?

10. What is the difference between discrete point and integrative testing techniques?

Part III:

11. What is validity? Define it in general and specify three types of validity.

12. What is reliability? Define it in general and specify five different types of reliability.

13. Define what norms are.

PRE-TEST ANSWERS

Part I:

1. Name the major purposes of testing in a Title VII program. (p. 7)

2. Give three different definitions of language proficiency. (p. 17)

3. Explain what properties a test should have to measure each of the different definitions of language proficiency. (p. 18-19)

4. Explain what different properties a test should have in order to be appropriate to 1) classification of students for eligibility for Title VII services; 2) for program evaluation; 3) for individual student diagnosis. (p. 7-10)

Part II:

5. What are the two basic ways of constructing tests? Describe each and contrast. (p. 11-13)

6. What is the difference between indirect and direct measures of language proficiency? (p. 14)

7. What are some common sources of bias and distortion in indirect measures? (p. 14-15)

8. What is the conclusion of research findings on the concept of relative language proficiency? (p. 16-17)

9. What are the major linguistic parameters? Name and describe. What are the pros and cons of using each to assess language? (p. 17)

10. What is the difference between discrete point and integrative testing techniques? (p. 21-22)

Part III:

11. What is validity? Define it in general and specify three types of validity. (p. 26-31)

12. What is reliability? Define it in general and specify five different types of reliability. (p. 33-35)

13. Define what norms are. (p. 36)

PART I: THE ROLE OF TESTING IN TITLE VII PROGRAMS

Testing is used for at least five different purposes in Title VII programs: 1) classification of students for program entry; 2) classification of students for transition (or exit) from the program; 3) diagnosis of student strengths and weaknesses; 4) program evaluation; and 5) program planning. Through careful planning, it is often possible to coordinate the testing program in a way that minimizes the number of tests given, and therefore the test cost and burden. These savings can be achieved when the data derived from one test can legitimately be used to serve more than one of the purposes listed above. However, each test is created to serve very specific purposes. If test results are used for purposes with which they are incompatible, only invalid and unreliable consequences can be expected.

The task of selecting a test which is appropriate to the purpose for which it will be used, and which has evidence of validity and reliability is a serious task. The stakes are high if errors of judgment are made. Children may be denied important educational opportunities. They may be subjected to educational experiences which inhibit their development. Or, they may be categorized in ways that give their teachers and parents inappropriate expectations for their growth - or worse yet, which stigmatize them. The inappropriate use of educational and psychological tests to classify children has an onerous history (e.g., Oakland, 1977). Many major law suits have been fought over the violation of children's civil rights brought on by the misuse of tests. Therefore, the choice of tests should always be considered as a decision of grave import - not only for the program, but for the welfare of the individual student as well.

The five purposes listed above may be simplified somewhat. Tests are viewed by many experts as serving two very broad functions. The first is to classify children in the sense of declaring their eligibility for placement in special programs. Most of the legislation, litigation, and

judicial action attending educational testing has focused on this issue of classification. The second broad function revolves around the planning and evaluation of curriculum and instruction, in which tests are used to develop and provide information to students, parents, and teachers for the purpose of describing the student's status and progress and to acquire information used to decide upon the subsequent content and methods of instruction. This information may also be used to evaluate educational programs and to plan for their future development or alteration. According to the National Educational Association (1973), it is this second broad function which should constitute the major use of tests:

> "The major use of tests should be for the improvement of instruction - for diagnosis of learning difficulties and for response to learning needs. They must not be used in any way that will lead to labeling and classifying of students, for tracking into homogeneous groups as the major determinants to educational programs, to perpetuate elitism, or to maintain some groups and individuals "in their place" near the bottom of the socioeconomic ladder. In short, tests must not be used in a way that will deny any student full access to equal educational opportunity."[1]

For Title VII programs, the task of selecting a test to classify students as eligible for services hinges on the language of the law itself. The Bilingual Education Act (1978) requires that bilingual educational programs be developed and provided to "children of limited English proficiency" in order to enable them, while using their native

---

[1] NEA's reference to "labeling and classifying students" reflects a concern for stigmatizing labels such as mentally retarded or culturally deprived. The statement about "tracking into homogeneous groups as the major determinants to educational programs" referred to the use at that time of IQ tests to track Black students into special education classes in which their educational opportunities were reduced. Although we seldom think of this situation as applicable to bilingual education programs, it serves as a warning as to the importance of ensuring quality in our bilingual education programs.

language, to achieve competence in the English language. The law went on to define "limited English proficiency" as:

1. individuals who were not born in the United States or whose native language is a language other than English;
2. individuals who come from environments where a language other than English is dominant, as further defined by the Commissioner by regulation, and;
3. individuals who are American Indian and Alaskan Native students and who come from environments where a language other than English has had a significant impact on their level of English language proficiency, subject to such regulations as the Commissioner determines to be necessary;
4. and, by reason thereof, have sufficient difficulty speaking, reading, writing, or understanding the English language to deny such individuals the opportunity to learn successfully in classrooms where the language instruction is English.

The 1978 Title VII legislative language provides several notions of limited English proficiency on which classification might rest: non-U.S. birth place; native language other than English; environments where a language other than English is dominant; environments where an American Indian or Alaskan Native language has significantly impacted English proficiency; or sufficient difficulty in speaking, reading, writing, or understanding English exists so as to deny the student of educational opportunity if placed in an English only classroom. Ultimately, the selection of a definition of limited English proficiency and the way it is operationalized into eligibility status is the responsibility of individual states. Exhibit I, P. 39 shows there is considerable diversity among the states on the issue of defining limited English proficiency and thereby determining under what conditions a student is eligible for bilingual education services.

As shown in Exhibit I, p. 39, most states depend heavily upon tests to identify students as LEP. Tests of oral language proficiency, which usually provide measures of speaking and understanding, are central to the determinations of many states. The relative standing of a potential Title VII student to a comparison group (such as a district or national norm group) on a standardized test of reading and writing achievement is part of the eligibility criteria of many states.

Title VII's legislative language gives few clues about how students in bilingual education programs should be re-classified. Exhibit II, p. 41 provides examples of how some states approach this other major classification task (Series B, Packet III provides more information on this topic). Again, tests of oral language proficiency as well as standardized tests of achievement play important roles in this process. While classification for entry appears to focus more often on the former, achievement testing is given increased emphasis in reclassification considerations. In some states (e.g., California) reclassification methods stress the desirability of using decision making teams which employ multiple indicators, and conducting follow-up checks on students' progress once exited in order to ensure that the proper placement has been made.

Very often, a single test is used for all purposes and functions. To judge whether it is appropriate to use a particular test for a particular function, consider the criteria presented in Exhibit III, p.42.

PART II: DIFFERENT TYPES OF LANGUAGE PROFICIENCY TESTS

Two Different Ways of Constructing Tests

In order to understand how to use tests and what to expect of their results, it is necessary to understand two fundamental ways of constructing tests: norm referenced tests and domain referenced tests.[1] As Hively (1974) has phrased it, "The world of psychometrics may be seen as a contrast between Domain Referenced Testing and Norm Referenced Testing." This same distinction was drawn by Glaser (1971), although he preferred the term criterion to domain.

Norm Referenced Testing. The goal of norm referenced testing is to differentiate among people. This goal is central to the way in which test items are created, refined, and selected or deselected. Since items which are answered correctly or incorrectly by most respondents do not achieve the goal of differentiation, most are eliminated. It is not important in norm referenced testing to measure what the majority of respondents can do or cannot do. Instead, focus is on what some can do and some cannot do. In practice, only items on which 40 to 60 percent of the respondents of a defined group (e.g., third graders) answer correctly are usually included in a norm referenced test.

This practice slants the focus of norm referenced tests. Since their function is to differentiate among people rather than to provide a representative measure of some body of knowledge or behavior, norm referenced tests may be rather fuzzy about their subject matter structure. It is of greater importance, in terms of their goal and function, to exhibit certain psychometric qualities such as internal consistency (success on each item is positively correlated with success on the test as a whole), comparability (individuals obtain similar scores on alternate

---

[1] We have adopted Hively's use of the term Domain Referenced because we agree with his statement that the term Criterion Referenced carries with it the surplus meaning of mastery learning which often leads to misinterpretation.

forms), stability (individuals obtain similar scores when retested), concurrent validity (individuals obtain similar or highly related outcomes on two or more other related measures), or predictive validity (individuals' norm referenced test score is correlated with events such as graduation, grade point average, or success in a related training program). However, the goal of differentiation does not require that the items of a norm referenced test present an unbiased picture of the over-all content of some body of knowledge or behavior which is obviously outside of this body. Also unnecessary to this goal are item qualities which facilitate content transfer or generalization, or general principles of item generation which could have relevance to instruction.

These common attributes of norm referenced tests often give short-shrift to planned correspondence between educational assessment and educational goals. Yet, if other classic psychometric qualities - construct and content validity - were given careful attention in norm referenced testing, this correspondence would, by definition, be improved. However, since a distinctive quality of norm referenced testing is that the way items are created does not lend itself to calling up an indefinite number of parallel tests by systematically sampling from the defined content structure, norm referenced tests are not appropriate to frequent (daily, weekly, or even monthly) checks on student progress and the instructional conditions being applied. They frequently are also not suited for decisions about individual students since they are group referenced. Perhaps their most appropriate uses are to be found in program evaluation, program planning, and gross screening activities. Domain referenced tests are more likely to be appropriate for student diagnosis and for decisions that call for information on individual student status and progress over time.

Domain Referenced Testing. The purpose of domain referenced testing is to measure proficiency on a specified set of concepts or behaviors.

Domain referenced testing begins with the development of a rationale or theory of the subject matter structure. The definition of this structure describes the components of a certain body of knowledge or behavior, covering the range of important actual situations in which these occur. This theory of subject matter structure may be based on empirical findings (e.g., the Bilingual Syntax Measure for oral language acquisition), upon the specifications of educational decision makers (e.g., SOBAR Reading Tests), or upon some logical system. Whichever approach is used to define the content structure, the direct correspondence between educational goals and educational assessment is maximized at the level of instrument construction in domain referenced testing.

The goal of domain referenced testing is to create an extensive pool of items that will constitute a representative sample of the greater body of concepts or behaviors. It is considered crucial that the items of a domain referenced test incorporate the qualities of demands or problems actually encountered in the field. This requirement is linked to the other objective in that items should have a high degree of transfer or generalization to the universe from which they are sampled. The same standards of psychometric quality apply to domain referenced tests that are applied to norm referenced tests. However, the distinctive goal of domain referenced tests emphasizes content and construct validity. By definition, domain referenced tests are intended to be of service primarily to the instructional applications rather than the differentiation functions of testing.

Indirect and Direct Measures of Language Proficiency

Over the years many different approaches have been taken to the problem of determining language proficiency. One of the fundamental differences between these approaches is whether the individual's language is directly observed, sampled, and appraised, or whether instead someone is asked to estimate what that individual's language proficiency is. In theory, at least, it is possible to obtain fairly reliable and valid assessments either way. In practice, however, indirect measures have never been validated, and the number of reliable and valid direct measures is also wanting. There is room for much improvement in the years to come for both indirect and direct measures of language proficiency.

Indirect Measures. Frequently, indirect measures of language proficiency are used to obtain national or state estimates of the number of individuals who may be counted as either part of the language minority population or as part of the limited English-proficienct population. In addition, indirect measures have often been used by Title VII programs as a quick and inexpensive method of estimating need for services.

Indirect measures of language proficiency should always be regarded with caution. They are subject to several major sources of bias and distortion. The partial list below might be used as a checklist when one is considering how much faith to place in the results of an indirect measure of language proficiency.

POSSIBLE SOURCES OF BIAS AND DISTORTION

1. The respondent may not be qualified to appraise the language proficiency of the student, in either or both L_1 or L_2.
2. There may be a socio-political reason why the respondent may not be entirely candid about the appraisal. Perhaps the respondent's answer is influenced by factors such as fear or a desire to conform in some way in order to aid the student.
3. The answer may depend heavily upon the context of the respondent. For example, a person in the home of the student may state that the student speaks primarily L_1, a fact which may be influenced by the language spoken by others in the home.

However, a person at the student's school may state that the student speaks primarily L_2, a fact which will be influenced by the language spoken by those in the school.

4. The questions asked in the indirect measure are often not very good reflections of language proficiency. The exact wording of the questions must be examined in order to make accurate interpretations. For example, questions that ask a respondent to judge the relative proficiency of a student's English and non-English language are asking for such a complex appraisal that only a highly trained, very proficient bilingual speaker would be qualified to provide a valid answer. A question that asks for the country of origin, and then uses the number of individuals reporting a non-English language country as a measure of the number of limited English-speaking students is a serious distortion.

5. The economic and political motives of the source of the indirect measure should always be examined. For example, some indirect measures of language minority students and limited English-proficient students will yield much larger estimates than will others. To what use will the results of the indirect measure be put? If the purpose is to estimate the number of students who are language minority, and therefore may have a need for bilingual education services, the measure that yields the most comprehensive count would be preferable since it would most nearly capture all students in need of educational assistance. If the purpose is to identify students for eligibility, placement, or reclassification, an indirect measure is not sufficiently refined or psychometrically sound to do the job.

Appendix A provides a few examples of indirect measures of language characteristics that have been used in the major national studies and in a few selected states. Also provided are a few comparative results which show how the different wording of these measures can affect the results that they yield.

The validity and reliability of an indirect measure of language proficiency could be improved if a strong association could be established between the indirect measure of individuals and direct measures of these same individuals which have been independently validated. Once this association is established, the indirect measure could be used thereafter with some confidence for such purposes as program planning which do not involve educational decisions about individual children.

Direct Measures. A direct measure of language proficiency is one in which a systematic approach is taken to the direct observation, sampling, and evaluation of the proficiency of a student in a particular language. Rather than being a report from memory or a remote judgment of the student's proficiency as in the indirect measure, it is a first-hand calibration of specific language behavior.

It should be noted that direct measures of language proficiency vary on several key dimensions. Whenever a Title VII program is deliberating the selection or development of a direct measure of language proficiency, the status of the measure on these dimensions should be considered in order that the instrument chosen is compatible with the purposes to which it is to be put.

1. Linguistic Systems Assessed: Speaking, understanding, reading, and writing each represent an important language system. Few language proficiency tests address all four of these systems. More commonly, one will find tests of oral language proficiency which purport to measure speaking and understanding, sometimes providing separate scores for each and often providing an integrated measure of the two. In addition, one will find separate measures of reading and writing. A few instruments, such as the Language Assessment Battery (LAB) have attempted to assess all four systems.

2. Languages Assessed: Somewhat more than one half dozen oral language proficiency instruments provide both an English and Spanish version. Some of these same instruments provide a measure of a language other than English or Spanish. Only a very few standardized tests of reading and writing provide measures in both English and a non-English language (usually Spanish).

Considerable discussion has arisen in recent years about the concept of relative language proficiency in limited English

proficient students. Burt and Dulay (1980), for example, have argued that there are large numbers of LEP students who are "English Superior" and that their curriculum should be dictated by this classification. Indeed, the proposed LAU regulations of 1980 which were eventually discarded would have reflected this claim and "English Superior" students would have been provided substantially different services than would have "L_1 Superior" or "Equally Limited." It is important to realize that we currently do not have a testing technology that is adequate for the purpose of establishing a student's relative language proficiency. In addition, the research conducted by Dulay and Burt (1980) on this issue has several very serious methodological flaws which have been discussed at some length by De Avila and Ulibarri (1981). Merino and Spencer (1980) have examined the comparability of five oral language proficiency tests which have both an English and Spanish version from both a linguistic and psychometric point of view. They concluded that none of the instruments were sufficiently comparable across languages to warrant claims of relative language proficiency.

3. Linguistic Parameters Assessed: The methods used by a test to assess language differ partly as a function of the component(s) of language that the test author has chosen as either representative of the linguistic system or as being of special interest. Dieterich and Freeman (1979) have described these aspects of language clearly:

"Every spoken language has a universal framework of properties, or components: pronunciation, grammar, vocabulary, forms of discourse, and rules for use. Together, these components comprise the linguistic system of a language, and the different aspects of the system form a hierarchy of levels so that units at each level are organized into larger units at the next level.

Thus, the sounds of English are organized into words, words combine into grammatical structures which are organized to express various meanings, and utterances are used systematically in social situations.

In general, when language is learned in natural social situations (i.e., in the absence of formal instruction), the whole of a language system is acquired in meaningful "chunks" as communicative situations require or permit. In acquiring overall control of a linguistic system, the learner is gaining control of various subsystems which can be analytically separated: pronunciation (phonology); the grammar proper (syntax); the vocabulary (lexicon); patterns of discourse beyond the sentence; meanings associated with the grammar, vocabulary, and patterns of discourse beyond the sentence; and the rules of use (pragmatics)."

The choice of which parameters shall be represented in a test of language proficiency is an important one. The authors of one of the most commonly used tests of oral language proficiency, the Bilingual Syntax Measure, have argued convincingly that syntax is the parameter that is the least likely to be affected by extraneous or irrelevant distortion and bias. A close examination of attempts to measure other parameters provides many examples of where such measures can go wrong. The use of phonology as a major ingredient in decisions about language proficiency is a good example of where problems can arise. Whether a speaker of English pronounces words precisely as a native English speaker does is not a very important facet of that person's ability to <u>function</u> in an English-speaking situation. Henry Kissinger and Zsa Zsa Gabor are examples of people who function very competently in an English-speaking society but who have a heavily accented mode of speaking English. Regional differences alone could cause points to be lost on many tests of phonology. Although few tests provide measures of semantics or pragmatics, tests of vocabulary are frequently found. Although vocabulary is

obviously an important ingredient, one cannot be proficient in a language by being proficient in vocabulary alone. Thus, any test which rests a substantial portion of its score on vocabulary is suspect. How was the vocabulary chosen? What made these words rather than others representative of the vocabulary with which students of this age should be familiar? Is it clear from the test instructions and discussion that a vocabulary score is insufficient as an indicator of overall language proficiency and should not be interpreted as such?

4. Domains of Language Assessed: Language must be appropriate to the particular linguistic and social context in which it is used. At least three common domains of language exist: home, school, and neighborhood. Even a cursory examination of tests of language proficiency will reveal that the test author has chosen one or more of these domains as the context for the language use being assessed. The objects, situations, people, and ideas found in the home are obviously different than those found at school, or those found in the neighborhood or community at large. Thus, the vocabulary appropriate to conversation and linguistic exchange varies considerably from one context to another. For example, a discussion of historical periods, of triangles and other geometric shapes, or of scientific terms and procedures are likely to be found in the school, and less likely to be a frequent part of home discourse. The basic syntactical features will probably vary little across domains, but even pronunciation is likely to change from home to school. The formality of speech appropriate to a school setting may appear inappropriately condescending or disrespectful at home.

In theory, at least, it should be possible to assess language proficiency in each of these key domains. Unfortunately, little has been done to date to gauge the extent to which any particular test is tapping any particular domain.

5. Different Language Assessment Techniques: Different tests employ different techniques and tasks to assess a student's control over various aspects of language. Listed below is a supplemented version of the list of techniques and tasks given by Dieterich and Freeman (1979).

- o Answering questions about pictures, about a discourse, or general questions;
- o Describing, or telling a story about, pictures, objects, places, or people;
- o Paraphrasing something which is said;
- o Grammatically mainipulating sentences--changing tense or number, conjugating verbs, changing sentence form;
- o Completing cloze passages or sentences;
- o Repeating words, sentences, or stories;
- o Recalling words from lists of words, generally presented in two languages;
- o Discriminating between words which are phonologically similar;
- o Pointing to or marking pictures, words, sentences, or objects which correspond in some specified way to an oral cue;
- o Naming objects in pictures or in the physical environment;
- o Performing commands;
- o Selecting from several written sentences one which corresponds to an orally given sentence cue;
- o Selecting a grammatically correct written passage;
- o Finishing written passages with grammatically correct language;
- o Choosing the correct answer to a question reflecting the content in a written passage;
- o writing a passage or story based on pictorial or auditory input.

A distinction is frequently made between two general language assessment techniques: 1) Discrete Point Testing (or, linguistic manipulation; and 2) Integrative Testing (or, natural communication).

Discrete Point Testing. The term discrete point refers to the testing technique in which discrete (specific) points in the language system are tested. Each item of the test is designed to test a specific structure or rule. Usually, each item tests a particular point, independent of other points. All items combined would then sample a particular set of language points, or at least the points within a particular linguistic parameter--such as grammar, vocabulary, phonology.

One of the advantages of discrete point testing is that it permits systematic control over the extent to which specific facets of language are tapped by the test. Its major drawback, however, is that there may be a lack of correspondence between the extent to which students can perform to highly structured formal context-free tasks such as those usually used in discrete point testing, and the extent to which they use language effectively in natural interaction. The danger inherent in some discrete point testing is that the test results will merely show how familiar a student is with a roster of precise rules of grammar, without providing evidence of whether the student can use them in meaningful conversation or of the stage of language acquisition at which the student is presently functioning. Dieterich and Freeman (1979, page 28) provide a quick review of the pitfalls and potentials of several discrete point tests.

Integrative Testing. A test using the integrative technique will have the student produce connected discourse in a mean-

ingful context. The purpose of this task is to obtain evidence of a student's overall control of the language in a natural situation. The major drawback is that information will be obtained only on those language features that happen to occur in the language sample obtained. Some features (e.g., perfect tenses) are very difficult to elicit through natural communication or integrative techniques. In reviewing instruments which use this technique, Dieterich and Freeman (1979, page 28) found that none yielded any useful information regarding the degree of control of grammatical structure or developmental level of English acquisition. Even more serious problems were found in the scoring systems of these measures: "The types of evaluations which are made of elicited discourse are either so gross as to be unrevealing, so subjective as to limit their value, or based on misguided notions about language." Improvements that are badly needed in tests using integrative techniques include: 1) a scoring system that provides diagnostic information about students' control of the language structure; 2) indication of the developmental level of language acquisition; and 3) since nonstructured communication techniques could require that very extensive language samples would have to be collected before a sufficient range of structures were recorded, it would be ideal if some approach which reduced the necessary size of the language sample could be devised.

6. Assessing BICS and CALPS: James Cummins (1981) has articulated a model of language development that distinguishes between Basic Interpersonal Skills (BICS) and those required for the development of literacy and other Cognitive Academic Language Proficiency Skills (CALPS). Cummins' work is discussed in greater detail in Series B, Packet III. However, the concepts discussed by Cummins have an important bearing on our under-

standing of what tests tell us about language proficiency. For example, according to Cummins, CALP in L_1 and L_2 are interdependent. Thus, if it is possible to measure L_1 CALP, it might be possible to predict when a student could best benefit from reading and writing instruction in L_2. Many of the tests of oral language proficiency focus on aspects of language that Cummins would view as BICS (e.g., accent, syntax, vocabulary). Cummins has not yet fully operationalized his concept of CALPS in ways that would permit the development of instruments that measure its various components. Some researchers (e.g., Hernandez-Chavez and Merino, 1980) have ventured that factors such as graphic sense and other literacy readiness skills may make up some of the components of CALP. Surely standardized achievement tests are measuring part of the CALP concept, but one would expect that the domain could be much more clearly defined than depending upon those measures alone. Interest in CALP is bound to have an impact on test development in the years to come. It will be interesting to see whether these developments mimic the past mistakes of the developers of IQ, aptitude, and standardized achievement tests.

PART III: DEMISTIFYING THE PSYCHOMETRIC QUALITIES OF TESTS

The term psychometric qualities refers to properties that a test may or may not have which provide evidence that it is measuring what it purports to be measuring, and that it is doing so consistently. The two major types of psychometric quality are validity and reliability. In addition, many norm referenced tests and some domain referenced tests attempt to develop norms which provide a representative picture of how a specific population performs on the test. The psychometric quality of educational and psychological tests have been important to psychologists and educators for many years. The misuse of tests and the serious effects that such misuse can have was discussed earlier. In order to provide professional standards that would allow test developers to construct valid and reliable tests, a joint committee of the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education prepared and published a manual of test standards: Standards for Educational and Psychological Tests (1974). With this manual and a few other standard references (e.g., Anastasi,1976; Cronbach, 1970), any serious test developer has the blueprint for developing valid and reliable tests.

The purpose of this section is to review the various types of validity and reliability from the test user's point of view. This is an attempt to demistify concepts that are often regarded by program staff as highly technical and beyond their reach. In reality, the reasoning behind psychometric qualities rests very much on common sense. They need not be presented or viewed as difficult or technical material. Moreover, it is important for educators to understand what makes a test good and what makes a test poor, from a psychometric point of view. Program staff are often in the position of making a decision about which instrument to select and purchase. Once used, the staff again must cope with the interpretation of the test results. Because language assessment

is still an enterprise in its infancy, there are few instruments available that have established adequate psychometric qualities. Most have only brief research histories. We can look forward to many years of new tests, each in the process of establishing its validity and reliability. In addition, some tests have simply been very rapidly and very poorly constructed. In some cases, educational value has been sacrificed on the alter of profit. Tests are expensive for a program to purchase, and their prices will continue to increase in the years to come. Their expense rests not only in the purchase of test materials, but in the personnel and vendor costs involved in scoring, analyzing, and interpreting their results. For all of these reasons, Title VII program staff should be wise consumers of tests. They should know how to review them for key features, and they should demand quality. When quality is lacking, they should hold out until the test is improved.

Validity

Each test of language proficiency should have a technical report. Within this technical report, there should be a section which clearly and straightforwardly discusses the validity of the instrument. Not only should this section set forth the test developer's claims about validity, it should present evidence that proves those claims to be true. In short, saying it is so is not sufficient. There must be evidence.

There are basically three different types of validity:

1. Construct Validity
2. Content Validity
3. Criterion Validity
   a. Concurrent Validity
   b. Predictive Validity

Construct Validity. Perhaps the easiest way to understand what construct validity represents, is to consider how one would go about developing the ideal test of language proficiency. Ideally, the first step would be to carefully and systematically define the domain to be tested. In order to do this, one would need as complete a picture as possible of what both L_1 and L_2 consist (presuming two languages are of interest), and how they are developed and acquired respectively. Knowledge of the structure of the language, of all its components, and the rules that relate these components would be necessary. Knowledge of how these processes of development and acquisition vary as a function of a child's age, language background, and context would all be important. The existing research on the language to be assessed would provide a framework of understanding about how the language acquisition process works, and how it is used in various settings for various purposes.

With a framework, the test developer could then set priorities about which language domain would be represented by the items on the test. It is useful, then, to think of the research information and the framework of understanding built upon it as the constructs underlying the test. These are the test developer's view of the behaviors and skills to be tested. The question of whether a test has construct validity is asking whether the test developer's view is, in fact, accurate and sufficient in light of the best research evidence and existing theory. If so, is the structure and approach of the test construction effort reflective of these underlying constructs?

Cronbach (1970) provides the following definition of construct validity:

> *"Whenever a test talks about what the score means psychologically, or what causes a person to get a certain score, then concepts are involved and construct validity is appropriate. Construct validity is a broad area into which several statistical procedures for analyzing a test fall. It refers to the extent to which a test may be said to measure a theoretical concept or trait. Usually, construct validity is based both on the psychometric properties of the test and an analysis of discriminant and convergent validities (as opposed to simple correlation). It involves deriving hypotheses about test behavior based on the theory or construct and verifying them empirically."*

For a good example of how construct validity arguments have been developed, and of how a technical discussion of them has been presented, the manual of the Bilingual Syntax Measure should be reviewed.

Some of the questions to ask regarding a test's construct validity would include:

1. Have the underlying hypothetical constructs been identified?
2. How does the construct address first and second language development?
3. Has empirical evidence pertaining to the construct been cited?
4. Does the test agree with the theory or expectations of how the scores should behave according to the theory or other tests or criteria designed to measure the same construct?

Content Validity. Content validity is related to construct validity. For a test to have content validity, the test items or the testing techniques must provide an accurate and sufficient representation of the construct which the test purports to measure. It is not the same as face validity. In face validity, someone merely judges that the test items appear to be relevant to the construct. To have content validity a set of operations must have been carried out which - by virtue of their rationale, care, and thoroughness - give confidence that the test items

or the testing techniques have elicited behaviors which constitute a representative sample of the behaviors defined in the domain to be tested. Questions about content validity that would be appropriate to a language assessment instrument would include:

1. Is there a clear definition of the universe of language behavior? Is it adequate?
   a. Does the universe match classroom behavior/use of the language (including age variations)? Is it adequate?
   b. Is the method of sampling the universe specified? Is it adequate? (i.e., does it include dialect variance; origin of vocabulary usage, use of "experts" and their qualifications)
2. Have possible sources of bias been examined? (e.g., related to age, sex, ethnic, cultural, regional, sociolinguistic factors)
   a. Have items and procedures been screened for bias?
   b. Is there evidence on and an explanation of any group differences?

Program staff, as well as parent and community representatives may all play a role in examining a test for content validity. The pictures and language used in the test should be examined for cultural bias and age appropriateness. Most of the oral language proficiency tests used with older children have been criticized by adults and students alike for the childish nature of the picture stimuli and some content. Many of the English oral language proficiency tests were developed with a specific non-English language group in mind. For example, the BOLT has both a Spanish and English version. Perhaps this accounts for the fact that the picture stimuli seem much more appropriate for children of Hispanic background than any other ethnic background. In theory, the BOLT could be used with Asian children to assess their

command of English. However, the lack of relevance of the BOLT pictorial stimuli should argue against that use. Wherever picture stimuli are used, their appropriateness should be a matter of concern. The LAB, which was developed for New York, has pictures of steeples and of Grant's tomb which would not elicit appropriate responses from students who have not lived in the East. Some tests give unbalanced representation to girls and boys. Some provide culturally stereotyped portrayals of various ethnic groups. For example, one test which uses pictures to elicit storytelling behavior presents a variety of Asian children and adults in pictures, but the context is almost always rural, making it difficult for urban Asian students to identify with the pictures.

The use of language tests is far flung. They are marketed not only throughout the continental U.S., but in Hawaii and Micronesia as well. A teacher from Truk, an island in the Western Pacific and a Trust Territory of the U.S., recently reviewed the IDEA test of oral language proficiency. He commented on how unusual the drawing of a farmer would be to his students. It would be most unusual indeed for a farmer to be sitting on a tractor in clean clothes. More appropriate to their island context would have been a farmer with a machete, with a background of tropical vegetation. References to furnaces were puzzling, as are references to umbrellas in locations such as New Mexico. Teachers from another micronesian island objected to the use of timed tests in some of the commonly used instruments. Putting a premium on speed of response is incompatible with important values of conduct in many Pacific cultures. Mastery is the more important virtue. Gallimore, Whitehorn Boggs, and Jordan (1974) reviewed several experiments in which the test performance of a group of Hawaiian American students significantly improved in the number of correct responses and in the amount of sustained task-oriented behavior when the testing situations were organized to emphasize team work and rewards (e.g., Kubany, 1971; MacDonald & Gallimore, 1971:104). Unless the content and procedures of a language instrument have been specifically developed with the salient experience and performance factors of a culture in mind, the test may have very poor content validity for that particular culture. When there is any question, cross-cultural validity evidence should be provided by the test publisher.

Criterion Validity. The American Psychological Association's Standards for Educational and Psychological Tests (1974) defines criterion validity as follows:

> *"Criterion-related validities apply when one wishes to infer from a test score an individual's most probable standing on some other variable called a criterion. Statements of predictive validity indicate the extent to which an individual's future level on the criterion can be predicted from a knowledge of prior test performance; statements of concurrent validity indicate the extent to which the test may be used to estimate an individual's present standing on the criterion. The distinction is important. Predictive validity involves a time interval during which something may happen (e.g., people are trained, or gain experience, or are subjected to some treatment). Concurrent validity reflects only the status quo at a particular time."*

Thus, there are two kinds of criterion validity: concurrent and predictive. When a test developer reports that the test correlates with another test, it is concurrent validity that is usually being discussed. However, when the results of the test have been shown to be associated with some later event, such as graduation from highschool or success in an English-only classroom, the developer is referring to predictive validity. Unfortunately, none of the language proficiency tests currently available have been studied for their ability to predict successful transition out of Title VII programs and into English only classroom instruction.

The two most common approaches taken to establishing criterion validity seem to be the correlation of the test score with the scores of other language proficiency tests, and the correlation of the test scores with teacher ratings. If the first of these approaches is taken, a test should be selected which already has an established research history on its own validity. It is meaningless to know that two tests of unknown validity and reliability correlate highly. Attempts to show that a test's results correlate with teacher judgments have

more often than not been marred by methodological errors. When considering this type of evidence for criterion validity, one should ask whether the teachers had a clear and uniform understanding of each of the ratings possible, if they each had the same type of information about the students being rated, if they were naive about the test scores to which their ratings were to be compared, or to associated curriculum levels if the test is linked to the curriculum. Some of the other questions one should ask when appraising a test's criterion validity include:

1. Is there a logical intrinsic relationship between the test and the criterion?
2. Is there documentation, such as appropriate kinds of correlation, on the relationship between the test and the criterion and the strength of this association?
3. Is the criterion fully described?
   a. Does the criterion itself meet standards of validity and reliability?
   b. Is the criterion unbiased?
   c. Is the criterion related to language proficiency?
   d. If expert judgement is used, is information given on the expert's background and the procedures used for conducting the evaluation?
   e. Is the criterion based on some measure of classroom competency or achievement?
4. Is the criterion score determined independently of other test scores to avoid contamination?
5. Are appropriate distinctions made between concurrent and predictive validities?
6. Is a detailed description provided of the sample used for comparison (e.g., size, age, sex, ethnicity, dialectic background) in the criterion validity study?

Reliability

When a test is reliable, there is evidence that its results are consistently obtained on different occasions, by different examiners and different scorers, and that it exhibits stability. Five different types of reliability are of interest to users of language proficiency tests:

1. Test-retest Reliability;
2. Inter-examiner Reliability;
3. Inter-scorer Reliability;
4. Alternate Form Reliability;
5. Internal Consistency.

Test-retest Reliability. A test should yield approximately the same results on two different occasions when the same students are tested, the interval between testing occasions is no longer than four to six weeks, and when there has been no focused intervening training that would bring the students' scores to a higher level. When this is the case, a test is said to have test-retest reliability. This can be ascertained by simply correlating the scores of a group of students that are obtained on two separate, but proximate occasions. This is one of the most important of the psychometric qualities for tests of language proficiency because the results of the test are so frequently used to place students in educational programs and to diagnose their strengths and weaknesses. Unfortunately, many currently available tests do not have evidence of test-retest reliability. Users should insist that this information be developed. It is a minimum assurance of the test's quality and it is fast and inexpensive for a developer to provide.

Questions one should consider when examining the test-retest reliability of a test include:

1. Is a description provided of the sample, conditions, and interval used during the reliability study?

2. Is information provided on the nature of educational activities provided to the students taking the test during the intervening period?

Inter-Examiner Reliability. Different test administrators should be able to obtain the same results when they give the test to the same group of students. If they cannot, the quality of the test and its usefulness is certainly in question. Underlying conditions that contribute to inter-examiner reliability include thorough and understandable instructions for administering the test. The examiners should have instructions for proceeding under certain problem situations. If the expected examiner behavior is made clear, it is more likely that different examiners will conduct the testing in the same way. Also beneficial is a formal examiner training procedure. Many of the most commonly used tests of oral language proficiency train examiners at little or no cost. Evidence of inter-examiner reliability will consist of a correlation of the test results of a group of students when tested by one examiner and also by another examiner. The scoring of the results should be done by a single individual or by random assignment if more than one scorer is used in order to eliminate any possible effects of different scorers.

Inter-Scorer Reliability. Different scorers should be able to obtain the same results when they score the same set of tests. This is also an extremely important type of reliability for language assessment instruments because they typically require some degree of subjectivity in the scoring procedures. In order to achieve a high degree of consistency across scorers, the directions for scoring must be very clear. It must provide examples of how to deal with ambiguous situations. Training for scorers is advised in order to identify dissimilarities in their approaches and to train them to use uniform methods. Evidence of inter-scorer reliability will consist of a correlation of the test results of a group of students when scored separately by two or more scorers.

Alternate Form Reliability. Some tests provide more than one version of the basic test form. The IDEA is one example of a test that does this. Alternate forms are a good idea because if they are truly comparable forms the test can be given on repeated occasions with less concern about a practice effect occurring. They also provide more security in that students who take the test first will not be able to communicate important information about the test to a group who takes it later if the two groups receive different forms of the test. However, there must be a thorough rationale for constructing alternate forms and evidence that they are indeed parallel and comparable in effect. They must have identical structures and test precisely the same domain. Only the specific nature of the items should be changed. If alternate forms are offered by a test developer, the technical report on the test must discuss clearly and completely how the comparability of the forms was built in during the test construction phase. In addition, information should be presented on the means, variances, and characteristics of items in the forms, including coefficients of correlation among their scores.

Internal Consistency. If a test is internally consistent there is evidence that some portion of its items correlate with the total test score, or that several sets of items are intercorrelated. It is one of the easiest to obtain indicators of reliabiltiy and is therefore frequently reported. However, it is probably not as important to the test user as the preceding types of reliability. Moreover, evidence of internal consistency can never compensate for the lack of the other types of reliability. The rationale underlying the concept of internal consistency is that if properly selected, all components of a test will be tapping the same constructs and should therefore correlate with one another. Although this is generally true, one should always consider how

reasonable it is, from what is known about the empirical evidence and the theory about the construct, for one part of a test to correlate highly with another. For example, it would not be reasonable for skill in principles of accounting to correlate highly with skill in the principles of invertebrate behavior. Of course, this is an extreme example and no test would combine these subjects into related scales. Yet it does alert one to the possibility that subscales may not correlate because they are tapping quite different kinds of knowledge or skills.

Norms

The term norm refers to the process of norming a test and to the use of the results - norms - once the process has been carried through. In the process of standardizing a test, it is administered to a large group of students which has been carefully sampled so as to be a representative sample of some greater population, such as a sample of all third graders in the United States. The norms obtained in this way will indicate the average performance of students on the test, as well as the relative frequency with which students deviate above and below the average (Anastasi, 1976). Thus, norms "permit the designation of the individual's position with reference to the normative or standardization sample. Few tests of language proficiency have established norms for performance on the test. The obvious exception to this statement are the standardized tests of reading and language achievement such as the CTBS English and CTBS Español. A few other instruments present data in sections labeled "norms" but fail in all ways to actually develop a systematic body of normative data. It is important to understand when reviewing a test's discussion of norms that field test data is not the same as norms. To deserve the label, a sample must have been systematically selected that represents students on a number of key dimensions such as age, race/ethnicity, language background, and

regional location. Rather than lumping all scores together, the norming studies should provide separate results for different subsets of the sample. For example, it should be possible to compare the mean scores of students from different regions, or from different language backgrounds, at different grades, with different racial/ethnic identities, and of boys separately from girls. Among some of the questions one should ask when considering the norms of a test are the following:

1. Are normative data provided for subgroups for which the test is to be used?
2. If not provided, is justification given for why there is no need for differential norms?
3. Are normative data provided for the final version as opposed to the earlier field versions of the test?
4. Is information provided on the conditions of the norming studies and characteristics of the norming group, including age, sex, socioeconomic status, locale, ability, size, ethnic group, language background, rural/urban nature of the population, and grade level?
6. Is a discussion provided on any possible biasing factors or conditions.

EXHIBIT 1

EXAMPLES OF HOW STATES DIFFER IN THEIR APPROACHES TO DEFINING STUDENTS AS LIMITED ENGLISH PROFICIENT FOR PURPOSES OF ELIGIBILITY FOR BILINGUAL EDUCATION SERVICES

CALIFORNIA	Each student identified by a home survey as having a language background other than English is tested with a State designated instrument of oral language proficiency. The student is classified as limited or fluent English speaking based on the test's classification system. Limited English speaking students are considered LEP and are eligible for bilingual services. Students in grades 3 - 12 who are fluent in English oral proficiency, but score below the district established standards in either reading or writing, are also classified as LEP.
TEXAS	LEP is defined by a student's oral language proficiency performance, as gauged by the instrument's particular approach to the definition of LEP. The Texas State Education Agency convenes a group of experts to review and approve tests for use by local education agencies. In addition, for students in grades 2 - 12, those scoring below the 40th percentile on the language arts and reading standardized achievement test are also identified as LEP.
NEW YORK	LEP students are those who by reason of foreign birth or ancestry speak a language other than English or come from a home where a language other than English is spoken, and: 1) either understand or speak little or no English, or; 2) score below the 20th percentile on the Language Assessment Battery (LAB).
NEW MEXICO	Students are identified as LEP who have no English, are limited in English facility, and who are bilingual but below grade level. Many different tests are used to assess the language proficiency of students. Local education agencies make all decisions regarding the selection and use of instruments, with the State education agency playing no test review role. Bilingual programs are offered only at the K-6 grade levels.
FLORIDA	LEP students have a limited ability or no ability to understand, speak, or read English and have a primary or home language other than English. In general, these students are monolingual speakers of a language other than English (speak no English) or bilingual speakers primarily proficient in a language other than English. Testing and identification decisions are left entirely to the local educational level.

Source: Telephone Survey, June - August, 1981.

EXHIBIT I (CONTINUED)

COLORADO	Previously, LEP was defined as any child in an eligible district who has a language background other than English, and who performs below the district mean at grade level on a standardized test in reading and language arts -- or in the absence of a test, is judged by a teacher to read below grade level. New legislation defines LEP to be "linguistically different" students who: a) are monolingual in a language other than English; b) are dominant in a language other than English; or c) are bilingual but whose dominance is difficult to determine.

EXHIBIT II

EXAMPLES OF HOW STATES DIFFER IN THEIR APPROACHES FOR DETERMINING WHEN STUDENTS SHOULD BE RECLASSIFIED

TEXAS	Students scoring at or above 40% of the national norms on the reading and language arts scales of a standardized achievement test, and who have parent permission, may be reclassified. Otherwise, LEP students receive mandated bilingual education until the third grade. All LEP students must be in an ESL program regardless of grade level.
NEW YORK	In New York City, in the recent past, students with scores above the 20th percentile on the LAB were considered eligible for reclassification. Upstate New York districts used procedures which varied from district to district.
NEW MEXICO	In transitional programs, students would be exited when they attain a criterion of proficiency set by the individual district. Most programs are maintenance. In these cases, students would be expected to remain in these programs for enrichment purposes.
COLORADO	Students who have received bilingual education service for two years would be reclassified.
CALIFORNIA	Recommended reclassification procedures would include the establishment of a reclassification team which would review multiple measures of each student's performance, and make transition decisions on the basis of the total array of information available. Such measures might include scores from oral language proficiency tests, achievement tests, criterion referenced measures of curriculum mastery, teacher observations and ratings, and parent observations.

Source: Telephone Survey, June - August, 1981.

EXHIBIT III

NECESSARY PROPERTIES FOR TESTS USED FOR VARYING FUNCTIONS AND PURPOSES

Function: Classification

Purpose:	Necessary Properties of Tests:
1. Classification of students as either eligible or ineligible for bilingual education services.	a. The test must measure the language attributes specified by the State's definition of limited English proficiency; b. There must be a clear method of interpreting the test score into classifications of eligible and ineligible; c. Since this purpose is concerned primarily with obtaining a census of all students eligible for service, the test need not be suited for making educational decisions about individual students. If it is not so suited, these test results should only be used for inferences about groups and as a coarse screening device - not for inferences about individuals.
2. Reclassification of students into English only programs.	a. The test must measure the language attributes specified by the State's definition of English language proficiency; b. There must be clear criteria for interpreting the test score as an indicator that the student is either ready or not ready for the transition out of bilingual education services; c. Since this purpose is concerned with the placement of individual students, the test used must be valid and reliable at the individual student level.

Function: Providing Information About Student and Program Status

Purpose:	
3. Student Diagnosis	a. The test must be valid and reliable at the individual student level; b. It should provide detailed rather than gross screening information about the student's relative strengths and weaknesses on one or more language parameter

EXHIBIT III (CONTINUED)

	(e.g., lexicon, syntax, phonology, semantics), in one or more of the four language systems (e.g., speaking, understanding, reading, writing);
	c. It would be useful if it provided empirically established prescriptive information on activities which could enhance proficiency in either the specific skill areas where weaknesses were identified, or through an integrative approach.
4. Program Evaluation	a. The test must provide a measure of student performance compatible with each major program objective for student performance;
	b. The test score should be sensitive to change that occurs as the result of learning that takes place within a single program year;
	c. Test scores should be amenable to meaningful pre- and post-program comparisons. Scores on a continuous quantitative scale are most appropriate for this. Test results in the form of a few ratings or levels are less appropriate;
	d. Since this purpose is concerned primarily with group progress, the test need not be suited for making educational decisions at the individual level.
5. Program Planning	a. The test must provide measures on the student skills in which program decisions hinge. For example, if the program suspects that there will be a need for the addition of another English reading classroom during the next year, it would want to have measures on the L_1 and L_2 oral proficiency as well as the L_1 reading proficiency of students at the end of the program year.
	b. Since this purpose is concerned primarily with group progress, the test need not be suited for making educational decisions at the individual level.

EXHIBIT IV

COMPARISON OF TWO MAJOR TYPES OF ORAL LANGUAGE ELICITATION TASKS: NATURAL COMMUNICATION AND LINGUISTIC MANIPULATION

ITEM	NATURAL COMMUNICATION	LINGUISTIC MANIPULATION
Definition	Taps student's unconscious use of grammatical rules to produce utterances in a conversation. Uses natural speech where student's focus is on communicating something.	Taps student's conscious application of linguistic rules to perform a noncommunicative task. Uses artificial "speech" where student's focus is on a given rule.
Some Types	Structured communication, nonstructured communication, and so on (See Table 3)	Imitation, translation, completion, transformation, substitution, and so on.
Advantages	The language sample obtained represents natural communication, the skill that is ultimately being assessed. The task is virtually free of confounding task biases.	Target structures seem to be readily obtained.
Disadvantages	Certain structures are extremely difficult to elicit naturally; e.g., perfect tenses (had seen).	Confounds conscious knowledge and use of grammar rules with ability to use the language for communication; results in qualitatively different language than communication tasks.

Source: H. Dulay, E. Hernandez-Chavez, and M. Burt (1978).

EXHIBIT V

COMPARISON OF STRUCTURED AND NONSTRUCTURED NATURAL COMMUNICATION TASKS

ITEM	STRUCTURED COMMUNICATION	NONSTRUCTURED COMMUNICATION
Definition	Natural conversation between student and examiner in which examiner asks student specific questions designed to elicit target structures naturally and systematically	Natural conversation between student and examiner or other person in which no intent exists to elicit specific structures.
Advantages	Target structures may be elicited selectively and quickly; more efficient than nonstructured communication	Structures that are difficult to elicit with specific questions may be offered by subjects spontaneously.
Disadvantages	Not all structures are easily elicited; e.g., yes-no questions	A great deal of speech must usually be collected before a sufficient range of structures is used by the student to permit assessment of linguistic proficiency. One cannot make any statements about the student's control over structures not offered during the collection periods (because one cannot be certain why a structure was not offered; i.e, whether the situations did not require it or whether the student did not know it).

Source: H. Dulay, E. Hernandez-Chavez, and M. Burt (1978).

EXHIBIT VI

RELATIVE PSYCHOMETRIC STATUS OF SEVERAL POPULAR
ORAL LANGUAGE PROFICIENCY TESTS - ENGLISH VERSIONS

ORAL LANGUAGE PROFICIENCY TESTS: ENGLISH / MAJOR REVIEW FACTORS	BINL	BOLT	BSM I	BSM II	IDEA	LAB	LAS I	LAS II
1. NES/LES/FES Classification	θ	θ	◒	θ	θ	0	◒	θ
2. Validity: - Criterion	0	θ	0	0	θ	0	◒	θ
- Content	0	θ	●	0	θ	θ	θ	θ
- Construct	0	θ	◒	0	0	0	θ	θ
3. Reliability: - Test-Retest	0	θ	θ	0	0	0	●	0
- Interscorer	0	θ	◒	0	0	0	θ	0
- Internal Consistency	0	θ	0	0	θ	θ	●	0
- Alternate Form	0	0	0	0	θ	0	0	0
4. Norms	0	0	0	0	0	θ	0	0

0 No Information Given
θ Information Given, But Not Adequate or Appropriate
◒ Information Satisfactory, But Not Thorough
● Information Thorough and Adequate

50

Source: Merino & Spencer, 1980.

PART IV: REVIEW OF SELECTED POPULAR LANGUAGE ASSESSMENT INSTRUMENTS

Several efforts have been made to review the linguistic and psychometric qualities of language assessment instruments. The bibliography provides references to most of these. In addition, Appendix B contains a format developed by the California State Department of Education's Instrument Review Committee for the evaluation of oral language proficiency instruments.

The purpose of this section is to present brief reviews of some of the most commonly used language proficiency tests. These reviews are based primarily on those conducted for the California State Department of Education by Spencer (1978) and the Instrument Review Committee (1980, 1981). In referring to these and other test reviews, it should be remembered that new information and research is constantly being assembled on these and other instruments. When these instrument reviews are used, efforts should be made to obtain the most current information from the test authors, ERIC, and State and Federal agencies.

BILINGUAL SYNTAX MEASURE I (BSM I)

Psychological Corporation
1001 Polk Street
San Francisco, California 94109
415-771-3100

AGE/GRADE: Grades K - 2

MEASUREMENT FOCUS: The BSM I focuses exclusively on English and Spanish syntax. It measures structural proficiency in English and Spanish.

PURPOSE: The BSM I is intended to assess the English and Spanish oral proficiency of school children and to determine English-Spanish language dominance.

ORGANIZATION: A Spanish and English set of administration procedures are both used with a single set of stimulus pictures. These same materials are used with children of all ages within the K-2 range. There are 22 items which test specific language features in a hierarchical manner.

ADMINISTRATION: The BSM I is individually administered, requiring approximately 10-15 minutes per child. Anyone who can speak either English or Spanish can administer the BSM I in that language.

SCORING: The BSM I yields five scores: Level 1 (no English/Spanish), Level 2 (receptive English/ Spanish only), Level 3 (survival), Level 4 (intermediate), Level 5 (proficient).

The BSM is based on a theory of language acquisition which posits that specific language features are acquired in a hierarchical manner with certain language structures occurring before others. On this basis, the BSM was constructed so that two hierarchically ordered subsets of items are included in the test. The scoring and ultimate classification procedure are based on these ordered subsets. Levels I and II correspond to Non-English/Spanish Speaking, Levels II and IV to Limited English/ Spanish Speaking, and Level V to Fluent English/ Spanish Speaking.

BILINGUAL SYNTAX MEASURE I (BSM I)

NORM/DOMAIN REFERENCE:

The BSM I is a domain referenced test of syntactic proficiency in Spanish and/or English.

VALIDITY/RELIABILITY:

Although the validity of BSM I levels has been criticized in the literature, a careful foundation has been layed for construct validity. A study of the raw scores obtained by 1st and 3rd grade language minority California students on the BSM I and the LAS I showed a high degree of correspondence between the two measures. The correlation of language proficiency categories showed less correspondence.

In a test-retest comparison of 147 pupils, 5 changed on the BSM-S. Using the Kappa Coefficient, the two tests were shown to be about equally reliable at levels above chance. Point-biserial correlations of each item with the Level 1 or 2 scale score resulted in coefficients of less than .50 in all cases.

POTENTIAL PROBLEMS/ LIMITATIONS:

The rating scale of the BSM I has been criticized for having too few levels, for identifying children with some understanding of English at Level I (no English), by attempting to identify receptive skills at Level II via oral production, and for Level 5 (native or near native) because only 60% of Anglo-American students received that rating.

Of the seven stimulus pictures, only one contains a female and no representatives of Blacks or Asians are provided.

The small number of items is a problem because control of each language feature is measured by so few items, and because of the likelihood of practice effects when students are given the test several times.

The BSM I is a highly focused test and is applicable only to the early elementary grades. Thus, a comprehensive testing program would have to select other tests appropriate to all ages and to other communication content areas.

BILINGUAL SYNTAX MEASURE II (BSM II)

Psychological Corporation
1001 Polk Street
San Francisco, California 94109
415-771-3100

AGE/GRADE	Grades 3 - 12
MEASUREMENT FOCUS:	The BSM II focuses exclusively upon English and Spanish syntax. It measures structural proficiency in English and Spanish, and the degree of maintenance or loss of basic Spanish structures.
PURPOSE:	The BSM II is intended to assess the English and Spanish oral proficiency of school children.
ORGANIZATION:	A Spanish and an English set of administration procedures are both used with a single set of stimulus pictures. These same materials are used with children of all ages within the 3-12 grade range. There are 26 items which test specific language features in a hierarchical manner.
ADMINISTRATION:	The BSM II is individually administered, requiring approximately 10-15 minutes per child. Anyone who can speak either English or Spanish can administer the BSM II in that language.
SCORING:	The BSM II yields six scores: Level 1 (no English/Spanish), Level 2 (receptive English/, Spanish only), Level 3 (survival), Level 4 (intermediate), Level 5 (proficient), and Level 6 (proficient II). The BSM is based on a theory of language acquisition which posits that specific language features are acquired in a hierarchical manner with certain language structures occurring before others. On this basis, the BSM was constructed s- that two hierarchically ordered subsets of items are included in the test. The scoring and ultimate classification procedure are based on these ordered subsets. Levels I and II correspond to Non-English/ Spanish Speaking, Levels III and IV to Limited English/Spanish Speaking, and Levels V and VI to Fluent English/Spanish Speaking.

54

BILINGUAL SYNTAX MEASURE II (BSM II)

NORM/DOMAIN REFERENCE:	The BSM is a domain referenced test of syntactic proficiency in Spanish and/or English.
VALIDITY/RELIABILITY"	The construct of language acquisition/ creative construction is addressed in general terms but is not specifically defined as it applies to the age range for which the BSM II is designed. Rather, the rationale used for the BSM I seems to have been adopted with little consideration for the developmental differences involved with this older group of students. No evidence is provided to show that the syntactic features selected are needed for purposes of engaging in classroom discourse between third grade and high school. There is some evidence that the BSM II is related to other measures of oral language proficiency, but not to standardized tests of achievement. None of the information provided on reliabiltiy is considered to be both satisfactory and thorough. Methodological problems in the test-retest study included the misclassification of a good number of children, the failure to include students from the entire grade range, and the absence of students across the entire continuum, particularly at the lower levels
POTENTIAL PROBLEMS/ LIMITATIONS:	There are no alternate forms. Given the small number of items, the possibility of learning the test is strong. Because of variability of responses in early acquisition of L_2, a small number of items may mask proficiency. The empirical justification of the NES/LES (LEP) and FES (FEP) classification is inadequate. The story line used in the BSM II is sometimes hard to follow for a beginner. The implications of using an extended narrative for the purpose of testing language proficiency are not discussed in the rationale. For example, memory would be a factor.

55

LANGUAGE ASSESSMENT SCALES I (LAS I)

DeAvila, Duncan & Associates
P. O. Box # 770
Larkspur, California 94939

AGE/GRADE: K - 6 grades

MEASUREMENT FOCUS: The LAS, a test of oral language, was designed to simultaneously measure four psycholinguistic subsystems of the English and Spanish language: the phonemic system (basic "sounds" of the languages); the referential system (the words of the language); the syntactical system (the rules for making meaningful sentences); and the pragmatic system (using the language to reach specific goals).

PURPOSE: The LAS I was developed for the purpose of providing an overall picture of a student's linguistic (oral language) ability by separately assessing the component parts of the language system. It was also designed to permit diagnostic interpretations of the linguistic problems of each student.

ORGANIZATION: The LAS I is organized into six parts as follows: Part I - 36 phoneme production items; Part II - 36 items measuring ability to distinguish minimal sound pairs; Part III - 10 items measuring lexical production; Part VI - 10 items measuring aural syntax comprehension; Part V - storytelling section measuring oral syntax production; and Part VI - 7 items assessing ability to use the language for pragmatic ends.

ADMINISTRATION: The LAS is individually administered and requires approximately 15 to 20 minutes per child. It can be administered by any school personnel who are fluent in the language in which the test is being administered. A tape recorder is required. All instructions are to be translated into the student's first language when necessary, in order to ensure that the student understand the instructions. Actual test items are given in the language being assessed. Emphasis is placed on having the child understand the task.

LANGUAGE ASSESSMENT SCALES I (LAS I)

SCORING:

The LAS I produces measures in English and Spanish for 1) phoneme production; 2) ability to distinguish minimal sound pairs; 3) oral lexical production; 4) aural syntax (sentence) comprehension; 5) oral syntax production; and 6) ability to use language for pragmatic ends.

A total composite score is available, as well as specific identification of the linguistic problems of each student and a comparison of linguistic development among students at similar grade/age levels.

Guidelines are also provided for interpreting scores in terms of an individual student's language proficiency (e.g., Non-English/speaking, Limited English/Spanish speaking, Near Fluent English/Spanish speaking, Totally Fluent English/Spanish).

NORM/DOMAIN REFERENCE:

Although reference to normative data is made in the technical notes, the LAS I is basically a domain referenced test. There is however, a substantial research history on the use of the LAS with a number of language groups.

VALIDITY/RELIABILITY:

The LAS has been criticized for insufficient and after-the-fact development of a construct validity rationale. However, authors have recently articulated the underlying constructs as: 1) that oral proficiency is related to school achievement, and 2) that the LAS can discriminate levels of oral proficiency. Evidence supports these positions. Critiques of content validity have focused on the lack of rationale for selecting the language parameters that are included, and particularly the use of minimum pair items which seem to be testing dialect variation in some cases (e.g., whether-weather). These items seem to have been selected because they would present difficulty for a second language learner. Items were not appropriate to Spanish on this basis. Studies have shown the LAS to correlate moderately to the BSM and the BINL at some grade levels.

NES/LES/FES classifications on LAS I were stable over a test-retest interval of two to three weeks. Correlations for subscale scores were also high for both English and Spanish raw scores over the

LANGUAGE ASSESSMENT SCALE I (LAS I)

VALIDITY/RELIABILITY: (cont.)

same interval. Interrater reliabilities were computed and the resulting correlations were high. One recurring criticism of the LAS is the method of scoring the story retelling task. The examiner is instructed to use the sample protocols as guides to sorting the respondent's story into the appropriate proficiency level. However, the ambiguity of these samples and the lack of further guidance or training leads to confusion. Since this task accounts for a larger portion of the total score, scoring problems could have substantial effects on test results, and hence, on student placement or diagnosis decisions. However, a study of inter-scorer reliability yielded high correlations.

The LAS has no alternate forms. A study of internal consistency resulted in high correlation coefficients for all applicable subscales.

LANGUAGE ASSESSMENT SCALES II (LAS II)

DeAvila, Duncan & Associates
P. O. Box 770
Larkspur, California 94939

AGE/GRADE: 6 - 12 grades

MEASUREMENT FOCUS: The LAS, a test of oral language, was designed to simultaneously measure four psycholinguistic subsystems of the English and Spanish language: the phonemic system (basic "sounds" of the language); the referential system (the words of the language); the syntactical system (the rules for making meaningful sentences); and the pragmatic system (using the language to reach specific goals).

PURPOSE: The LAS II was developed for the purpose of providing an overall picture of the older student's linguistic (oral language) ability by separately assessing the component parts of the language system. It was also designed to permit diagnostic interpretations of the linguistic problems of each student.

ORGANIZATION: The LAS II is organized into six parts as follows: Part I - 24 items measuring ability to distinguish minimal sound pairs; Part II - 19 items measuring lexical production; Part III - 33 phoneme production items; Part IV - Sentence comprehension section measuring aural syntax comprehension; Part V story-telling section measuring oral syntax production; Part VI - an optional section measuring written production.

ADMINISTRATION: The LAS is individually administered and requires approximately 15 to 20 minutes per child. It can be administered by any school personnel who are fluent in the language in which the test is being administered. A tape recorder is required. All instructions are to be translated into the student's first language when necessary in order to ensure that the student understands the instructions. Actual test items are given in the language being assessed. Emphasis is placed on having the child understand the task.

LANGUAGE ASSESSMENT SCALES II (LAS II)

SCORING:

The LAS II produces measures in English and Spanish for 1) phoneme production; 2) ability to distinguish minimal sound pairs; 3) oral lexical production; 4) aural syntax (sentence) comprehension; 5) oral syntax production; and 6) written production.

A total composite score is available, as well as specific identification of the linguistic problems of each student and a comparison of linguistic development among students at similar grade/age levels.

Guidelines are also provided for interpreting scores in terms of an individual student's language proficiency (e.g., Non-English/speaking, Limited English/Spanish speaking, Near Fluent English/Spanish speaking, Totally Fluent English/Spanish.

NORM/DOMAIN REFERENCE:

Although reference to normative data is made in the technical notes, the LAS is basically a domain-referenced test.

VALIDITY/RELIABILITY:

The empirical evidence for the validity of the LAS II is considered to be uneven, according to critiques by Review Committees in California. The description of validity studies lack important detail in some cases. Content validity has been questioned due to the lack of a rationale for choosing the particular types of tasks and items that are represented in the test. The construct validity of the story retelling task has been cited several times as a source of concern, partly because a large proportion of the total score is based on this subscale and because its relevance to the construct of oral language proficiency has never been fully addressed by the authors. Criterion validity has been based primarily on correlation of the LAS with achievement measures and is considered to be fairly good. Studies have also shown a moderate relationship between the LAS II and the BSM and BINL, depending upon grade of the student.

The reliability of the LAS II is considerably weaker than that of the LAS I. There is no evidence of test-retest data, although studies of this nature are currently being conducted. Internal consistency correlation coefficients were high for all subscales. Interrater reliabilities were also high. The ambiguity of scoring protocols for the story retelling task are reported by users in the field to be a problem, undoubtedly contributing something to instability in the scoring of this task.

BASIC INVENTORY OF NATURAL LANGUAGE (BINL)

CHECpoint Systems
1558 N. Waterman Avenue, Suite C
San Bernardino, California 92404
714-883-3093

AGE/GRADE:	K - 12 grades
MEASUREMENT FOCUS:	The BINL measures oral language proficiency in English and Spanish. Some work with Cantonese, French, German, Greek, Italian, Japanese, Korean, Portuguese, and Vietnamese is claimed by the author.
PURPOSE:	The BINL was designed to measure the oral language proficiency of students in grades k-12.
ORGANIZATION:	The BINL consists entirely of a procedure for acquiring a natural language sample. A set of 40 story starter posters are provided as stimuli for the language samples. No guidelines are provided for selecting one of these stimuli over another.
ADMINISTRATION:	Previously, the BINL was administered to small groups of students. It is now more commonly administered on an individual basis. Test time is not specified. Each student is required to make up and tell a short story about the story posters. The examiner records the student's story verbatim until 10 or more sentences of phrases are provided. A tape recorder is essential.
SCORING:	The BINL provides scores in Fluency, Level of Complexity, and Average Sentence Length. Test authors claim that these scores will show the dominant language of a bilingual student, the degree of communication in either or both languages of a particular student, a student's degree of fluency in either language, and a student's level of complexity for either language.

BASIC INVENTORY OF NATURAL LANGUAGE (BINL)

SCORING: (cont.)	Scoring may be done either by hand or tests may be sent to the publisher for scoring. There is a cost associated with the latter. Serious questions have been raised about the procedures used in this centralized scoring process. The exact nature of this process is unknown, and there is evidence that it has fluctuated without warning or without psychometric reason.
NORM/DOMAIN REFERENCE:	The BINL is a domain referenced instrument. No normative information is available.
VALIDITY/RELIABILITY LIMITATIONS:	Several reviews of the BINL have found it to be severely wanting in psychometric quality. The following weaknesses have been cited: Scoring protocols are based on a construct (sentence complexity defined as T units), which does not discriminate developmentally, as well as other types of analysis. The relationship of the BINL with school achievement is negligible and inconsistent, contrary to expectations. Administration is lengthy if recommended procedures are followed. The test is expensive especially if machine scored. Inter-examiner reliability is critical and has not been addressed. Assignment to NES/LES/FES (LEP/FEP) categories is not empirically justified. Machine scoring will prolong time elapsing in placement of students. A stable computer program for machine scoring is not available.

LANGUAGE ASSESSMENT BATTERY (LAB)

Houghton Mifflin
777 California Ave.
Palo Alto, California 94304
415-324-4777

AGE/GRADE: K - 12

MEASUREMENT FOCUS: Reading, writing, listening, comprehension, and speaking proficiency in English and Spanish.

PURPOSE: The LAB was developed by the New York City Board of Education to assess the abilities of hispanic children, with the aim of identifying those children who cannot participate effectively in English, and provide a comparable measure of their communication skills in Spanish.

ORGANIZATION: A separate set of tests (one test for each skill in English and Spanish) for each of three grade levels: K-2, 3-6, 7-12.

ADMINISTRATION: Administration for all forms may be conducted by a non-expert examiner. The test of English is administered entirely in English and the test of Spanish is administered entirely in Spanish. The K-2 form is administered individually and requires approximately 5 to 10 minutes per child (40 items). The 3-6 and 7-12 forms are group administered and require approximately 41 minutes (92 items).

SCORING: The LAB yields separate scores in Listening/Speaking, Reading, and Writing at Level I (K-2) and separate scores in Listening, Reading, Writing, and Speaking Levels II and III.

NORM/DOMAIN REFERENCE: The LAB was field tested in English with 12,532 New York students in 45 schools with a population of at least 75% mono-lingual English. The Spanish version was field tested on 6,721 New York students in schools having a Hispanic population of over 70% It does not qualify as a norm-referenced test; however, the domain is not clearly articulated.

LANGUAGE ASSESSMENT BATTERY (LAB)

VALIDITY/RELIABILITY:

Very little is known about the validity of the LAB. Test authors claim content validity based on test development work using explicit learning objectives and a panel of item writers who are experts in the four communication skills. However, there is no information on how these objectives were established. The universe appears to be grossly inadequate for the age groups at every level. For example, the speaking test at Level I tests for lexicon only, when a native speaker of the age of 5 could be expected to be proficient in many more features. A child of this age would be diagnosed as fully effective before s/he is in any way comparable to a native English speaker of the same age. Similar problems exist at the higher levels.

No information is available on criterion validity, test-relability, or inter-examiner or inter-scorer reliability. Split-half correlations of all forms (English and Spanish) exceeded .87. Other measures of internal consistency (KR_{20} and KR_{21}) exceeded .80 for all forms excent kindergarten (KR_{21} = .43; KR_{20} = .82) Standard Error of Measurement ranged from 1.80 to 3.48 across all levels in English and Spanish.

POTENTIAL PROBLEMS/ LIMITATIONS:

The lack of satisfactory validity and reliability information makes reliance on the LAB for important educational decisions prohibitive. Moreover, the regional biases present in the item and in the field test sample do not bode well for use of the instrument beyond New York.

BAHIA ORAL LANGUAGE TEST (BOLT)

P. O. Box 9337
North Berkeley, California 94709

AGE/GRADE: 7 - 12 grades

MEASUREMENT FOCUS: Oral language skills in English or Spanish. Scores are based on whether or not the student answers a simple question correctly in the target language.

PURPOSE: To objectively classify a student in any one of four language categories (or into a supplementary level), for the purpose of aiding the teacher in placing the student.

ORGANIZATION: BOLT-English and BOLT-Spanish each consist of two sections. Section I contains four simple questions to determine if the test should be continued. Section II consists of a series of questions about picture stimuli.

ADMINISTRATION: The BOLT is administered individually by a non-expert examiner who is proficient in the target language. Administration of each test takes about six minutes.

SCORING: The BOLT yields five language classification scores: Level I - Non-Eng/Sp Speaking, Level I-S - Supplementary to level I (receptive skills apparent), level II - Very Limited Eng/Sp Speaking, Level III - Limited Eng/Sp Speaking, Level IV - Eng/Sp Speaking.*

NORM/DOMAIN REFERENCE: The BOLT is a domain referenced test.

VALIDITY/RELIABILITY: The underlying hypothetical constructs are not clearly set forth. The BOLT's validity claim seems to be keyed to the BSM II. However, the validity and reliability of that instrument are not certain. The information provided on content validity is also inadequate. The definition of the universe did not cite empirical evidence on syntax acquisition in second lan-

* The term Eng/Sp is used to indicate fluency in English when the English form is administered, and in Spanish when the Spanish form is administered.

BAHIA ORAL LANGUAGE TEST (BOLT)

VALIDITY/RELIABILITY: (cont.)

guage learners. The assumption was made that second language learning is the same as first language learning. Items were generated from first language acquisition research only. Research indicates that the order of acquisition in first and second language learners of English is not always parallel, especially at the upper ages which are the subject of this test. The rationale and procedures for selecting items from the universe are not anchored to research, and appear to have been done in a very casual way.

POTENTIAL PROBLEMS/ LIMITATIONS:

The BOLT is a focused test which addresses only oral language. It requires comprehension of oral questions about picture stimuli and the production of appropriate oral answers. It does not address reading or writing and is not suited for children in the elementary grades. Risks are high in using the BOLT for important educational decisions until more satisfactory and thorough validity and reliability evidence is available.

INDIVIDUALIZED DEVELOPMENTAL ENGLISH ACTIVITIES (IDEA)

Ballard & Tighe, Inc.
Oral Language Programs
7814 S. California Ave.
Whittier, CA 90602
213-947-6746

AGE/GRADE: K - 8 grades

MEASUREMENT FOCUS: The IDEA measures English production and comprehension through a discrete point technique. It focuses on vocabulary, comprehension, syntax, and verbal expression.

PURPOSE: The IDEA language proficiency instrument was initially used to determine at which level of the IDEA Oral Language Management Program a student should be placed. It is now promoted for use in determining level of oral language proficiency for Title VII eligibility and other instructional placement purposes.

NORM/DOMAIN REFERENCE: The IDEA does not qualify as a norm referenced test, yet its domain has not been clearly articulated either.

VALIDITY/RELIABILITY: The underlying hypothetical construct is superficially addressed and includes statements that do not reflect empirical evidence on language acquisition. For example, the statement that much instruction in the classroom is required for second language acquisition, is not supported by the research evidence (e.g., Ervin Tripp, 1974; Fillmore, 1976). Some of the six constructs mentioned do not appear in the instrument; e.g., "language is used to communicate in social situations." Pragmatics do not appear in the instrument.

Though the universe defined for oral language proficiency includes syntax, lexicon, phonology, morphology, comprehension, and production, no rationale or empirical evidence is given in terms of developmental order. Some areas (e.g., oral expression) are addressed by one item only. The inclusion of items in specific levels appears unrelated to relative difficulty. In some instances levels do not coincide with sequential order of phonology as known. More complex syntax,

INDIVIDUALIZED DEVELOPMENTAL ENGLISH ACTIVITIES (IDEA)

VALIDITY/RELIABILITY: (cont.)

vocabulary, and pragmatics are not addressed. Methodological problems involved in the three studies in which criterion validity could be evaluated render the results of these studies uninterpretable.

No information is available on test-retest, inter-examiner, or inter-scorer reliability. Internal consistency correlations yielded high positive associations. Although the IDEA has alternate forms, there is no alternate form reliability evidence available, and inspection of the items on these forms raises questions of their comparability.

POTENTIAL PROBLEMS/ LIMITATIONS:

If the IDEA test is used where the IDEA KIT is being used as the ESL curriculum, the possibility of contamination is present, as students could be "trained" to the test. Considerable risk is involved in using the IDEA as the basis for important educational decisions due to the lack of satisfactory and thorough psychometric qualities.

COMPREHENSIVE TESTS OF BASIC SKILLS (CTBS) ESPAÑOL

CTB McGraw-Hill
Del Monte Research Park
Monterey, California 93940
408-649-8400

AGE/GRADE: K - 12.9 grades

MEASUREMENT FOCUS: Reading and mathematics. In reading, tests of word recognition, vocabulary, and comprehension are provided.

PURPOSE: The purpose of the CTBS Español is to provide a Spanish language adaptation of the CTBS/S, a comprehensive, norm-referenced test of reading and mathematics. It is a measure of group skill rather than individual mastery.

ORGANIZATION: CTBS Español is a Spanish-language adaption of the CTBS/S Reading and Mathematics Tests. CTBS/S consists of seven overlapping levels: Level A (K.0-1.3); Level B (K.6-1.9); Level C (1.6-2.9); Level 1 (2.5-4.9); Level 4 (8.5-12.9). CTBS Español was adapted for levels B through 3, with Level A omitted because it is an English reading readiness test and Level 4 omitted because Level 3 is appropriate to most advanced Spanish-speaking students. CTBS Español includes the following tests Word Recognition I (Level B); Word Recognition II (Level B); Reading Vocabulary (Levels C-3); Reading Comprehension (Level-B-3); Mathematics Computation (Level B-3); Mathematics Concepts and Applications (Levels B-3). The items in each of the skills areas measure the following five test objectives: 1) the ability to recognize or recall information; 2) the ability to translate or convert concepts from one kind of language to another (i.e., verbal or symbolic); 3) the ability to comprehend concepts and their interrelationships; 4) the ability to apply techniques, including performing fundamental operations; 5) the ability to extend interpretation beyond stated information. Level B excludes objectives 3 and 4.

COMPREHENSIVE TESTS OF BASIC SKILLS (CTBS) ESPAÑOL

ADMINISTRATION:

The CTBS/S and CTBS Español are group administered tests. Recommended group size is 15 for grade 1 and 35 for all other grades. Testing is spaced over two or three days. The examiner must be highly skilled in reading and speaking Spanish and English. Two to four proctors are required. Carefully controlled administration procedures are necessary. Total testing times range from 19 to 45 minutes, with actual student working times ranging from 14 to 40 minutes. Several of the tests are "timed" tests, meaning that a definite time limit is placed on students' performance.

SCORING:

The CTBS/S in English and the CTBS Español provide three reading achievement scores: vocabulary, comprehension, and total Reading. According to test authors, these scores are not designed as measures of individual mastery of content; but rather, as a measure of group skills.

NORM/DOMAIN REFERENCE:

CTBS Español is a norm referenced test. The Reading and Mathematics CTBS/S tests of Levels B through 3 were adapted to Spanish by the Norwalk-La Mirada Unified School District. Following two field tests with subsequent item refinement phases, a standardization edition was administered to 5,200 Spanish-speaking students in the United States. Students were given both CTBS/S and CTBS Español.

VALIDITY/RELIABILITY:

The publisher is currently engaged in research to statistically equate CTBS/S with CTBS Español. A technical report with equated norms tables was unavailable at the time of review, but it is expected in the fall of 1978.

POTENTIAL PROBLEMS/ LIMITATIONS:

The CTBS Español is limited to measuring reading achievement of groups. It does not address oral language domains or writing and its authors caution that it is not suitable as a measure of individual mastery.

APPENDIX A: EXAMPLES OF INDIRECT MEASURES OF LANGUAGE

Survey Questions and Criteria Used by Three State Departments of Education to Estimate the LMP Population in the US.

State	Survey Questions	Criteria
California	1. Which language did your son or daughter learn when he or she began to talk? _____ 2. What language does your son or daughter most frequently use at home? _____ 3. What language do you use most frequently to speak to your son or daughter? _____ 4. Name the language most often spoken by the adults at home: _____	
	The HLS is sent to the home of each new kindergarten pupil, and each newly enrolled pupil whose files do not contain evidence of his/her being surveyed in 1977-78, and any special education pupils who were not surveyed in 1977-78. Any students returning the HLS with a language other than English indicated will have their oral language efficiency assessed.	
Texas	1. Does your child hear a language other than English spoken at home? Yes ____ No ____ 2. If yes, what is the other language that your child hears? _____ 3. Does your child hear this language spoken? a. most of the time _____ b. some of the time _____ c. not very often _____ 4. When this language is spoken, does your child understand? a. most of what is said _____ b. some of what is said _____	Indication that a language other than English is heard or spoken by the student.

State	Survey Questions	Criteria
	c. very little of what is said _____ d. nothing of what is said _____ 5. Does your child speak this language? The child is tested if 1. is YES and 2. is named; if 5. is YES; if 5. is NO but Questions 3. and 4. are answered a. or b.	
Colorado (prior to repeal of state bilingual education legislation)	1. Describe the language spoken by your child: a. speaks only _____ b. speaks mostly _____ c. speaks some _____ and some _____ d. speaks mostly English e. speaks only English 2. Do you have the advantage of having a language other than English spoken in your home? (options similar to A-E) 3. Describe the language understood by the child a. understands only _____ b. understands mostly _____ c. understands mostly English and some of _____ d. understands only English 4. Does your child have an opportunity to play with children or others who speak languages other than English? Yes _____ No _____	Indication that any language other than English is spoken, read, or understood.

Survey Questions and Criteria Used in three National Studies to Estimate the US LMP Population

Study	Survey Questions	Criteria
AUI CESS and NCES/ORA Review of CESS	1. What language do the people in this household usually speak at home?* 2. Do the people in this household often speak any other language here at home? If yes, which language?* 3. If yes to 1 and children 5-18 live in the household, a series of household enumeration questions were given. 4. What language does (each individual in household) usually speak? . What other language does _____ speak? . Was _____ born in the U. S?? . Where was _____ born?	

*Respondent is asked to select answer from precoded alternatives.

IPS

1. If the usual household language is other than English, or
2. If the second household language is other than English, or
3. If the individual's usual language is other than English, or
4. If the individual's other spoken language is other than English, or
5. If the individual's mother tongue is other than English.

POST-TEST

Part I:

1. Name the major purposes of testing in a Title VII program.

2. Give three different definitions of language proficiency.

3. Explain what properties a test should have to measure each of the different definitions of language proficiency.

4. Explain what different properties a test should have in order to be appropriate to 1) classification of students for eligibility for Title VII services; 2) for program evaluation; 3) for individual student diagnosis

Part II:

5. What are the two basic ways of constructing tests? Describe each and contrast.

6. What is the difference between indirect and direct measures of language proficiency?

7. What are some common sources of bias and distortion in indirect measures?

8. What is the conclusion of research findings on the concept of relative language proficiency?

9. What are the major linguistic parameters? Name and describe. What are the pros and cons of using each to assess language?

10. What is the difference between discrete point and integrative testing techniques?

Part III:

11. What is validity? Define it in general and specify three types of validity.

12. What is reliability? Define it in general and specify five different types of reliability.

13. Define what norms are.

POST-TEST ANSWERS

Answers will be found on page 5.

BIBLIOGRAPHY

American Psychological Association. Standards for educational and psychological tests. Washington, D.C.: APA, 1974.

Anastasi, A. Psychological Testing, fourth edition. New York: Macmillan Publishing Co., 1976.

Bilingual Education Act. In Bilingual education series: 9, the current status of bilingual education legislation, an update. Arlington, VA: Center for Applied Linguistics, 1980.

Dulay, H. and Burt, M. The relative proficiency of limited English proficient students. In J.E. Alatis (Ed.), Georgetown University Round Table on Languages and Linguistics 1980: Current Issues in Bilingual Education. Washington, D.C.: Georgetown University Press.

California State Department of Education, Office of Program Evaluation and Research. Language Proficiency Instrument Review by the Language Proficiency Instrument Review Committee. Sacramento, CA: 1980, 1981.

Cohen, A. D., and Roll, C. L. Assessing bilingual speaking skills: in search of natural language. In A. D. Cohen, M. Bruck, and F. V. Rodriguez-Brown, Evaluating Evaluation, Bilingual Education Series: 6, Arlington, VA: Center for Applied Linguistics, 1979.

Cronbach, L. J. Essentials of psychological testing, third edition. New York: Harper & Row, 1970.

Cummins, J. The role of primary language development in promoting educational success for language minority students. Paper prepared for the California State Department of Education, Compendium on Bilingual-Bicultural Education. Sacramento, CA: 1981.

De Avila, E. and Ulibarri, D. Nabe News, Volume IV, Number 3, January, 1981.

Dieterich, T. G. and Freeman, C. Language in Education: Theory & Practice, 23, a linguistic guide to English proficiency testing in schools. Arlington, VA: Center for Applied Linguistics, 1979.

Gallimore, R., Whitehorn Boggs, J., and Jordan, C. Culture, Behavior, and Education, A Study of Hawaiian-Americans. Beverly Hills: Sage Publications, 1974.

Gillmore, G., and Dickerson, A. The relationship between instruments used for identifying children of limited Engish speaking ability in Texas. Houston, Texas: Region IV Education Service Center, 1750 Seamist, 77008.

Glaser, R. A criterion-referenced test. In Popham, W.J. (Ed.), Criterion-Referenced Measurement. Englewood Cliffs, N.J.: Educational Technology Publications, 1971.

Hively, W. Domain-referenced testing. Educational Technology, June 1974.

Kubany, E.S. The effects of incentives on the test performance of Hawaiians and Caucasians. Dissertation, University of Hawaii, Department of Psychology, Honolulu, 1971.

MacDonald, S. and Gallimore, R. Introducing experienced teachers to classroom management techniques. Journal of Educational Research, 1971, 65, 420-424.

Merino, B. and Spencer, M. The comparability of English and Spanish versions of oral language proficiency instruments. Paper presented to the Association of Mexican American Educators, State of California, San Diego, California, November 14, 1980.

National Education Association: Task Force and other reports. Presented to the fifty-second Representative Assembly of the National Educational Association, July 3-6, 1973. Washington, D.C.: National Education Association, 1973.

Oakland, T. (Ed.) Psychological and Educational Assessment of Minority Children. New York: Brunner/Mazel, 1977.

Rosansky, E. J. A review of the Bilingual Syntax Measure. In B. Spolsky (Ed.), Papers in Applied Linguistics, Advances in Language Testing, Series: 1. Arlington, VA: Center for Applied Linguistics, 1979.

Spencer, M. Selective review of Spanish/English Tests of achievement and oral language development. California State Department of Education, Office of Program Evaluation and Research, July, 1978.