EAC-West, New Mexico Highlands University, Albuquerque, December 1995

HANDBOOK OF ENGLISH LANGUAGE PROFICIENCY TESTS

Ann Del Vecchio, PhD
Michael Guerrero, PhD

Evaluation Assistance Center - Western Region
New Mexico Highlands University
Albuquerque, New Mexico

December, 1995

Table of Contents

Introduction

The purpose of this handbook is to provide educators who have the responsibility for assessing the English language proficiency of limited English proficient (LEP) students with information about commercially available, standardized English language proficiency tests. The majority of the information in this handbook concerns the description of 5 standardized English language proficiency tests. The handbook includes information to facilitate informed test adoption. No particular endorsement for the use of any of these tests is intended.

In addition to providing information about these tests, background information about language proficiency testing is included. Language proficiency testing is a complex undertaking that continues to stir much debate among language researchers and test developers. Major differences of opinion concern the exact nature of language proficiency and how to best assess it. More importantly, while this debate takes place, educators are pressed into choosing and administering language proficiency tests to make programmatic decisions about limited English proficient students. We hope that this background information will allow English language proficiency test users to better understand the strengths and weaknesses of the tests described in this handbook.

The Legal Mandate

Regardless of state laws, all states, local education agencies, and schools in the United States must have and must implement a legally acceptable means of identifying LEP students. This obligation is enforced by various federal laws upheld by the Office for Civil Rights. If a student comes from a home where a language other than English is used, the school must assess the student's oral language proficiency, including reading and writing skills, in the English language (Roos, 1995).

A number of states have explicit procedures schools must follow to identify potential LEP students. Generally these procedures entail the use of a Home Language Survey which consists of questions designed to determine whether or not the student comes from a non-English speaking background. If there is an indication that the student comes from such a background, then the student's English language proficiency must be measured. Some states also endorse the use of specific English language proficiency tests. It is the responsibility of the readers to become informed about their state's policy with regard to the assessment of English language proficiency. There is no doubt that with the passage of the Improving America's Schools Act (IASA), the need for educators to select and use English proficiency tests will increase. For example, the new Title I legislation clearly indicates that:

- LEP students are eligible for educational services on the same basis as other students served by Title I.
- English language proficiency results must be disaggregated within each State, local educational agency, and school.
- The English language assessments used must be valid and reliable and consistent with relevant, nationally recognized, professional and technical standards for such assessments.
- English language proficiency results must be reported annually.

Whenever a potential LEP student has been identified, the local education agency has the legal responsibility of assessing the student's English language proficiency. Other state and federally funded educational programs intended for LEP students set forth criteria which must also be considered and

followed with regard to the assessment of the learner's English language proficiency. In both cases, the educator responsible for assessing the learner's English language proficiency must adopt a test for this purpose.

Definitions of Language Proficiency

Before engaging in a discussion of what it means to be limited English proficient, it is first necessary to understand what language proficiency encompasses. Unfortunately, it is at this point in the assessment of language proficiency that a lack of consensus begins. Language researchers openly acknowledge this dilemma.

Cummins (1984), for example, states that the nature of language proficiency has been understood by some researchers as consisting of 64 separate language components and by others as consisting of only one global factor. Valdés and Figueroa (1994) indicate that:

> ...what it means to know a language goes beyond simplistic views of good pronunciation, "correct" grammar, and even mastery of rules of politeness. Knowing a language and knowing how to use a language involves a mastery and control of a large number of interdependent components and elements that interact with one another and that are affected by the nature of the situation in which communication takes place. (p. 34)

Oller and Damico (1991) succinctly state that the nature and specification of the elements of language proficiency have not been determined and there continues to be debate among academicians and practitioners about the definition.

The complexity of language and the lack of consensus as to the exact nature of language proficiency is critical for one fundamental reason. Each language proficiency test should be based on a defensible model or definition of language proficiency. The question then becomes, which definition? Language proficiency tests have been developed based on a plethora of definitions and theories. Additionally, the test developer may indicate that a test is based on a particular model of language proficiency but it remains to be seen just how successfully the model was actually operationalized in the form of a test. In other words, describing the theoretical model of language proficiency in a technical manual does not mean that the test exemplifies the model.

What does it mean to be limited English proficient? Not surprisingly, there is also no common operational definition used by all states to define what it means to be limited English proficient (Rivera, 1995). However, existing federal education legislation, state education staff, and academicians have set forth general and consistent parameters for defining limited English proficiency and fluent English proficiency.

Under Section 7501 of the Bilingual Education Act, reauthorized in 1994 under IASA, a limited English proficient (LEP) student is a student who:

- was not born in the United States or whose native language is a language other than English and comes from an environment where a language other than English is dominant; or
- is a Native American or Alaska Native or who is a native resident of the outlying areas and comes from an environment where a language other than English has had a significant impact on such an individual's level of English language proficiency; or
- is migratory and whose native language is other than English and comes from an environment where

a language other than English is dominant; and

- who has sufficient difficulty speaking, reading, writing, or understanding the English language and whose difficulties may deny such an individual the opportunity to learn successfully in classrooms where the language of instruction is English or to participate fully in our society.

This legal definition for LEP students has been used to determine the eligibility of students for bilingual education services funded through federal Title VII legislation. It also is used by districts without Title VII funding to design entry/exit criteria for English as a Second Language (ESL) programs or district-funded bilingual programs. It merits highlighting that any determination of limited English proficiency entails assessing language in each of the four modalities (i.e., speaking, listening, reading and writing).

The Council of Chief State School Officers (CCSSO) defines English language proficiency in this way:

> A fully English proficient student is able to use English to ask questions, to understand teachers, and reading materials, to test ideas, and to challenge what is being asked in the classroom. Four language skills contribute to proficiency as follows:
>
> 1. *Reading* - the ability to comprehend and interpret text at the age and grade-appropriate level.
>
> 2. *Listening* - the ability to understand the language of the teacher and instruction, comprehend and extract information, and follow the instructional discourse through which teachers provide information.
>
> 3. *Writing* - the ability to produce written text with content and format fulfilling classroom assignments at the age and grade-appropriate level.
>
> 4. *Speaking* - the ability to use oral language appropriately and effectively in learning activities (such as peer tutoring, collaborative learning activities, and question/answer sessions) within the classroom and in social interactions within the school. (1992, p. 7)

The CCSSO adds to the definition of English language proficiency by identifying limited English proficient students as:

> having a language background other than English, and his or her proficiency in English is such that the probability of the student's academic success in an English-only classroom is below that of an academically successful peer with an English background (1992, p. 7).

Captured again within the CCSSO definition of language proficiency is the necessity of language ability in the four modalities and the need to assess each of these four skills.

Canales (1994) offers an equally practical definition of English language proficiency. Her definition of language usage (proficiency) is predicated on a socio-theoretical foundation. What this means is that language is more than just the sum of discrete parts (e.g., pronunciation, vocabulary, grammar). It develops within a culture for the purpose of conveying the beliefs and customs of that culture. Anyone who has ever tried to translate an idiom from one language to another understands this premise. A "bump on a log" in English means someone who is lazy or a do-nothing, but the non-English speaker has to assimilate the idiom

rather than the meaning of each individual word in order to make sense of the phrase. Canales says that language usage is:

- dynamic and contextually-based (varies depending upon the situation, status of the speakers, and the topic);
- is discursive (requires connected speech); and
- requires the use of integrative skills to achieve communicative competence. (p. 60)

In other words, language proficiency is a coherent orchestration of discrete elements, such as vocabulary, discourse structure and gestures, to communicate meaning in a specific context (e.g., the school).

Consider the kinds of linguistic abilities that underlie the successful academic performance of students. Students must be able to orally respond to teacher and peer queries for information, ask probing questions, and synthesize reading material. They must be able to understand routine aural instructions in a large group setting and peer comments in a small group setting. In terms of reading skills, students are required to extract meaning from a variety of text types including trade books, textbooks from across the curriculum, reference books and environmental print. The continuum of writing skills students need in order to succeed academically is equally broad. For example, students must be able to write short answers, paragraphs, essays and term papers. Moreover, the successful language user also knows the social and cultural rules governing these and many other language mediated activities.

Each of these educationally driven conceptions/definitions of language proficiency share at least two critical features. First, each definition accommodates the four linguistic modalities: speaking, listening, reading and writing. Second, each definition places language proficiency within a specific context, in this case the educational setting. Consequently, an English language proficiency test should utilize testing procedures that replicate -- as nearly as possible -- the kinds of contextualized language processing that is used in mainstream English speaking classrooms.

Valdés and Figueroa (1994) maintain that language proficiency testing should require this kind of contextualized language processing. They take the position that it is feasible to:

> ...identify the levels of demand made by such contexts and the types of language ability typical of native, monolingual English speaking children who generally succeed in such contexts. From these observations, one could derive a set of criteria against which to measure the abilities of non-native English speaking children in order to decide whether to educate them in English or their home language. (p. 62)

The reason for this recommendation is obvious. An English language proficiency test score is intended to assist educators in making an accurate judgment regarding which students need English language instructional assistance or no longer need such assistance. Making such a judgment becomes difficult when the language tasks underlying the test score bear little resemblance to the language tasks characteristic of a mainstream classroom.

General Nature of Language Proficiency Tests

Oller and Damico (1991) indicate that language proficiency tests can be associated with three schools of thought. The first of these trends, the discrete point approach, was based on the assumption that language proficiency:

> ...consisted of separable components of phonology, morphology, lexicon, syntax, and so on, each of which could be further divided into distinct inventories of elements (e.g., sounds, classes of sounds or phonemes, syllables, morphemes, words, idioms, phrase structures, etc) (p. 82).

They describe language tests based on the discrete point approach in the following way:

> Following the discrete point model, a test could not be valid if it mixed several skills or domains of structure (Lado, 1961). By this model, presumably the ideal assessment would involve the evaluation of each of the domains of structure and each of the skills of interest. Then, all the results could be combined to form a total picture of language proficiency. (p. 82).

A discrete point language proficiency test typically uses testing formats such as phoneme discrimination tasks where the test taker is required to determine whether or not two words presented aurally are the same or different (e.g., /ten/ versus /den/). A similar example might be a test designed to measure vocabulary which requires the test taker to select the appropriate option from a set of fixed choices.

The authors conclude that the weaknesses leading to the demise of such thinking centered upon evidence such as:

- the difficulty of limiting language testing to a single skill (e.g., writing) without involving another (e.g., reading);
- the difficulty of limiting language testing to a single linguistic domain (e.g., vocabulary) without involving other domains (e.g., phonology); and
- the difficulty of measuring language in the absence of any social context or link to human experience.

According to Damico and Oller (1991), these limitations gave rise to a second trend in language testing, the integrative or holistic approach. This type of testing required that language proficiency be assessed "in a fairly rich context of discourse" (p. 83). This assumption was based on the belief that language processing or use entails the simultaneous engagement of more than one language component (e.g., vocabulary, grammar, gesture) and skill (e.g., listening, speaking). Following this logic, an integrative task might require the test-taker to listen to a story and then retell the story or to read the story and then write about the story.

The third language testing trend described by the Damico and Oller (1992) is referred to as pragmatic language testing. It differs from the integrative approach in one fundamental way, an ostensible effort is made to link the language testing situation with the test-taker's experience. As Oller and Damico (1991) state, normal language use is connected to people, places, events and relations that implicate the whole continuum of experience and is always constrained by time or temporal factors. Consequently, pragmatic language tasks are intended to be as "real life" or authentic as possible.

In contrast to an integrative task, a pragmatic approach to language testing might require the test-taker to engage in a listening task only under the contextual and temporal conditions that generally characterize this activity. For example, if the test-taker is going to listen to a story and then retell the story, the following conditions might apply. From a pragmatic perspective, language learners do not generally listen to audio-taped stories; they more commonly listen to adults or competent readers read stories. In this sense a story-retell listening task which uses a tape-mediated story falls short of meeting pragmatic criteria. A pragmatic approach to story retelling might take on the following features:

- normal visual input is provided (e.g., the reader's gestures, the print on the page, an authentic number of story linked pictures in the text);
- time is managed differently in that the learner may have opportunities to ask questions, make inferences, or react in a normal way towards the content of the story; and
- the story, its theme, the reader, and the purpose of the activity form part of the learner's experience.

Oller and Damico (1991) make an interesting observation regarding the power of pragmatic language testing. The researchers state:

> What was more important about pragmatic tests, and what is yet to be appreciated fully by theoreticians and practitioners (e.g., Spolsky, 1983), is that all of the goals of discrete point items (e.g., diagnosis, focus, isolation) are better achieved in the full rich context of one or more pragmatic tests... As a method of linguistic analysis, the discrete point approach had some validity, but as a practical method for assessing language abilities, it was misguided, counterproductive, and logically impossible. (p. 85)

In other words, if the intent is to measure the learner's proficiency in the areas of grammar, vocabulary and pronunciation, for example, this is best achieved through a pragmatic language approach as opposed to a discrete point approach.

Language proficiency testing approaches tend to fall into one or more of the three trends just described. It is, however, the pragmatic language approach which seems to meet the demands of educators as set forth by federal education mandates, state education staff and academicians previously described. Nonetheless, educators will be limited to the use of currently available tests which may or may not measure language proficiency in a pragmatic, "real life" manner.

Limitations of Existing English Language Proficiency Tests

In the Introduction, it was clearly stated that the intent of this handbook is not to critique the five English language proficiency tests described. The following information is intended to inform the test user (i.e., educators) about the limitations of these tests in general and to help explain these limitations.

It has already been stated that language proficiency tests need to based on a particular theory or model of language proficiency. However, it was also stated that there is no consensus among researchers regarding the nature of language proficiency. The result has been the development of language proficiency tests which differ in many fundamental ways from one another. More important is the fact that different language proficiency tests have been shown to generate different language classifications (e.g., non-English speaking, limited English speaking and fully English proficient) for the same students (Ulibarri, Spencer & Rivas, 1981). Valdés and Figueroa (1994) report:

> So great indeed were the discrepancies between the numbers of children included in NES and LES category by different tests that cynical consultants often jokingly recommended one "state approved" instrument or another to school districts depending on whether administrators wanted to "find" large or small numbers of LES children. (p. 64)

Unfortunately, it is not only the test qualities with which educators must be concerned.

Related to the design of language proficiency tests, there may be a propensity for test developers to use a

discrete point approach to language testing. Valdés and Figueroa (1994) state:

> As might be expected, instruments developed to assess the language proficiency of "bilingual" students borrowed directly from traditions of second and foreign language testing. Rather than integrative and pragmatic, these language assessments instruments tended to resemble discrete-point, paper- and-pencil tests administered orally. (p. 64)

Consequently, and to the degree that the above two points are accurate, currently available language proficiency tests not only yield questionable results about student's language abilities, but the results are also based on the most impoverished model of language testing.

In closing this section of the handbook, consider the advice of Spolsky (1984):

> Those involved with language tests, whether they are developing tests or using their results, have three responsibilities. The first is to avoid certainty: Anyone who claims to have a perfect test or to be prepared to make an important decision on the basis of a single test result is acting irresponsibly. The second is to avoid mysticism: Whenever we hide behind authority, technical jargon, statistics or cutely labelled new constructs, we are equally guilty. Thirdly, and this is fundamental, we must always make sure that tests, like dangerous drugs, are accurately labelled and used with considerable care. (p. 6)

In addition, bear in mind that the above advice applies to any testing situation (e.g., measuring intelligence, academic achievement, self-concept), not only language proficiency testing. Remember also that the use of standardized language proficiency testing, in the context of language minority education, is only about two decades old. Much remains to be learned. Finally, there is little doubt that any procedure for assessing a learner's language proficiency must also entail the use of additional strategically selected measures (e.g., teacher judgments, miscue analysis, writing samples).

The Tests Described

The English language proficiency tests presented in this Guide are the:

1) Basic Inventory of Natural Language (Herbert, 1979);
2) Bilingual Syntax Measure (Burt, Dulay & Hernández-Chávez, 1975);
3) Idea Proficiency Test (Dalton, 1978;94);
4) Language Assessment Scales (De Avila & Duncan, 1978; 1991); and
5) Woodcock-Muñoz Language Survey (1993).

With the exception of the Woodcock-Muñoz Language Survey, the other 4 tests included in this handbook are the most commonly used tests in Title VII funded bilingual education programs (CCSSO, 1992). This was the criterion used for selecting these particular tests. The Woodcock-Muñoz Language Survey was included because it represents one of the more recently developed tests.

Obviously, it is beyond the scope of this guide to present all of the commercially available English language proficiency tests. Figure 1 in the next section of the handbook provides the reader with a list and a brief description of the five tests that will be described in more detail in the following sections.

[table of contents]

Organization of the Language Proficiency Test Handbook

This section of the handbook contains a brief description of the 5 language proficiency tests presented in the handbook. We have provided information about the test materials, cost, test administration and other pragmatic concerns. This information is provided as a "how to" guide. It is meant to be neither a critique nor an endorsement of any of the tests included in this handbook. The reader should acquire sufficient information to assist with the selection of a test or tests for assessing the language proficiency of students.

Some of the information included in this handbook relates to "technical standards". Technical standards are, " those guidelines related to an assessment that are specified by psychometricians, statisticians, test publishers, and specialists in the domains covered by the assessment." (Wheeler & Haertel, 1993, p.139). Usually technical standards include the reliability and validity tests conducted for any instrument. When considering the technical standards of a test it is important to keep in mind the characteristics of the children to be tested. By characteristics we mean the ages, grades, language background, exposure to United States public school systems, including the contexts and values of these systems, and other factors that may have an impact on the students' ability to perform on the test. The validity of the test may be compromised if the students in your school or program are not represented in the norm group for the test.

In order to make the meaning of these standards clear and accessible to the reader, we also have included a glossary of terms in this section. If you are interested in additional information on the topic of testing and technical standards, we recommend the following useful references:

American Psychological Association. (1985). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

Anastasi, A. (1988). *Psychological Testing* (sixth edition). New York, NY: MacMillan Publishing Company.

Durán, R.P. (1988). Validity and Language Skills Assessment: Non-English Background Students. In H. Wainer & H.I. Braun (Eds.), *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Wheeler, P. & Haertel, G.D. (1993). *Resource Handbook on Performance Assessment and Measurement: A Tool for Students, Practitioners, and Policymakers*. Berkeley, CA: The Owl Press.

There are many useful publications available on testing practice and ethics. As minority language students are sometimes not able to protect themselves adequately from unfair testing practices we consider it very important to be careful, informed consumers of standardized language proficiency tests and other tests that might adversely impact a student's individual educational progress. The glossary of terms, following Figure 1, which presents information on how to find/buy these tests, should provide the reader with background information necessary to understanding the descriptive information included in this handbook on the tests. If you are already familiar with these terms, please skip the glossary and begin with the test review sections.

Test Descriptions and Publisher Information

Figure 1:

Five Standardized English Language Proficiency Tests Included in this Handbook

Assessment Instrument	**General Description**
Basic Inventory of Natural Language (BINL) CHECpoint Systems, Inc. 1520 North Waterman Ave. San Bernadino, CA 92404 1-800-635-1235	The BINL (1979) is used to generate a measure of the K-12 student's oral language proficiency. The test must be administered individually and uses large photographs to elicit unstructured, spontaneous language samples from the student which must be tape-recorded for scoring purposes. The student's language sample is scored based on fluency, level of complexity and average sentence length. The test can be used for more than 32 different languages.
Bilingual Syntax Measure (BSM) I and II Psychological Corporation P.O. Box 839954 San Antonio, TX 78283 1-800-228-0752	The BSM I (1975) is designed to generate a measure of the K-2 student's oral language proficiency; BSM II (1978) is designed for grades 3 through 12. The oral language sample is elicited using cartoon drawings with specific questions asked by the examiner. The student's score is based on whether or not the student produces the desired grammatical structure in their responses. Both the BSM I & BSM II are available in Spanish and English.
Idea Proficiency Tests (IPT) Ballard & Tighe Publishers 480 Atlas Street Brea, CA 92621 1-800-321-4332	The various forms of the IPT (1978 & 1994) are designed to generate measures of oral proficiency and reading and writing ability for students in grades K through adult. The oral measure must be individually administered but the reading and writing tests can be administered in small groups. In general, the tests can be described as discrete-

	point, measuring content such as vocabulary, syntax, and reading for understanding. All forms of the IPT are available in Spanish and English.
Language Assessment Scales (LAS) CTB MacMillan McGraw-Hill 2500 Garden Road Monterey, CA 93940 1-800-538-9547	The various forms of the LAS (1978 & 1991) are designed to generate measures of oral proficiency and reading and writing ability for students in grades K through adult. The oral measure must be individually administered but the reading and writing tests can be administered in small groups. In general, the tests can be described as discrete-point and holistic, measuring content such as vocabulary, minimal pairs, listening comprehension and story retelling . All forms of the LAS are available in Spanish and English.
Woodcock-Muñoz Language Survey Riverside Publishing Co. 8420 Bryn Mawr Ave. Chicago, IL 60631 1-800-323-9540	The Language Survey (1993) is designed to generate measures of cognitive aspects of language proficiency for oral language as well as reading and writing for individuals 48 months and older. All parts of this test must be individually administered. The test is discrete-point in nature and measures content such as vocabulary, verbal analogies, and letter-word identification. The Language Survey is available in Spanish and English.

Note: Addresses and telephone numbers current as of November, 1995.

Glossary

The following terms and definitions should prove useful in helping to understand the information to follow:

Basal and Ceiling Rules are guides for minimizing testing time. Test items are arranged in order of difficulty, with the easiest item first and the most difficult item last. The chronological age of the examinee is used to identify the basal or easiest items that examinee is expected to answer. The ceiling for a particular

examinee is identified after a specific number of items are missed consecutively (anywhere from 4 items to an entire page of items may need to be answered incorrectly to identify the ceiling).

Correlation Coefficient is a statistical measure of the linear or curvilinear relationship between two variables, scores, or assessments. The correlation coefficient ranges from -1.0 to +1.0; when there is no relationship between two variables, it equals 0. A negative value indicates that as the value of one variable increases, the other variable tends to decrease. A positive value indicates that as one increases in value, so does the other and that as one decreases in value, so does the other. Correlation is used in both reliability and validity studies for tests.

Criterion Referenced Test (CRT) is a test developed and used to estimate how much of the content and skills covered in a specific content area have been acquired by the examinee. Performance is judged in relation to a set of criteria rather than in comparison to the performance of other individuals tested with a norm-referenced test (NRT).

Discrete Point refers to test items that measure a single unit of content knowledge in a particular domain. These items are usually multiple choice, true-false, or fill-in-the-blank and allow for only one correct answer.

Holistic Score is the assignment of a single score that reflects an overall impression of performance on a measure. Scores are defined by prescribed descriptions of the levels of performance, examples of benchmarks at each level, or by scoring rubrics.

Interpolation is a process of using available points based on actual data to estimate values between two points. Interpolation is usually done by calculation based on the distance between two known points but also can be done geometrically by connecting the points.

Language Dominance refers to the general observation that a bilingual or multilingual individual has greater facility in one language as opposed to the others. However this linguistic facility can vary based on the context of language use (e.g., school, church) and linguistic skill (speaking, writing).

Lexical refers to the lexicon of a language and is roughly equivalent to the vocabulary or dictionary of a language. Another term used by linguists at this level of a language is *semantics*. Semantics is the study of meanings at the word or sentence level.

Morphological refers to how words are constructed; a morpheme is essentially the smallest unit of language which conveys meaning. Some morphemes can stand alone (e.g., test) while others can only appear in conjunction with other morphemes (e.g., ed).

Norm Group is the sample of examinees drawn from a particular population and whose test scores are used as the foundation for the development of group norms. *Group norms* are the statistical data that represent the average (usually mean score) results for various groups rather than the scores of individuals within one of these groups or individuals across all groups. Only to the degree that persons in the norm group are like the persons to whom one wishes to administer a test can proper interpretation of test results be made. In other words, the test is not valid for persons who are not like the norm group.

Norm Referenced Test (NRT) is an instrument developed and used to estimate how the individuals being assessed compare to other individuals in terms of performance on the test. Individual performance is judged

in comparison to other individuals tested, rather than against a set of criteria (criterion referenced test) or in a broad knowledge area (domain referenced test).

Normal Curve Equivalent (NCE) is a transformation of raw test scores to a scale with a mean of 50 and a standard deviation of 21.06. NCEs permit conversion of percentile ranks to a scale that has equal intervals of performance differences across the full range of scores and which can be arithmetically manipulated. Percentile ranks can not be used in arithmetic calculations.

Phonological (graphonic) refers to the smallest, distinguishable unit of sound in a language which help convey meaning (i.e., a phoneme). In isolation, however, the phoneme /p/ means nothing. Graphonic refers to the visual, graphic representation of the phonological system of a language which make reading and writing possible.

Pragmatic is the dimension of language which is concerned with the appropriate use of language in social contexts. In other words, pragmatics has to do with the variety of functions to which language is put (e.g., giving instructions, complaining, requesting)

and how these functions are governed depending on the social context (e.g., speaking to a teacher versus a student; at school versus at home).

Reliability is the degree to which a test or assessment consistently measures whatever it measures. It is expressed numerically, usually as a correlation coefficient. There are several different types of reliability including:

> *Alternate Forms Reliability* is sometimes referred to as parallel forms or equivalent forms reliability. Alternate forms of a test are test forms designed to measure the same content area using items that are different yet equivalent. This type of reliability is conducted by correlating the scores on two different forms of the same test.
>
> *Intra Rater Reliability* is the degree to which a test yields consistent results over different administrations with the same individual performing at the same level by the same assessor (intra-rater).
>
> *Inter Rater Reliability* is the degree to which an instrument yields the same results for the same individual at the same time with more than one assessor (inter-rater).
>
> *Internal Consistency Reliability* is sometimes called split half reliability and is the degree to which specific observations or items consistently measure the same attribute. It is measured in a variety of ways including Kuder Richardson 20 or 21, Coefficient Alpha, Cronbach's Alpha, and Spearman Brown Prophecy Formula. These methods yield a correlation coefficient that measures the degree of relationship between test items.

Rubric is sometimes referred to as a scoring rubric and is a set of rules, guidelines, or benchmarks at different levels of performance, or prescribed descriptors for use in quantifying measures of attributes and performance. Rubrics can be holistic, analytic or primary trait depending upon how discretely the defined behavior or performance is to be rated.

Stratified Sampling Design is the process of selecting a sample in such a way that identified subgroups in

the population are represented in the sample in the same proportion that they exist in the population.

Syntactic level of a language is equivalent to the grammar of the language. This level (sometimes referred to as syntax) involves the way words are combined in rule-governed order.

Validity is the degree to which a test measures what it is supposed to measure. A test is not valid per se; it is valid for a particular purpose and for a particular group. Validity evidence can come from different sources such as theory, research or statistical analyses. There are different kinds of validity including:

> *Content Validity* is the degree to which a test measures and intended content area. Item validity is concerned with whether the test items represent measurement in the intended content area. Sampling validity is concerned with how well the test samples the total content area. Content validity is usually determined by expert judgement of the appropriateness of the items to measure the specified content area.
>
> *Construct Validity* is the degree to which a test measures an independent hypothetical construct. A construct is an intangible, unobserved trait such as intelligence which explains behavior. Validating a test of a construct involves testing hypotheses deduced from a theory concerning the construct.
>
> *Concurrent Validity* is the degree to which the scores on a test are related to the scores on another, already established test administered at the same time, or to some other valid criterion available at the same time. The relationship method of determining concurrent validity involves determining the relationship between scores on the test and scores on some other established test or criterion. The discrimination method of establishing concurrent validity involves determining whether test scores can be used to discriminate between persons who possess a certain characteristic and those who do not, or those who possess it to a greater degree. This type of validity is sometimes referred to as criterion-related validity.
>
> *Predictive Validity* is the degree to which a test can predict how well an individual will do in a future situation. It is determined by establishing the relationship between scores on the test and some measure of success in the situation of interest. The test that is used to predict success is referred to as the predictor and the behavior that is predicted is the criterion.

Organization of the Test Information

The next section of the handbook will provide the reader with thorough reviews of all five of the language proficiency tests. Each review consists of 10 points of information that include some very concrete items such as the purpose of the test, administration time, cost and scoring. In addition, the theoretical foundation for the test and the technical standards of reliability and validity are addressed. The final point of information provided directs the reader to critiques of the test.

The authors reviewed all test materials provided by the publishers including actual test items, technical and administration manuals and scoring forms, guidelines and computer programs. Test critiques for each test were reviewed. The tests were administered to a small non-random sample of students when necessary to understand the administration procedures and scoring. Some of the descriptions of the tests vary as a result of the information available about the tests in the manuals and other documentation. For example, the Woodcock-Muñoz Language Survey includes a wider variety of validity studies. The validity studies for

this test were reported as described in the test manual. Some of the other tests conducted fewer validity studies but more reliability studies. In each case, we reported the information available in the test manuals and so, each of the test descriptions varies dependent upon the test documentation.

Although we wanted to make the theoretical foundation and the reliability and validity sections in this handbook parallel for each test, this was not possible. The test developers used the theoretical foundation and conducted the reliability and validity studies which they felt necessary to the construction of their instrument. All test developers presented some evidence to support both the reliability and validity of their instruments. The reader is urged to consult one of the sources cited earlier in this section (p. 16) for additional information about the minimum requirements for establishing test reliability and validity and compare these guidelines to the information presented in the technical standards manuals for any of these tests.

[table of contents]

Basic Inventory of Natural Language (BINL)

Purpose

According to the *Technical Report* (CHECpoint Systems, 1991), the BINL can be used as an indicator of oral language dominance and/or language proficiency. It also can be used as a tool to monitor language development. The BINL scoring system leads to the following four language proficiency classifications:

- non-English speaking (NES)
- limited English speaking (LES)
- fluent English speaking (FES)
- proficient English speaking (PES)

Age and Language Groups

The BINL is appropriate for students in grades K through 12 between the ages of 4.9 and 17.8 years. The test was normed on a sample of 8,134 judged to be native English speaking students. The norming sample roughly approximated the population distribution described in the United States 1980 Census. The following data typify the norming sample:

- 51.8% of the sample was female; 48.2% was male;
- .Samples from four geographic regions were taken (NY, MA, CN, NJ; OH, IL, MI; TX, LA, FL; and CA, NM, AZ, CO, HI, OR);
- 73% of the scores came from urban sites and 27% came from rural sites;
- 67% of the sample was White, 6% Black, 18% Hispanic, 6% Asian and 3% American Indian.

The BINL can also be used to assess the oral language ability of students in 32 different languages. However, the test administrator must be able to speak the non-English language well enough to administer the test in a valid and reliable manner.

Administration Time

The author of the BINL recommends that the test users take some time to familiarize themselves and the students with the test format (i.e., an oral picture describing type activity). Familiarization with the test can be done with small groups as part of regular classroom activity. The author estimates the average administration time to be about 10 minutes per student.

Cost

The cost of the complete BINL Kit for elementary (Grades K-6, Forms A or B) or secondary students (Grades 7-12, Forms C or D) is currently $59.00. The kit includes: Examiner's Manual, 400 oral score sheets, class oral language profile card, 20 full-color poster story starters and 80 talk-tiles to elicit speech. The oral language samples can be machine scored through the publisher at a cost of 80 per score sheet. School districts may also opt for purchasing a rapid scoring program diskette for $285. Other BINL products (e.g., instructional materials) and services (e.g., scoring written samples in English and other languages) are also available.

Test Administration

An oral language sample is elicited from the student using real-life, colored posters approximately 16" X 11" in size. The technique used to elicit the oral sample is referred to as a "free-sampling technique." Essentially, the intent of the test is to obtain as natural an oral sample from the student as possible within a picture-description type of test format. The student selects 3 to 5 pictures s/he prefers to talk about. The person administering the test must tape-record the student's oral sample and must avoid using direct questions to prompt the student or correct the student. The test administrator is asked to keep the student in an information-giving role and to maintain the testing situation as a logical communication activity.

Items

Because the BINL uses a free-sampling technique, there are no specific test items. However, the student's oral language sample is scored based on three linguistic features:

- fluency or the total number of words used in the language sample;
- level of syntactic complexity (i.e., grammatical competence); and
- average sentence length.

Scoring

The oral language sample can be either hand or computer scored. However, the author suggests computer-scoring when a large number of students are tested and staff are not trained in grammatical analysis. In either case, the scoring of the oral language sample hinges on the three factors previously mentioned: fluency, level of syntactic complexity and average sentence length. A fluency score or value is determined based on the total number of different words used in the language sample. The average sentence length value is generated based on the fluency count and the number of phrases or sentences used by the student.

The level of syntactic complexity score is calculated based on a predetermined numerical value assigned to various word classes. For example, an adjective is assigned a value of 14, an adverb a value of 15, and a preposition a value of 20.

Ultimately, the fluency, level of syntactic complexity and average sentence length values are combined to generate a language proficiency classification.

Test Design and Theoretical Foundation for the Test

According to the author, the BINL is a criterion referenced instrument. The intent of the test is to obtain a natural oral sample from the student within a picture-description type of test format. Unlike other tests, the BINL does not use specific test items to which students must respond and which are later scored as either correct or incorrect. Rather, the test design of the BINL is intended to measure language proficiency through the process of eliciting a natural language sample which is later examined for specific linguistic qualities.

The theoretical foundation underlying the BINL is best characterized as one which firmly rests upon the varying complexity of syntax or grammatical structures. In the *Technical Report* for the BINL (CHECpoint Systems, 1991), proficiency is defined as "the degree of command of a particular language or dialect; that is, the ability a student has in using the various components of language such as vocabulary, structure and morphology" (p. 1).

The author of the BINL bases the design of the instrument on linguistic research which supports the following three findings. Vocabulary, syntactic (grammatical) structures and sentence length vary on a developmental continuum. In other words, research has shown that a young child's breadth of vocabulary, use of complex syntactical structures and ability to use longer and longer sentences develops gradually and predictably as the child grows older. However, it is the child's ability to use syntactic structures (i.e., various word classes, types of phrases and sentence types) of varying complexity that provides the theoretical foundation for the BINL.

Reliability and Validity

Evidence for the reliability of the BINL is based on two correlation analyses. The language classification of a student is based on what is called the complexity level score. This score is derived by scoring ten orally produced sentences. The author of the test performed a split-half correlation. That is, the author examined the correlation between the complexity level score of the first five sentences with the complexity level score for the remaining five sentences. A correlation of .925 was found in this analysis involving 2808 students in grades K through 12.

Test-retest analyses were conducted also. A total of 118, K-6 students with a non-English language in the home were tested then retested two to four weeks later. The correlation coefficient was .80. In other words, and in this case, some students scored differently on the retest; mean scores increased from 31.89 to 43.13.

Regarding the construct validity of the BINL, the author sets forth these three pieces of evidence:

(1) 116 holistically scored *writing samples* from a southern California high school were scored as BINL oral language production samples; the correlation between the two sets of scores was .81.

(2) Teacher *oral proficiency judgments* of 263 limited English proficient and fluent English proficient were compared with the same two classifications using BINL scoring procedures; overall agreement was 64%.

(3) Three separate studies using *reading passages* (taken from the Gilmore Oral Reading Test, Diagnostic Reading Scales and Longman Structural Readers) of varying and graduated syntactic complexity were conducted; the studies showed high correlations (.98 to .99) between the BINL complexity scores and the reading passages

Critiques

Critiques of the BINL can be found in the following three sources:

Carrasquillo, A. (1992). Basic Inventory of Natural Language. In D.J. Keyser & R.C. Sweetland (Eds.), *Test Critiques Volume IX*. Austin: Pro-ed publications.

Guyette, T.W. (1985). Review of the Basic Inventory of Natural Language. In J.V. Mitchell (Ed.), *The Ninth Mental Measurements Yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.

Hargett, G. (1987). Basic Inventory of Natural Language. In J.C. Alderson, K.J. Krahnke, & C.W. Stansfield (Eds.), *Review of English Language Proficiency Tests*. Washington, D.C.: Teachers of English to Speakers of Other Languages.

[table of contents]

Bilingual Syntax Measure I and II (BSM I & II English)

Purpose

Based on the *Technical Manual* of the English version of the BSM I (Burt, Dulay, & Hernández-Chávez, 1976), the purpose for this test is to provide a measure of oral language proficiency. The *Technical Handbook* for the BSM II (Burt, Dulay, & Hernández-Chávez, & Taleporos, 1980) indicates that this version of the test was designed for the same purpose as the BSM I. The target population for the BSM I includes students in grades K through 2 and the BSM II is for students in grades 3-12. Both tests lead to the following language proficiency classifications:

- Level 1: No English
- Level 2: Receptive English Only
- Level 3: Survival English
- Level 4: Intermediate English
- Level 5: Proficient English

The BSM II uses the same classifications as the BSM I from Level I to Level 4. The BSM II differs from BSM I in that two additional classifications are possible:

- Level 5: Proficient English I
- Level 6: Proficient English II

The results from either of the two tests can be used to make placement decisions, provide formative and summative program evaluation data, make diagnoses, and collect language acquisition research data.

Age and Language Groups

According to the *Technical Handbook* for the BSM I which was designed for younger students in grades K-2, the test was field tested in 1974 on 1,371 students having some of the following characteristics:

- Some of the students were native English speakers and others were native Spanish speakers; The age of the students ranged between 69.4 and 93.8 months;
- 25.9% of the sample came from urban communities, 31% from rural communities, 20.9% from suburban communities and 22.2% came from an urban-rural mixed community;
- 16% of the sample came from the North Atlantic region of the U.S., 13.2% from the Southeast, 16% from the Great Lakes & Plains, and 54.8% came from the West and Southwest;
- 47.6% of the sample were classified as Chicano, 8.4% Cuban, 10.8% Puerto Rican, 2.9% Other Spanish, 5.1% Afro-American, 23.9% Anglo-American, and 1.3% were classified as Other non-Spanish.

The BSM II, intended for learners in grades 3-12, was field tested in 1977 on 758 Spanish and English speaking students sharing some of the following characteristics:

- The age of the students ranged between 8.7 and 17.9 years of age;
- No reference is made to the distribution of the sample drawn from urban, rural, suburban or urban-rural sites;
- The sample came from the same geographic regions as the BSM I; however, no percentages are provided; and
- 60% of the sample were classified as Mexican-American, 12% other Hispanic, 1.8% Black, 17%

White (non-Hispanic), 3.4% Other and 5.8% no response.

Administration Time

The administration time for the BSM I is approximately 10 to 15 minutes per child and may vary depending on how quickly the child responds, how long the responses are and how quickly the examiner can write the child's responses down. Neither the *Technical Handbook* (1980) nor the *Manual* (1975) for the BSM II indicate how long it takes to administer the BSM II. Hayes-Brown (1987) indicates it takes 10-15 minutes to administer and Carrasquillo (1988) states from 15 to 20 minutes.

Cost

The current cost for a complete kit of the BSM I or BSM II, which includes the English and Spanish editions of the test, is $302.50. The kit includes a storage box, picture booklet, administration manuals in both English and Spanish, *Technical Handbook*, 35 response booklets in both languages and 2 class record sheets. If only the English version of the BSM I or BSM II is needed, the above items must be purchased separately for a total cost of $206.50.

Test Administration

The *Administration Manual* for the BSM I in English describes the administration of the test as similar to "having a chat with a child about some pleasant pictures" (1975, p. 6). The tool used to elicit the oral language sample from the student is a series of multi-colored, cartoon-like pictures. There are a total of 25 test questions, some of which correspond only to specific pictures. Some questions also require the examiner to point to specific aspects of the pictures as the questions are asked. Further, these questions are designed to elicit the use of specific English syntax structures. The examiner must write the student responses exactly as they are produced on the appropriate line in the test response booklet. The examiner must be able to read and write English and maintain a natural flow of conversation throughout the administration of the test.

Because the BSM II is intended for older students, a few aspects of the test administration are different than the BSM I. For example, the cartoon storyline for the BSM II is more sophisticated and contains a beginning, middle and end. Again, the questions asked by the test administrator elicit more complex grammatical structures appropriate for older students. Nonetheless, the administration procedure for the BSM II is much like the procedure described for the BSM I. That is, the test administrator is asked to try to establish a smooth and real conversational exchange with the student.

Items

The items on the BSM I are the 18 (scored) test questions. Other questions are asked but not scored. Some of the test items have underlined words which should be stressed. As stated previously, the test questions are designed to elicit the use of specific syntactic structures. For example, the BSM I is designed to elicit the use of the present progressive tense, plural forms, auxiliaries, past irregular verb forms, and the conditional perfect. It is also important to indicate that the first test question is not necessarily going to require the use of the simplest syntactic structure. In other words, the test questions do not move from the use of simple syntactic structures to progressively more and more complex syntactic structures.

The test items on the BSM II consist of 22 (scored) questions also orally presented by the examiner and

intended to elicit the use of specific syntactic structures appropriate to learners in grades 3 through 12. Additional test questions are asked but need not be scored. The grammatical constructions elicited by the questions entail structures such as complex sentence types, conditional subordinate clauses, and a variety of phrase structures. Similar to the BSM I, the syntactic structures elicited vary in their complexity but do not necessarily appear in an increasing order of difficulty.

Scoring

The BSM I and the BSM II are hand scored. The procedures for determining a student's proficiency level in English require basically the same three steps. First, the scorer must make a determination as to whether or not a child responded to a minimum number of test questions in English. If this criterion is met, then the scorer proceeds to step two. At this point, the scorer makes a determination as to the number of grammatically correct student responses to a given set of the test questions. Specific guidelines are provided for determining grammaticality. Finally, the student's level of proficiency is determined based on the number of correct responses to specific test items. In effect, those students demonstrating competency with more advanced syntactic structures are classified as more orally proficient in the English language.

Test Design and Theoretical Foundation for the Test

As its name suggests, the Bilingual Syntax Measure I & II are tests whose foundation rests on syntax or what is sometimes referred to as grammar. The *Technical Handbook* (1980, p. 3) for the BSM II states:

> The Bilingual Syntax Measure series (both BSM I and BSM II) is built upon two fundamental concepts about language development:
>
> 1) Children acquire language largely through a process of "creative construction."
>
> 2) Language structures are acquired in an ordered, hierarchical fashion: certain grammatical structures are learned before other more complex structures.

Briefly, the Bilingual Syntax Measure I & II are based on research which demonstrates that children acquire syntactic structures in a systematic and hierarchical manner as they progress through stages in language acquisition. Further, the authors drew on research which has shown that even children from diverse language backgrounds appear to acquire certain syntactic structures in a common order.

There is one fundamental difference between the test design of the BSM I and the BSM II. According to the BSM II *Technical Handbook* (1980), the BSM II is based on *Item Hierarchies* as opposed to *Structure Hierarchies*. In other words, the raw data which underlies the design of the BSM II are based on the responses field test students gave to pairs of items and not on a structural analysis of their responses.

Reliability and Validity

Two analyses were conducted to demonstrate the reliability of the BSM I. The first analysis was a test-retest reliability procedure. 147 students in grades K-2 were tested twice, two weeks apart, in order to examine consistency in their language proficiency classification. The results of this procedure indicated that only five children were classified more than one level apart on the two administrations.

The second procedure entailed a measure of inter-rater reliability. In this case, two scorers were given the

same set of student responses to score and the student classifications were then compared for consistency. Two hundred and seventy one tests from students in grades K-2 were scored by two independent scorers. The scorers agreed on the student classification in 83.8% of the cases.

The reliability of the BSM II included various analyses. An analysis of internal consistency reliability yielded coefficients ranging from .80 to .90, depending on the item groups or clusters. Analyses of test-retest reliability indicate that 58 of the 85 children were classified at the same exact level. The students took the tests from one to two weeks apart.

The *Technical Manual* (1976, p. 32) of the BSM (I) sets forth three pieces of evidence which the test developers believe support the construct validity of the test:

1. the body of psycholinguistic theory and research supporting the concept of natural sequences of acquisition;

2. the statistical procedures and research which have resulted in the production of the acquisition hierarchy and of the scoring system which is based on the hierarchy; and

3. evidence that the BSM classifications reflect the relationships expected to be found among bilingual children.

The *Technical Handbook* for the BSM II (1980) indicates that the techniques for the construct validation of this test were similar to those used for the BSM I. Again, the main difference lies in the fact that the BSM II is based upon what the authors call an *Item Hierarchy*. Briefly, the essence of this hierarchy is linked to how the field test students responded to pairs of items. The response patterns of the students allowed the test developers to ascertain which items were more difficult than others and consequently to determine their value within the scoring system.

Critiques

Critiques of the Bilingual Syntax Measure I and II can be found in:

Bejar, I. (1978). Review of the Bilingual Syntax Measure. In O.K. Buros (Ed.), *The Eighth Mental Measurements Yearbook*. Highland Park, NJ: Gryphon.

Carrasquillo, A. (1988). Bilingual Syntax Measure II Test. In D.J. Keyser and R.C. Sweetland (Eds.), *Test Critiques Volume VII*. Kansas City, MO: Test Corp. of America.

Cziko, G. (1987). Bilingual Syntax Measure I. In J.C. Alderson, K.J. Krahnke, & C.W. Stansfield (Eds.), *Review of English Language Proficiency Tests*. Washington, D.C.: TESOL.

García, E. (1985). Review of the Bilingual Syntax Measure II. In J.V. Mitchell (Ed.), *The Ninth Mental Measurements Yearbook Volume I*. Lincoln, NE: University of Nebraska Press.

[table of contents]

IDEA Proficiency Tests (IPT)

Purpose

The IPT was designed to assist districts with the process of identifying limited English proficient students (LEP) and with redesignating these students as fluent English proficient for placement in mainstream classes after a period of instruction in special programs for LEP students. The IPT scoring system leads to the following categories of oral language proficiency:

IPT Oral Language Proficiency *Designations:*	IPT Reading Proficiency *Designations:*	IPT Writing Proficiency *Designations:*
non-English speaking (NES) limited English speaking (LES) fluent English speaking (FES)	non-English reader (NER) limited English reader(LER) competent English reader (CER)	non-English writer(NEW) limited English writer (LEW) competent English writer (CEW)

In 1979, the Oral IPT was designed to identify NES and LES students for special assistance or redesignate them for mainstream programs. A version of the oral test for older students was developed and published in 1983. The Reading and Writing test was developed in 1992 especially for language minority students to assess reading and writing and to complement the oral language proficiency test. Together the IPT Oral and the Reading and Writing Tests assess overall language proficiency.

Age and Language Groups

Several different versions of the IPT were developed for specific age groups. The Pre-IPT English was designed to test the oral language proficiency of preschool students. The IPT I was designed to assess oral proficiency for students in grades

K - 6. The IPT II is used for students in grades 7 - 12. The Reading and Writing Tests are sold in three versions: IPT 1 is for grades 2 - 3; the IPT 2 is for grades 4 - 6; and the IPT 3 is for grades 7 - 12.

These versions of the IPT Oral Proficiency Tests and Reading and Writing Tests are available in English and Spanish. The English versions may be used to assess English language proficiency for any language group and the Spanish tests can be used to assess Spanish proficiency for any language group.

Each of the versions of the test has information about the configuration of the norm group sample including names of schools, districts, and states participating, the age, gender, and ethnicity of the students in the norm

group, and the primary languages of the students. As there are so many versions of the tests available, the information about the norm group for the IPT I (K-6 grade Oral Proficiency Test) is presented for illustrative purposes. The norm group was characterized by:

- 44.9 % of the sample was male and 55.1% of the sample was female;
- states included in the sample were AZ, CA, FL, IL, MI, NM, TX, and WI;
- student ages ranged from 5 years (11.2% of the sample) to 12 years (7.0% of the sample). The percentage of students for each age between 5 and 12 years was reported;
- the ethnic groups included in the sample were: Spanish (53.3%), English (25.6%), Korean (4.9%), Chinese (4.9%), Japanese (3.7%), Southeast Asian (3.4%), and "Other" (6.7%);
- students in kindergarten comprised 15.1% of the sample, first grade 17.4%, second grade 15.8%, third grade 14.2%, fourth grade, 14.2%, fifth grade, 13.4%, and sixth grade, 9.9%.

The reader is urged to review the norm group information for the specific form of the IPT to be used.

Administration Time:

The length of time involved in the administration of the oral form of the IPT varies depending upon two variables: (1) the student's English language proficiency and (2) the level at which to start testing for LES/FES students. Administration time ranges from 5 minutes for a non English proficient student (NES) to 20 minutes for a fluent English proficient student (FES). Average test time is 14 minutes.

The Reading and Writing Tests are untimed. The *Examiner's Manual* for the various forms provides a table of approximate time ranges for each section of the reading/writing tests. The Reading Test will take from 45-70 minutes to administer (Amori, Dalton, & Tighe, 1992, p. 7).

READING:

Part 1: Vocabulary	5-10 minutes
Part 2: Vocabulary in Context	5-10 minutes
Part 3: Reading for Understanding	10-18 minutes
Part 4: Reading for Life Skills	8-15 minutes
Part 5: Language Usage	4-9 minutes

The Writing Test will take from 25-45 minutes to administer (Amori, Dalton, & Tighe, 1992, p. 8).

WRITING:

Part 1: Conventions	4-9 minutes
Part 2: Write a Story	4-10 minutes
Part 3: Write Your Own Story	12-20 minutes

The IPT publisher recommends allowing a break between the Reading and Writing Tests. The Writing test can be given later in the day or on the next day.

Cost:

The cost of the Oral Proficiency Test set with 50 test booklets is $105.00. An additional 50 test booklets is $29.00. Components of the Oral Language Test set can be purchased separately but if purchased individually the entire test would cost $109.50. All forms (A, B, C, D) of the IPT I and the IPT II in either Spanish or English are priced the same ($105 for the tests set and $29 for additional test booklets). The Reading and Writing Test Set prices vary. For the IPT Reading & Writing Tests 2 and 3 (grades 4-6 and 7-12 respectively), each set includes 50 reading test booklets, 50 reading test answer sheets, 50 writing test booklets (consumable), a scoring template, and an *Examiner's Manual* and a *Technical Manual* for $160.00. The IPT Reading & Writing Test 1 set (for grades 2 and 3) includes the same components except that there are 50 consumable reading test booklets and no test answer sheets (students write their answers in the test booklets) for $150.00. The publisher offers a lower priced ($125.00) version of the IPT 1 for Form A which can not be machine scored. The set components are sold separately. Extra reading test answer sheets are $25.00 and extra test writing test booklets are $37.00. The technical and examiner's manuals are $29.00 and $39.00 respectively.

Test Administration

The IDEA Oral Language Proficiency Test assesses four basic areas of English oral language proficiency: Vocabulary, Comprehension, Syntax, and Verbal Expression which includes Articulation. There are six levels of difficulty tested (A - F) and all students are tested individually. A student advances through the levels of the test until the test is completed or stops at a proficiency level as indicated by information in the score box at the end of each level.

The IPT Reading and Writing Test is a group administered standardized test. The Reading Test consists of five parts: Vocabulary, Vocabulary in Context, Reading for Understanding, Reading for Life Skills, and Language Usage. The Writing Test Has three parts: Conventions, Write a Story and Write Your Own Story. Three different reading level designations (a non-English reader, a limited English reader, and a competent English reader)and three writing level designations (a non-English writer, a limited English writer and a competent English writer) are identified after the Reading and Writing Test is scored.

Items

In the appendix of the *Examiner's Manual* for the IPT I Oral Proficiency Test is a matrix of test items for each skill area and developmental level. The oral proficiency items consist of questions and statements to the student (e.g., "Stand up and turn around", or "What is your name?") and colorful picture cards for identification of an action verb (e.g., riding a bike, were singing/sang) or a noun (e.g., helicopter, stove).

The items for the Reading and Writing Test are presented in a traditional multiple choice, bubble in the correct answer format. Reading test items have four possible responses among which to choose. The writing multiple choice items have three possible answers. The Writing test includes two sections that require the student to write a story. The first section has a series of three pictures that "tell" a story. The student is expected to write about the sequence of events depicted. The second writing sample has two different

picture prompts (for Form 1A English, Writing one picture is of several children watching television and the other picture is of a girl opening a present) and the student is expected to write a story about one of the prompts. Other forms of the Reading and Writing Tests include similar items that are appropriate developmentally for the students' age. The test consumer needs to keep in mind the educational experiences and background of the examinee for the Reading and Writing test to be valid.

Scoring

Scoring the Oral Language Proficiency Test is a straightforward procedure. Each item has a corresponding response line which gives the desired answer. It is numbered the same as the item. Each item is followed by two boxes: one to check if the item is correct and the other to check if the item is incorrect. The Oral Test is divided into six levels, A-F. At the end of each level, the examiner totals the number or errors or incorrect responses. Each level gives a maximum number of errors possible to continue the testing at the next level. If the student exceeds an error threshold, they are classified at that level or the previous level depending upon the total number of errors made. These six levels can then be translated in the NES, LES, or FES categories.

The multiple choice portions of the Reading and Writing Tests can be hand scored using a template or they can be machine scored by the publisher. The total score for this test is the total number of correct answers in each of the five subtest sections. The number correct then can be compared with the required number of correct test items for the Non-English reader, Limited English reader and Competent English Reader designations. A table in the Appendix of the *Examiner's Manual* can be used to interpret the raw scores as standard scores, percentile ranks, or normal curve equivalents (NCE).

Scoring the Writing Test involves a similar process for the multiple choice items but the two sections that involve actual student writing samples completed for the test booklet prompts, require the examiner to use the IPT Rubrics and student examples in the *Examiner's Manual* to rate the writing samples. The rating given to these two sections using the rubric is considered the score. The scores for the three writing sections are compared to a range of scores that are required for classification as a non-English writer, a Limited English Writer or a Competent English writer. The writing test can be machine scored; however, the writing samples must be rated by the examiner. Care must be taken to train examiners to use the rubric so that subjectivity and rater bias are not introduced into the scoring process. To ensure the reliability of these ratings, two examiners should rate each writing sample. If there is a discrepancy, a third rater should score the paper independently. Training scorers adequately using clear benchmark papers (samples are provided in the manual) will ensure inter-rater reliability.

Test Design and Theoretical Foundation for the Test

The IPT is made up of a series of reading, writing and oral language proficiency tests designed for use with students from Kindergarten to adult. The following forms and versions (Spanish or English) are available:

Oral Language Proficiency Tests

Pre-IPT English (for preschool students)
Pre-IPT Spanish (for preschool students)
IPT English forms A, B, C, D
IPT Spanish form
IPT II English forms A and B
IPT Spanish

Reading and Writing Proficiency Tests

IPT -1 English (grades 2-3) forms 1A and 1B
IPT -2 English (grades 4-6) forms 2A and 2B
IPT -3 English (grades 7-12) forms 3A and 3B

Six basic beliefs or premises comprise the underlying structure for the IPT family of tests: 1) language is developmental; 2) language is incremental; 3) language is systematic; 4) language is symbolic; 5) language is used to communicate in a social context; and 6) language involves both receptive and productive skills. The theoretical foundation for the IPT series is based on the four stages of language acquisition: babbling; echolalic; telegraphic; and syntactic. The babbling stage is from birth to 12 months of age and is the period during which infants play with sounds such as gurgling, cooing, raising and lowering pitch, laughing and crying. These sounds are one way the infant communicates with other people in the environment. Meaning is not attached to infant sounds until infants accidentally articulate sounds to which parents attach meaning (e.g., mammaa). At this point parents tend to repeat the sounds, smile at the child, and reinforce the sounds in a variety of other social ways. The child is then more likely to repeat the sounds.

In the echolalic stage, the child begins to use inflections, stops, pitch, pauses, and stresses in imitation of the sound patterns of the language the child hears spoken. It is during this stage that the child acquires and practices the various phonemes (individual sound units) and morphemes (syllables) of the language. During this stage, the child begins to connect a series or combination of sounds with the objects in the environment for which they stand. Labeling begins and the child uses these labels to name and describe the things around in the environment and to express needs and feelings. This stage is called the telegraphic stage because the child uses one and two word combinations to convey meaning.

The final stage is the syntactic stage. During this stage the child begin to assimilate the rules or grammar of the language. At first, simple two and three word sentences are constructed. Later as the child acquires new adjectives, adverbs prepositions and verbs, more complex communication becomes possible. Further interaction with proficient adult speakers allows the child to refine language use and communicate more clearly with those important people in the world around.

Reliability and Validity

Reliability of the various Oral Proficiency Test forms (A,B,C, & D) is extremely high. Reported internal consistency reliability, based on inter item correlations, was .99 which is unusually high. Split-half reliability for the Oral Test (form C) was lower with a .79 coefficient reported for Part 1 and .51 reported for Part 2. The scale reliability overall (both Parts 1 and 2) was .65.

The reliability for the Reading and Writing Test was reported for each of the different subtests and forms. The reliability coefficients quoted in this section are for Forms 1A and 2A English Reading and Writing tests. The reliability analyses for the Reading subtests are reported first and include the conventions Writing subtest. Reliability was established in a different manner for the other two Writing subtests separately from the Reading and Writing conventions reliability coefficients. Internal consistency was measured for each of the Reading subtests and the Writing Conventions subtest and ranged from .68 for the Reading for Understanding subtest to .84 for the Vocabulary subtest. Internal consistency for the entire Reading battery was .76. These coefficients indicate a moderate to high degree of internal consistency. Test retest reliability was measured also and ranged from .43 for Vocabulary to .71 for Vocabulary in Context. The test-retest

reliability for the entire battery was .87 which indicates that the results of comparing scores taken two weeks apart are very consistent.

Reliability for the Writing subtests involved establishing inter-rater reliability between two different raters of the examinee's writing samples. Correlations between scores given by two different readers were extremely high for all items of the writing samples ranging from a low of .90 for item 2 to a high of .98 for items A and B. Inter-rater correspondence ranged between 79% exact agreement for item A to 86% exact agreement for item 1.

Validity for the Oral Language Test included discussion of content validity, construct validity and criterion (concurrent) validity. For content validity, the reader is provided with a chart that displays the relationship between each test item and the six behavior domains measured by the test (syntax, lexicon, phonology, morphology, comprehension, and oral expression). The percentage of items devoted to measuring each domain is reported at the end of the chart and this information is offered as evidence of content validity. Several different sources of evidence are provided for construct validity. The first method used correlates the child's chronological age with knowledge and language ability (as measured by the IPT). A moderately low positive correlation of .37 for form C and .40 for form D was found. The student's grade level was also compared to IPT results and was correlated .40 for form C and .47 for form D. Grade level tests were repeated for the English only speaking students included in the norm group. Grade level was compared to IPT results and the correlations were .58 for form C and .59 for form D. Finally, criterion validity (sometimes called concurrent validity) was tested by comparing teacher predictions of student performance with IPT results. Teachers' perception (opinion) of student language proficiency and IPT results was strongly related for both test form, .73 for form C and .72 for form D. District designation of students as NES/LES/FES, as measured using methods other than the IPT, was correlated with the IPT to provide evidence for criterion validity as well. These correlations were .65 for form C and .66 for form D. Overall, many different methods and sources for evidence of the validity of the Oral IPT were presented and the reader is urged to consult the *Technical Manual* for the specific form of the IPT to be used (or purchased) and to review the reliability and validity evidence presented.

Validity for the Reading and Writing tests included the same types and data presentation format as was described for the Oral IPT. Construct, content and criterion validity tests were conducted. A matrix of the items included in the Vocabulary, Vocabulary in Context, Reading for Understanding, Reading for Life Skills, and Language Usage subtests is organized by the objective competency or concept that each item tests. This matrix is provided to the test user as evidence of the content validity of the Reading and Writing Tests.

The subtest inter-correlations are provided in a table as evidence of the construct validity of the Reading and Writing Tests. Correlation coefficient range from a low of .56 for the correlation between the Vocabulary and Writing Conventions subtests to a high of .75 between the Vocabulary in Context and Reading for Understanding subtests. The correlations in this table show the relationship between the various subtests. There are moderate correlations between the Reading subtests indicating that they are measuring related but not identical domains. The Reading subtests correlate higher among themselves than they do with the Writing Conventions subtest which is anticipated, as the writing domain is somewhat different from the reading domain.

Finally, criterion validity is presented for the Reading and Writing Test. It indicates how strongly the IPT Reading and Writing tests correlate with other independent measures of what the tests are designed to assess. Teacher ratings of students' reading and writing ability and percentile scores for the Comprehensive

Tests of Basic Skills (CTBS) were used as independent sources of evidence for the purpose of comparison to IPT Reading and Writing Tests. The CTBS/IPT norms correspondence correlation coefficients ranged from .50 to .86. When teacher ratings were compared to IPT Reading and Writing test scores, the correlation coefficients ranged from .45 to.63.

Critiques

Critiques for the IDEA Proficiency tests can be found in these critiques:

McCollum, P. (1983). The IDEA Oral Language Proficiency Test: A Critical Review. in S.S. Seidner (Ed.). *Issues in Language Assessment, Volume 2*. Chicago, IL: Illinois State Board of Education.

McCollum, P. (1987). IDEA Oral Language Proficiency Test. in J.C. Alderson, K.J. Krahnke, & C. W. Stansfield, (Eds.). *Review of English Language Proficiency Tests*. Washington, D.C.: Teachers of English to Speakers of Other Languages.

Rivera, C., & Zehler, A. (1987). IDEA Proficiency Test II. in J.C. Alderson, K.J. Krahnke, & C. W. Stansfield, (Eds.). *Review of English Language Proficiency Tests*. Washington, D.C.: Teachers of English to Speakers of Other Languages.

Stansfield, C.W. (1992). IDEA Oral Language Proficiency Test. in D.J. Keyser & R.C. Sweetland (Eds.). *Test Critiques: Volume IX*. Austin, TX: Pro-Ed.

Stansfield, C.W. (1992). IDEA Oral Language Proficiency Test-II . in D.J. Keyser & R.C. Sweetland (Eds.). *Test Critiques: Volume IX*. Austin, TX: Pro-Ed.

[table of contents]

Language Assessment Scales (LAS)

Purpose

For the purposes of this handbook, only the most recent forms of the LAS tests for grades 1-6 are described. These include: LAS-Oral (Forms 1C & 1D) and LAS-Reading/Writing (Forms 1A/1B & 2A/2B). Collectively, the LAS consists of an oral, reading and writing language proficiency assessment system

available in English and Spanish. According to the LAS Preview Materials Booklet (1991), LAS results may serve several purposes including: assessing the learner's language proficiency, placement decisions, reclassification, monitoring progress over time and pinpointing a learner's instructional needs. Because the three parts of the test can be combined in different ways, at least three proficiency classifications are possible: a LAS-Oral Score, A LAS-Reading and Writing Score, and a Language Proficiency Index (LPI) which combines the LAS-Oral, Reading and Writing scores. The proficiency levels are:

LAS Oral Score:
1 Non Speaker
2 Limited Speaker
3 Limited Speaker
4 Fluent Speaker
5 Fluent Speaker

LAS Reading and Writing Score:
1 Non Literate
2 Limited Literate
3 Competent Literate

Language Assessment Scales Language Proficiency Index Classifications

1/2	LEPa	low level reading and writing skills
1/3		mid level ("limited") listening and speaking skills
1/4	LEPb	low level reading and writing skills
1/5		high level ("proficient") listening and speaking skills
2/2	LEPc	mid level reading and writing skills
2/3		mid level ("limited") listening and speaking skills
2/4	LEPd	mid level reading and writing skills
2/5		high level ("proficient") listening and speaking skills
3/2	LEPe	high level reading and writing skills
3/3		mid level ("limited") listening and speaking skills
3/4	FEP	high level reading and writing skills
3/5		high level ("proficient") listening and speaking skills

Age and Language Groups

Different levels of the Language Assessment Scales are available depending on the age and grade of the

learners. The following represents a breakdown of the different levels of the test designed for specific grades and ages:

Grades	Instruments
Pre-K, K, Grade 1	Pre-LAS (ages4-6)
1	LAS Oral Level 1 (age 7+)
2-3	LAS Oral Level 1 LAS Reading/Writing, Level 1
4-6	LAS Oral Level 1 LAS Reading/Writing, Level 2
7-12	LAS Oral Level 2 LAS Reading/Writing, Level 3
12+	Adult LAS

The LAS-Oral (Level I) and LAS-Reading/Writing (Forms 1A/1B & 2A & 2B) for students in grades 1 through 6 were normed on approximately 1,671 learners (*Oral Technical Report,* 1990) sharing some of the following characteristics:

- No data are available on the age of the norming population; the report indicates that 55.9 % of the total norming sample (n= 3560) were in grades 1 through 6.
- 8.63% of the sample was from southern California, 34.75% from Texas, 6.33% from Northern California, 4.26% from New York and 46.04 % from Illinois and Wisconsin.
- No demographic data (e.g., urban, rural, suburban) were reported.
- No ethnicity data were reported; students tested by home language data indicate that 33% reported English as a home language and 61% came from a Spanish home language background. The remaining 6% is spread across 8 or more non-English languages.

The *Oral Technical Report* (1990) includes additional sample data such as the percentages of students by length of residency in the U.S., language status and sex.

Administration Time

According to the *Oral Administration Manual* (1990), there are no specific time limits for administering this part of the test. The instructions indicate that each student should be given a "reasonable amount of time to respond to each item" (p. 6). It should be noted that the oral part of the test can be administered in its long form (i.e., Vocabulary, Listening, Story Retelling, Minimal Pair Sounds and Phonemes) or its short form (i.e., Vocabulary, Listening, and Story Retelling). Obviously, less time will be required to administer the short form of the oral test.

The Reading and Writing Tests may be group administered and may be timed or untimed activities. Duncan and De Avila (1988) indicate that time administration limits will vary depending on the competency of the student. Nonetheless, the authors do set forth some basic guidelines for allocating time for this portion of the LAS system.

For grades 2-3, the authors recommend that the reading and writing tests be administered in two 45 to 55 minute sessions for students with language difficulties. In addition, the authors recommend that the first four subtests (i.e., Vocabulary, Fluency, Reading for Information, Mechanics and Usage) be administered in the first session and subtests five and six (i.e., Finishing Sentences and What's Happening?) be administered in the following session. Grades 4-6 follow the same procedures except under the writing portion. In this case students take the Finishing Sentences, What's Happening and Let's Write subtests during the second testing session.

Cost

Currently the cost of the Language Assessment Scales-Oral Level I, Form C (Grades 1-6) examiner's kit is $100.00. The kit includes: an *Administration Manual*, a Scoring and *Interpretation Manua*l, a cassette with Pronunciation subtest items, a cue picture booklet, and six audio cassette tapes. Answer documents contain 50 individual answer documents and must be purchased separately at a cost of $26.50.

Testing materials for the Reading and Writing parts of the Language Assessment Scales are not available in a kit form. For testing with English Level I, Form A (Grades 2-3) 35 consumable reading test booklets at a price of $66 and 35 consumable writing test booklets at a price of $32.25 must be purchased. For testing reading and writing at Level II, Form A (Grades 4-6), the following materials are needed and must be purchased at the indicated prices:

35 reusable reading test booklets $42.00
35 consumable writing test booklets $32.25
50 student answer documents $26.50

The reader should note that the *Examiner's Manual* for the Reading and Writing Tests is only included if the reading tests are purchased; otherwise, the manual must be purchased separately. Many other LAS products are available such as alternate forms of the above tests, student profile sheets, and training videos.

Test Administration

The LAS-Oral (Forms 1C & 1D) must be individually administered. The authors recommend that the test administrator (1) rehearse administering the test with family or friends, (2) read all materials, and (3) ensure that the students understand what they are expected to do before beginning each subtest. Further, the test giver should be a literate and fluent, proficient speaker of standard American English and able to discriminate between correct and incorrect responses. However, the general instructions for the test may be provided in whatever language, combination thereof or dialect. Naturally, the oral part of the test should be administered in a quiet area.

As noted earlier, the LAS-Oral consists of five separate subtests:

- Vocabulary
- Listening Comprehension

- Story Retelling
- Minimal Sound Pairs
- Phonemes.

The Vocabulary subtest consists of pictorially represented items which the test giver must point to and the student must orally identify. The Listening Comprehension subtest consists of a tape-recorded dialog to which the student must first listen and then respond to ten yes-no questions. Pausing the tape between questions is permitted. The Story Retelling subtest requires the learner to first listen to a tape-recorded story and then retell the story with the support of four corresponding picture cues.

The two optional subtests, Minimal Sound Pairs and Phonemes, are also tape-mediated. In the Minimal Sound Pairs subtest the student hears a word pair and must determine whether they sound the same or different. The Phonemes subtest simply requires the test taker to repeat a word, phrase or sentence.

Again, the Reading and Writing tests (i.e., Vocabulary, Fluency, Reading for Information, Mechanics and Usage, Finishing Sentences and What's Happening) may be administered in a small group setting but each student must generate his/her own answers in the provided test booklets. Proctors are recommended for groups of more than 15 students. The test administrator must follow specific instructions including reading instructions aloud to the students; instructions may also be provided in the student's native language. In general, sample items are provided to ensure that students are familiar with the task and marking their answers in an appropriate manner in the answer booklet before attempting each subtest. The test administrator is advised to follow the same recommendations mentioned in reference to the administration of the oral portions of the test.

Items

The Vocabulary subtests consist of a total of 20 items. The first 10 items comprising the Name That Picture subtest are concrete nouns generally found in a public school setting. These items also represent nouns which vary in their frequency and were selected from the Kucera-Francis (1967) word list. The remaining 10 items on the Action Words subtest are gerund forms of verbs (e.g., think + ing) commonly used in conversation by mainstream students in grades 1 through 6.

The Listening Comprehension subtest is essentially a tape-recorded dialog between two people supported by a cue picture. The 10 test items on the Listening Comprehension subtest require only a yes or no answer to questions asked in the present, past and future tenses.

The Story Retelling subtest has no specific test items. However, the learner's story retell is examined from various perspectives including: vocabulary, syntax, organization, transitional words, and fluency.

The test items on the Reading and Writing subtests can be described in the following manner. The Vocabulary subtest consists of 10 items each of which is pictorially represented and then followed by four printed options. These items focus on nouns and adjectives. The Fluency subtest consists of 10 individually presented sentences with a blank. Each sentence is followed by 4 options from which the student must select the most appropriate response. These items focus mainly on verbs and adjectives.

The test items on the Reading for Information subtest are 10 statements which follow a short reading passage. The student must answer True or False to each statement. The items center on factual recall. The Mechanics and Usage subtest consists of 10 items. Each item contains a blank and is followed by three

options from which the student must choose. About half the items are aimed at punctuation and the remaining items test the student's knowledge of verb tenses and pronouns.

In the Finishing Sentences subtest, the student is presented with 5 incomplete sentences. The student must write a grammatically acceptable ending for each sentence. The items are aimed at the student's ability to produce dependent and independent clauses. In the final subtest for Level 1, What's Happening, there are five picture-cued items. The items are designed so as to require the student to produce written sentences which describe what is happening in the picture. These items are designed to elicit contextually appropriate responses written in a grammatically acceptable form. The final writing subtest for Level 2, Let's Write, is designed to generate an essay. The student writes guided by a series of four sequenced picture- cues and a lead-in dependent clause.

Scoring

The oral portions of the LAS I (Grades 1-6) are scored as each subtest is administered, with the exception of the Story Retelling subtest. The Answer Booklet used by the test administrator contains the correct responses for the Vocabulary, Listening Comprehension, Minimal Sound Pairs and Phonemes subtests. As the test administrator proceeds from item to item, s/he must darken in the "bubble" for incorrect responses only.

The Story Retelling subtest is holistically scored. The examiner must first write down the student's oral story retell as accurately as possible. Tape recording the retell is advised. The student's retell is then scored using a pre-established scoring rubric on a scale of 0 to 5. Note that the examiner must first undergo what the authors call a reliability exercise in order to ensure consistency in scoring this part of the test. The authors also indicate that the scorer must be a proficient, literate speaker of the English language.

Once the incorrect items have been tallied and a holistic rating has been assigned to the oral retell, the test administrator can determine the student's total score, level, and normal curve equivalent (NCE). Each of the five parts of the oral test are weighted and simple calculations are required in order to generate these final scores.

The Story Retelling score accounts for 50% of the student's total oral score. It is important to keep in mind that if the short form of the LAS-Oral is administered, the student's score is based on only the Vocabulary, Story Retelling and Listening Comprehension subtests; the long form would include the Phonemes and Minimal Pairs subtest scores. In either case, the Story Retell accounts for one half of the student's final oral score.

The reading subtests (i.e., Vocabulary, Fluency, Mechanics and Usage and Reading for Information) for grades 2-3 are scored as correct or incorrect according to the Answer Key. The Finishing Sentences and What's Happening writing subtests are holistically scored according to a rating scale with values from 0 to 3. The Let's Write subtest, which applies only to grades 4-6, is holistically scored using a rubric that ranges from 0 to 5. As in the holistic scoring of the Oral Retell, scorers of writing subtests should participate in the reliability exercises and be proficient, literate speakers of English.

Once each of the reading and writing subtests have been scored, various scores can be produced including standard scores, competency levels and reading and writing categories. Recall that these reading and writing scores can also be combined with the LAS-Oral score to generate the overall Language Proficiency Index previously referred to in Section A. Purposes.

Test Design and Theoretical Foundation for the Test

The theoretical foundation on which the test design of the Language Assessment Scales-Oral is based, is explained in two technical reports (De Avila & Duncan, 1982; 1990). The essence of the theoretical model of language proficiency which guided the development of the oral portions of the test is captured in the authors' statement:

> The development of the LAS was based on a view of language as consisting of four linguistic aspects: phonology (phonemes, stress, rhythm and intonation), the lexicon (the "words" of the language), syntax (the rules for comprehending and producing meaningful utterances) and pragmatics (the appropriate use of language to obtain specific goals) (De Avila & Duncan, 1982, p. 8).

The authors also set forth the rationale for the selection of the test items for each of the linguistic subsystems. Briefly, they are as follows:

- the phonemic items were selected based on the fact that phonemic differences between languages exist and these differences affect meaning; preliminary research suggests a relationship between English phonemic ability and academic achievement;
- the lexical items were selected based on the fact that they are commonly acquired by native speaking students irrespective of environment and the items represent five levels of frequency;
- no specific rationale is given for the selection of the ten comprehension of syntax items (see De Avila & Duncan, 1982, p. 54); (note that these items have been dropped from the most recent version of the LAS-Oral);
- the pragmatic dimension of the LAS-Oral is optional; the authors recommend its use when students score in the "gray area bandwith"; the 10 teacher judgment items cover a range of sociolinguistic tasks underlying school success.

The authors also maintain that the LAS-Oral represents a convergent approach to language assessment. That is, while each of the above linguistic subsystems are assessed independently and weighted differently, "it is the combined total score which is of ultimate critical importance for identification and programmatic treatment" (De Avila & Duncan, 1982, p. 8). Most recently, the authors refer to the LAS family of tests as "a comprehensive set of measures designed to assess the probability of success in an American mainstream classroom" (De Avila & Duncan, 1990, p. 10).

The LAS Reading and Writing tests also represent a convergent approach to language assessment (Duncan & De Avila, 1988); total scores are derived from various content areas and scoring procedures (i.e., multiple choice and holistic). The authors maintain that the test design of these two tests was guided by the review of various state guidelines, curriculum guides, expected learning outcomes and the scope and sequence of many commonly used ESL instructional programs. The authors make no specific reference to any particular theory of reading or writing development in the *Technical Report* for these portions of the test.

Reliability and Validity

Based on the *Oral Technical Report* (De Avila & Duncan, 1990) for the most recent forms of the LAS-Oral, Forms C & D (Level I), reliability correlation coefficients ranged from .87 to .88 (Form C) and .87 to .89 (Form D). However, the listening subtests for forms C and D yielded considerably lower correlation coefficients, .48 and .38 respectfully. Recall that test scorers must undergo training in order to establish the

inter-rater reliability for the Oral Story Retell. The inter-reliability of the oral part of the LAS, which carries the most weight, is contingent upon scorers developing precision in scoring the oral samples. The authors recommend an inter-rater reliability of at least .90 for this particular subtest. No test-retest reliability data were available for Forms C & D.

The reliability coefficients for the reading and writing portions of the LAS, (Forms 1A/B, 2 A/B) are as follows. The range of correlation coefficients for Form 1A was between .76 and .91; the range for Form 1B was between .75 and .90; Form 2A reliability coefficients ranged between .75 and .88; Form 2B ranged between .73 and .88. Recall that the Finishing Sentences, What's Happening and Let's Write writing subtests are each holistically scored. Again, inter-rater reliabilities on these three subtests are contingent upon trained, skilled scoring.

Evidence for the construct validity (i.e., does the test measure what it purports to?) of the LAS-Oral is both theoretical and empirical. The LAS-Oral test design is based on the assumptions that linguistic subsystems (e.g., vocabulary, syntax, phonological) can be assessed independently of one another and subtest scores can be weighted and combined to produce an overall, oral language proficiency rating.

The *Technical Manual* for the LAS-Oral provides statistical evidence to support the validity of the oral portions of the test. Correlation coefficients among the different oral subtests for Form 1C range between .58 and .30; correlation coefficients for Form 1D range between .58 and .25. The reader should bear in mind that two subtests that assess similar or overlapping skills (e.g., phonemes and minimal pairs) should correlate moderately with each other; conversely, subtests purporting to measure more distinct skills (e.g., phonemes and vocabulary) should demonstrate a low correlation with each other.

De Avila and Duncan (1990) also offer additional construct related evidence which links oral language proficiency (as measured through the LAS) with the Comprehensive Test of Basic Skills (CTBS Form U) total reading scores. That is, the authors hypothesize an intimate relationship between level of oral language proficiency and academic achievement (i.e., reading achievement). The authors have conducted various studies and analyses which indicate that as the student's oral proficiency increases (as measured by the LAS), his/her percentile score on the CTBS reading achievement subtest also increases. However, at Level 4 the student's reading score begins to plateau or even out. From the test authors point of view, this diminishing predictive validity is to be expected since there are native speakers of English who do not consistently score at the 50th percentile on norm referenced tests.

As previously stated, the authors of the Language Assessment Scales do not adhere to any particular theory of literacy development (e.g., phonemic approach, whole language) which would influence the manner in which reading and writing are assessed. They adopt various commonly accepted discrete point and open-ended methods for assessing literacy skills such as the use of multiple choice, true and false, fill in the blank, sentence completion and short essay test formats.

Empirical evidence supporting the construct validity of the Reading and Writing portions of the LAS are also correlational and associated with CTBS Total Scores. The authors state that the correlational data for Level 1 and 2 reveal "moderate to high degrees of association within and between different versions of the LAS R/W" (De Avila & Duncan, 1988, p. 60). The authors also offer evidence that demonstrates a linear relationship between the reading and writing levels on the LAS with CTBS Total Scores. That is, percentile scores on the achievement test increase as the student's LAS reading and writing levels increase with a leveling off effect.

Critiques

Carpenter, C.D. (1994). Review of the Language Assessment Scales, Reading and Writing. In *Supplement to Eleventh Mental Measurements Yearbook*. Buros Institute.

Guyette, T.W. (1994). Review of the Language Assessment Scales, Reading and Writing. In *Supplement to Eleventh Mental Measurements Yearbook*. Buros Institute.

Haber, L. (1985). Review of Language Assessment Scales. In J.V. Mitchell (Ed.), *The Ninth Mental Measurements Yearbook Volume I*. Lincoln, NE: University of Nebraska Press.

[table of contents]

Woodcock-Muñoz Language Survey

Purpose

As is the case in the *Comprehensive Manual for the Woodcock-Muñoz Language Survey*, this test will be referred to as the "Language Survey" throughout this test handbook review. It is a recent arrival (1993) to the library of standardized language proficiency tests and a short acronym for the test has not been adopted yet by test users although the test authors sometimes refer to it as the LS-E. The review of this test is based upon the test materials and use of the test with a small selected group of examinees. At this writing, published test critiques and other sources of information about the Language Survey were not available.

Several purposes were identified for the Language Survey. It was specifically designed to measure CALP -- Cognitive AcademicLanguage Proficiency. The Language Survey yields the following language proficiency classifications:

- Level 5: Advanced English CALP
- Level 4: Fluent English CALP
- Level 3: Limited English CALP
- Level 2: Very Limited English CALP
- Level 1: Negligible English CALP

The *Comprehensive Manual for the Woodcock-Muñoz Language Survey* (referred to as the Manual

hereafter) indicates that the wide range and breadth of test items allows the Language Survey to be used with confidence for the following purposes:

- To classify a subject's English or Spanish language proficiency,
- To determine eligibility for bilingual services,
- To help teachers understand a subject's language abilities,
- To assess a subject's progress or readiness for English-only instruction,
- To provide information about program effectiveness, and/or
- To describe the language characteristics of subjects in research studies.

The Manual provides a paragraph description of each of these purposes and a rationale for how the Language Survey is able to meet each purpose.

Age and Language Groups

The Language Survey was designed to be administered individually to any student older than 48 months of age. It was not designed to be used with any particular language group and can be administered to students who speak any non-English language to assess English language proficiency. There is also a Spanish version of the test that can be administered to students from any language group to measure their proficiency with the Spanish language.

The norm group for the English form of the Language Survey included 6,359 subjects aged 24 months to 90 years of age. Subjects for the norm group were chosen randomly within a stratified sampling design. This design controlled for ten specific community and subject variables. The sampling design insured that the norm group subjects approximated the population distribution in the United States. Oversampling for small groups (for example, American Indians) was adjusted during data analyses to ensure that the norming data closely approximated the exact distribution in the U.S. population for all 10 norming variables. The ten variables used to construct the system of stratification were:

- census region,
- community size,
- sex,
- race,
- Hispanic,
- funding of college/university,
- type of college/university,
- education of adults,
- occupational status of adults, and
- occupation of adults.

In addition to the ten sampling variables used to stratify the normative sample, communities were selected with respect to four socio-economic status variables. These variables were:

- years of education,
- household income,
- labor force characteristics of adults, and
- occupation of adults.

The Language Survey norms are based on the distribution of scores at the subject's exact chronological age or grade placement resulting from a continuous-year norming procedure. Norms are based on data collected throughout the school year rather than on one or two data collection points during the year.

Administration Time

The Language Survey requires approximately 15 to 20 minutes to administer the entire battery (four subtests). More time may be required with non-responsive young children. Some students may produce a scattering of correct responses requiring that a number of items be administered. Allow a reasonable amount of time for the student to respond and then move on to the next item even though the student may have responded correctly if given unlimited time to answer.

Cost

The initial investment in either form of the Language Survey is $147 for the complete English battery including manual, test book, 25 student test forms, and computer scoring and reporting program. If additional scoring forms are needed, the cost is $21 for an extra package.

Test Administration

The Language Survey-English is administered using an easel format. An easel format is a hard-bound book designed to fold open like a tent and sit on the table without being held. The examinee can see the test items on one side and the examiner can read the answer on the other side (which can not be seen by the examinee). The easel book allows the examiner to see both the items and the correct responses and allows for protecting the scoring form from the examinee's view behind the easel. Two or three administrations allow the examiner to become comfortable with the materials and administration of the test. It can be administered within 20 minutes using the "brisk" method of administration. However, students with a long wait time will take longer as will very young and very proficient students.

In most cases, the four subtests should be administered in the order they are presented in the test easel. However, the subtests can be administered in any order and a testing session can be stopped after the administration of any test. There are four subtests (referred to as "tests" in the Manual) and they are:

- Picture Vocabulary,
- Verbal Analogies,
- Letter-Word Identification, and
- Dictation.

Basal and ceiling rules are used to guide the amount of total time devoted to testing. Test items are arranged in order of difficulty with the easiest items first and the most difficult item last. Each subtest includes items that span a range of difficulty. By not administering items that are too easy or beyond the student's capabilities, the number of items actually administered can be minimized. The basal and ceiling guidelines allow the examiner to estimate the score that would be obtained if every item were administered. Testing is conducted by complete pages in the easel book. As the examinee does not see items below the basal or above the ceiling, they are unaware that other test questions exist. This method of administration spares the examinee embarrassment when an item is too difficult.

Items

The Picture Vocabulary subtest presents the examinee with colorful drawings of things that might be considered typical and not-so-typical in the American school classroom, in homes, and in the mainstream, urban American community. The examiner asks the student to "point to the bowl of soup". The bowl of soup would be one of five food objects depicted. More difficult Picture Vocabulary items ask the examinee to provide the name for an object. Some of the more difficult items to be labeled are "pillory" , "pagoda", and "yoke".

The Verbal Analogies subtest presents printed analogies such as "sky.....blue" "tree.....". The examinee must respond with "green" to complete this analogy correctly. Items become progressively more difficult and tap the examinee's knowledge of culturally specific analogies and the examinee's ability to reason through the use of analogy as well. The examiner prompts each item by reading, "Sky is to blue as tree is to (pause)." No other prompt is allowed.

The Letter-Word Identification subtest starts with simple drawings of several different items paired with one large realistic picture. The examinee is asked to point to the cartoon picture that tells about the big, realistic drawing. Later items on this subtest require the examinee to read letters and then words. The examinee is prompted with "What is this word? Go ahead with the others (on the page). Don't go too fast."

The final subtest is Dictation. The examiner directs the student to write down responses to a range of written grammar and punctuation exercises on an answer sheet. Easy items require the student to copy a line or a circle after the examiner makes one. More difficult items require the examinee to spell words like "arrogance" and "night's" used as a possessive noun in a sentence.

Scoring

The test booklets allow for quick collection of the test data and rapid scoring of answers. Item scoring is done during test administration. The manual provides clear directions on the process for transformation of the raw scores to scores for use in interpretation. Calculation of the raw score for each subtest is done after testing is completed. In all cases the test item is scored by placing a 1 or 0 in the appropriate space on the Test Record form; 1 is correct and 0 is incorrect.

Several scores are generated for each subtest, the two language clusters (oral, and reading-writing) and the overall test battery. They include a *W* score, age equivalents, grade equivalents, relative proficiency index, and CALP level scores. Explanations for how to convert raw scores to these other scores and definitions of these scores are included in the Manual. To those familiar with the Language Assessment Scales (LAS), the CALP level scores will look familiar. The manual describes each kind of score in detail as well as how to interpret and apply the scores. The CALP level scores allow for quick program placement while the other scores provide more specific information about student strengths and weaknesses in particular areas (for example reading-writing). The Language Survey comes with computer disks that include a scoring program that will convert the raw scores for each subtest into *W* scores, grade equivalents and other types of scores. If a computer is not available, the Manual provides specific and clear instructions for converting raw scores for the clusters to CALP levels or age and grade equivalents. The charts and tables needed for manual scoring are found in Appendices A and B in the Manual.

Test Design and Theoretical Foundation for the Test

The Language Survey uses Cummins' (1984) BICS and CALP distinction as the theoretical foundation for item selection and overall test design. BICS are Basic Interpersonal Communication Skills and CALP is

Cognitive Academic Language Proficiency. The English and Spanish versions of the Language Survey are parallel forms of the same test and assess the students knowledge of "academic" language in either Spanish or English. The items included look like the type of language that would be used in classroom and advanced academic settings as opposed to the type of language used in everyday social exchange.

Reliability and Validity

There is extensive evidence in the manual for the reliability and validity of the Language Survey. Internal consistency coefficients and split half procedures were used for the four subtests, two language and the broad English ability scores. Median internal consistency coefficients were reported and ranged from .87 for the Picture Vocabulary subtest to .96 for the broad English ability score (this is the total score). The reliabilities are generally in the high .80s and low .90s for the subtests and in the mid .90s for the clusters. The reliability for this survey is highest for the oldest subjects for all scores (subtest, cluster, overall).

Evidence for content, concurrent, and construct validity is provided in the Manual. Content validity evidence included a brief description of the rationale for item selection including the use of item validity studies (documentation for this process was not included in the Manual) and expert opinion/judgement.

A number of other language proficiency tests were compared to the Language Survey to measure the extent to which scores on the Language Survey are related to these criterion measures (the other language proficiency tests). For the youngest students (pre-school and kindergarten), a *Language Rating Scale* (Houston Independent School District), the *Pre-Language Assessment Scales* (Duncan & De Avila, 1985, 1986) and the *Woodcock Language Proficiency Battery-Revised* (Woodcock, 1991) were used as the criterion measures. Correlation coefficient ranged from .64 between the Language Survey Reading-Writing Cluster score and the *Pre-Language Assessment Scales* total score to .93 between the Oral Language cluster on the Language Survey and the *Pre-Language Assessment Scales* total score. A concurrent validity study was also conducted for older students and the criterion tests were the *Language Assessment Scales* (Duncan & De Avila, 1987, 1988, 1990), the *IDEA Oral Language Proficiency Test I -- IPT i* (Ballard & Tighe, 1980), *Woodcock Language Proficiency Battery-Revised* (Woodcock, 1991), and the *Language Rating Scale* (Houston Independent School District). Correlation coefficients ranged from a low of about .50 between the Pronunciation subtest of the *Language Assessment Scales* (LAS) and the Reading-Writing cluster of the Language Survey to a high of .98 between the Broad English Ability score of the *Woodcock Language Proficiency Battery-Revised* and the Broad English Ability score for the Language Survey.

The concurrent validity studies for the Language Survey are extensive and also include comparisons between the Language Survey scores and a number of standardized aptitude and achievement tests such as the *Stanford Binet-IV* and the *Kaufman Assessment Battery for Children*. These studies are not reported here -- rather, the reader is referred to the Language Survey Manual for this information.

Construct validity was also addressed in the Manual. One part of construct validation for this test involved comparing Language Survey scores to five broad achievement measures as assessed by the *Woodcock Johnson Psycho-Educational Battery-Revised* (WJ-R)(Woodcock & Johnson, 1989). The intent of these comparisons was to delineate the relationship between academic achievement and CALP as tested by the Language Survey. Correlations between the various subtests of the Language Survey and the WLPB-R (*Woodcock Language Proficiency Battery-Revised*) ranged from .24 between the Picture Vocabulary of the Language Survey and the Basic Writing Skills of the WLPB-R to a high of .92 between the Broad English Ability score on the Language Survey and the Broad English Ability score on the WLPB-R. Multiple comparisons were made with other achievement and aptitude measures such as the *Wechsler Intelligence*

Scale for Children-Revised (WISC-R), the *Peabody Individual Achievement Test* (PIAT), the *Stanford Binet-IV*, and the *Basic Achievement Skills Individual Screener* (BASIS). The validity coefficients for these comparisons ranged from a low of .14 between the Language Survey's Letter-Word Identification subtest and the BASIS Reading subtest to a high of .73 between the Language Survey's Broad English Ability score and the PIAT Reading Recognition subtest.

Inter-correlations between the Language Survey's various subtests, cluster scores and total Broad English Ability score were offered also as evidence of construct validity. The correlations became larger with the age of the test subject which suggests that levels of oral language and reading-writing become more similar in individuals as they progress through school and mature into adulthood. Several tables of these inter-correlations are presented by the age of the students tested and most correlations are moderate. The reader is directed to the Manual for the specific results of this study.

The last type of evidence offered regarding construct validity was a table of correlations between the Language Survey Clusters and school achievement in reading, writing, mathematics, and total achievement. Grades were correlated with the Language Survey Cluster scores for all grades from K - college. Median correlations were as low as .32 between writing and Oral Language in kindergarten and as high as .91 between total achievement and the Reading-Writing Language Survey score in the first grade. Most median correlations for these comparisons were between .52 and .89.

Critiques

As the Language Survey was released by Riverside Publishing Company in 1993, published test critiques are not available yet. The reader is urged to consider the types of issues raised in the test critiques published about some of the other tests reviewed in this handbook and apply the types of issues and questions raised to the Language Survey.

[table of contents]

Summary and Test Selection Checklist

Summary

Five standardized English language proficiency tests were reviewed. Three definitions of language proficiency were provided for the readers' consideration. In addition, the legal mandate to assess English language proficiency for Title I and Title VII federally funded programs was described in some detail. The rationale for the review of the five tests selected (why these tests were chosen) was discussed. Finally, a section of the handbook provided the reader with a brief description of each test and the test publisher's address and telephone number.

For each of the tests, the same selected topics were chosen for discussion. We chose to include pragmatic information related to the administration of the test, the time it takes to administer, test cost, test items, test scoring and the purpose for each test. Additionally, information was provided on the theoretical design of the test and the technical standards for the test. Two of the tests, the BINL and the BSM (I & II) assess only oral language (speaking). The other three (IPT, LAS, & the Woodcock-Muñoz Language Survey) assess reading, writing, listening and speaking. The information on each test depended on the amount of

information provided in the manuals and test materials. Therefore, the descriptions of each test varied some related to the test documentation.

No judgement was made with regard to the "quality" of the test. These test reviews should **not** be interpreted as **test critiques**. With the exception of the Woodcock-Muñoz Language Survey, all the tests have been critiqued and a reference list of these critiques was provided at the end of each test section. The reader is urged to read the critiques about any specific test prior to purchase of the test. In this way, the reader can be assured of being an *informed consumer*. The testing industry and test products should be considered in much the same way you would consider purchasing any large and costly service or piece of equipment such as a car or a copy machine. In a report from the *National Commission on Testing and Public Policy* (1990), this statement is made, " Today, those who take and use many tests have less protection than those who buy a toy, toaster, or plane ticket" (p. 21). Carefully weigh the pros and cons before purchase and choose the product best suited to the needs of your students, teachers, programs, school and district.

To facilitate the process of choosing, we provide the reader with a checklist to assist them with the selection of an English language proficiency test. The items on the checklist parallel the 10 categories we used to describe each test. Information about the test purpose, its method of administration, the cost, the time it takes to administer the test and so forth should be considered and rated for each of the 4 language modalities (listening, speaking, reading, and writing. A column is available for an overall rating as well. This checklist can be used for rating any language proficiency test or assessment and need not be considered only for the 5 tests described in this handbook. Some tests will not assess all 4 language modalities in which case the rater will check rate only the modality columns on the checklist appropriate for that test. Use the checklist as you review the "examination" kit for any test. Most test publishers will allow a 30 day examination period free of charge. Simply contact the publisher using the telephone number included in Figure 1 of this handbook and ask them for a test examination kit.

Checklist for English Language Proficiency Test Selection

Rate each item using the following criteria:

- 5 meets need exceptionally well
- 4 meets the need
- 3 some problems but acceptable
- 2 problems detract form overall utility
- 1 major problems with the instrument
- NA does not apply to this instrument

Some tests assess all four language modalities (reading, writing, listening, speaking) and some do not. Columns have been provided for each of the language proficiency modalities so that the instrument can be rated for each of these areas if desired. One column is available for an overall rating

topic/item	**rating**	**speaking**	**listening**	**reading**	**writing**	**comments…**
The purpose of the test is clear and meets our definition of language proficiency.						

The test administration directions are specific and understandable.						
The test administration guidelines can be used by teachers and teacher assistants.						
The cost of the test is OK.						
The test items seem to assess our definition of language proficiency.						
The test items can be used to help the teacher design appropriate instruction for the individual student.						
The test scores are useful for program placement.						
The test scores can be used for evaluating the program.						
There are multiple forms of the test so that it can be used for Pre/Post testing to evaluate the program.						
The amount of time it takes to administer the test is OK.						
The type of administration format (individual or group) is acceptable.						
The theoretical foundation of the test fits our definition of English language proficiency.						
There is adequate explanation of the theoretical foundation for the test.						
The test offers adequate evidence of reliability.						
The type of reliability evidence provided fits the design of the test.						
The type of validity evidence is adequate.						
The type of validity evidence provided makes sense with regard to the						

purpose(s) for which the test was designed.						
Major problems with the test, as identified by test critiques, do not compromise the positive elements of the test.						

[table of contents]

References

American Psychological Association. (1985). Standards for Educational and Psychological Testing. Washington, DC: American Psychological Association.

Amori, B. A., Dalton, E.F. , & Tighe, P.L. (1992). IPT 1 Reading & Writing, Grades 2-3, Form 1A, English. Brea, CA: Ballard & Tighe, Publishers.

Anastasi, A. (1988). Psychological Testing (sixth edition). New York, NY: Macmillan Publishing Company.

Ballard, W.S., Tighe, P.L., & Dalton, E. F. (1979, 1982, 1984, & 1991). Examiner's Manual IPT I, Oral Grades K-6, Forms A, B, C, and D English. Brea, CA: Ballard & Tighe, Publishers.

Ballard, W.S., Tighe, P.L., & Dalton, E. F. (1979, 1982, 1984, & 1991). Technical Manual IPT I, Oral Grades K-6, Forms C and D English. Brea, CA: Ballard & Tighe, Publishers.

Burt, M.K., Dulay, H.C., & Hernández-Chávez, E., (1976). Bilingual Syntax Measure I, Technical Handbook. San Antonio, TX: Harcourt, Brace, Jovanovich, Inc.

Burt, M.K., Dulay, H.C., Hernández-Chávez, E., & Taleporos, E. (1980). Bilingual Syntax Measure II, Technical Handbook. San Antonio, TX: Harcourt, Brace, Jovanovich, Inc.

Canale, M. (1984). On some theoretical frameworks for language proficiency. In C. Rivera (Ed.), Language proficiency and academic achievement. Avon, England: Multilingual Matters Ltd.

Canales, J. A. (1994). Linking Language Assessment to Classroom Practices. In R. Rodriguez, N. Ramos, & J. A. Ruiz-Escalante (Eds.) Compendium of Readings in Bilingual Education: Issues and Practices. Austin, TX: Texas Association for Bilingual Education.

CHECpoint Systems, Inc. (1987). Basic Inventory of Natural Language Authentic Language Testing Technical Report. San Bernadino, CA: CHECpoint Systems, Inc.

Council of Chief State School Officers (1992). Recommendations for Improving the Assessment and

Monitoring of Students with Limited English Proficiency. Alexandria, VA: Council of Chief State School Officers, Weber Design.

CTB MacMillan McGraw-Hill (1991). LAS Preview Materials: Because Every Child Deserves to Understand and Be Understood. Monterey, CA: CTB MacMillan McGraw -Hill.

Cummins, J. (1984). Wanted: A theoretical framework for relating language proficiency to academic achievement among bilingual students. In C. Rivera (Ed.), Language proficiency and academic achievement. Avon, England: Multilingual Matters Ltd.

Dalton, E. F. (1979, 1982, 1991). IPT Oral Grades K-6 Technical Manual, IDEA Oral Language Proficiency Test Forms C and D English. Brea, CA: Ballard & Tighe, Publishers.

Dalton, E. F. & Barrett, T.J. (1992). Technical Manual IPT 1 & 2, Reading and Writing, Grades 2-6, Forms 1A and 2A English. Brea, CA: Ballard & Tighe, Publishers.

De Avila, E.A. & Duncan, S. E. (1990). LAS, Language Assessment Scales, Oral Technical Report, English, Forms 1C, 1D, 2C, 2D, Spanish, Forms 1B, 2B. Monterey, CA: CTB MacMillan McGraw-Hill.

De Avila, E.A. & Duncan, S. E. (1981, 1982). A Convergent Approach to Oral Language Assessment: Theoretical and Technical Specifications on the Language Assessment Scales (LAS), Form A. Monterey, CA: CTB McGraw-Hill.

De Avila, E.A. & Duncan, S. E. (1987, 1988, 1989, 1990). LAS, Language Assessment Scales, Oral Administration Manual, English, Forms 2C and 2D. Monterey, CA: CTB MacMillan McGraw-Hill.

Duncan, S.E. & De Avila, E.A. (1988). Examiner's Manual: Language Assessment Scales Reading/Writing (LAS R/W). Monterey, CA: CTB /McGraw Hill.

Durán, R.P. (1988). Validity and Language Skills Assessment: Non-English Background Students. In H. Wainer & H.I. Braun (Eds). Test Validity. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

National Commission on Testing and Public Policy. (1990). From Gatekeeper to Gateway: Transforming Testing in America. Chestnut Hill, MA: National Commission on Testing and Public Policy.

Oller, J.W. Jr. & Damico, J.S. (1991). Theoretical considerations in the assessment of LEP students. In E. Hamayan & J.S. Damico (Eds.), Limiting bias in the assessment of bilingual students. Austin: Pro-ed publications.

Rivera, C. (1995). How can we ensure equity in statewide assessment programs? Unpublished document. Evaluation Assistance Center-East, George Washington University, Arlington, VA.

Roos, P. (1995). Rights of limited English proficient students under Federal Law -- A guide for school administrators. Unpublished paper presented at Weber State University, Success for all Students Conference, Ogden, UT.

Spolsky, B. (1984). The uses of language tests: An ethical envoi. In C. Rivera (Ed.), Placement procedures in bilingual education: Education and policy issues. Avon, England: Multilingual Matters Ltd.

Ulibarri, D., Spencer, M., & Rivas, G. (1981). Language proficiency and academic achievement: A study of language proficiency tests and their relationship to school ratings as predictors of academic achievement. NABE Journal, Vol. V, No. 3, Spring.

Valdés, G. and Figueroa, R. (1994). Bilingualism and testing A special case of bias. Norwood, NJ: Ablex Publishing Corporation.

Wheeler, P. & Haertel, G.D. (1993). Resource Handbook on Performance Assessment and Measurement: A Tool for Students, Practitioners, and Policymakers. Berkeley, CA: The Owl Press.

Woodcock, R. W. & Muñoz-Sandoval, A.F. (1993). Woodcock-Muñoz Language Survey Comprehensive Manual. Chicago, IL: Riverside Publishing Company.

[table of contents]
