

How Well Do Commonly Used Language Instruments Measure English Oral-Language Proficiency?

Lisa Pray
Utah State University

Abstract

This research examines three tests commonly used to assess the English oral-language proficiency of students who are English language learners (ELLs): the Language Assessment Scales—Oral, the Woodcock-Muñoz Language Survey, and the IDEA Proficiency Test. These tests were given to native English-speaking non-Hispanic White and Hispanic students from varied socioeconomic levels. Since these tests use native-language proficiency as the standard by which responses are evaluated, it is reasonable to expect native English speakers to perform extremely well on these instruments. The extent to which the native speakers of the language do not perform well on these instruments calls their validity into question. Findings indicated that none of the native English-speaking children who took the Woodcock-Muñoz Language Survey scored in the “fluent” or “advanced fluent” English ability. One hundred percent of the students scored in the “fluent English speaking” range of the Language Assessment Scales—Oral, and 87% of the students scored in the “fluent English speaking” range of the IDEA Proficiency Test.

Introduction

This study examines the validity of commonly used language assessments and determines how accurately the English versions of these assessments measure English proficiency by testing native English speakers on the English version of the assessments. Because these tests were developed using knowledge of English as the criterion, and because normally developing monolingual English speakers should have the required knowledge of English to speak and understand, their performance on these tests measures the degree

to which the tests have construct validity. If native English speakers do not receive scores in the “native speaking ability” range on the assessment, the chances of an English language learner (ELL) achieving a score that accurately reflects his or her English proficiency is diminished. No matter what theory of language ability is used, a test that identifies native speakers of English as limited or even non-speakers of English has failed to accurately measure oral English-language proficiency.

This research continues to probe the “non–non”¹ crisis first investigated by MacSwan (2001), Ortiz (2001), Mahoney (2001), and Fey (2001). All four researchers analyzed the scores of native Spanish-speaking students who were assessed using the Spanish version of the Language Assessment Scales—Oral (LAS–O), the IDEA Proficiency Test (IPT), and the Woodcock–Muñoz Language Survey (WMLS). Fey analyzed the WMLS, Mahoney the IPT, and Ortiz the LAS–O. MacSwan compared these results with a natural language sample taken from the children and found that large proportions of students were designated as non-speakers of their native language although their natural language sample revealed few errors. The present research is similar in that it assesses the English-speaking ability of children using English versions of the tests. A native speaker of English receiving a result of non-fluent would be essentially the same as obtaining a false negative.

Critical analysis of language assessment is necessary to inform educators who place students in language support programs according to their level of English-language proficiency, and those who refer ELL students into special education. Language assessment scores are heavily weighted in the decision to place students in these kinds of programs. A child has an increased likelihood of being referred to special education if he or she is listed as non-fluent in his or her native and second language. Thus, these assessments are high-stakes measures that play a major role in a child’s success in school.

Literature Review

Research indicates that teachers inappropriately use oral-language proficiency as an indication of the child’s overall academic performance (Limbos & Geva, 2001). ELL students, especially those categorized as non-fluent in both their native language and English by commonly used language assessments, are very likely to be placed in special education (Artiles, Rueda, Salazar, & Higaeda, 2000). Thus, language assessments that do not accurately reflect a student’s language ability may contribute to the misplacement of ELL students into special education and contribute to the limitation or exclusion of future opportunities that would otherwise be available to these students (Zehler, Hostock, Fleischman, & Greniuk, 1994).

ELL students, by virtue of their developing English-speaking skills alone, are no more “at risk” for language disorders, impairments, or disabilities than

are their native English-speaking peers. The vast majority of ELL children with limited English proficiency are normal language learners. It is not unusual for children who are learning a second language to be temporarily delayed in both their native and second languages (Ochoa, Galarza, & González, 1996). Assessing language proficiency is difficult because language proficiency is not a static state but instead a state of constant fluctuation (Ochoa et al., 1996).

Many language tests follow a psychological rather than linguistic theoretical framework, evidenced by the use of a single modality (such as a paper-and-pencil test that ignores spoken and oral comprehension). This effort at defining language proficiency using discrete psychometric properties is flawed and fails to assess language used in naturally occurring contexts and has little to do with real-world communication (Sommers, 1989; Geisinger & Carlson, 1992; Hernández, 1994). As such, traditional language assessments are poor predictors of language and communication abilities because they focus instead on ensuring replicable results based on a standardized, statistically norm-referenced model and ignore the cognitive nature of bilingualism (Hernández, 1994; Figuera, 1991). Traditional (psychometric) assessment practices often do not assess a student's ability to use his or her language to solve real-world problems and are therefore incomplete (i.e., they do not assess all that should be assessed) or inappropriate (i.e., they do not assess what is really needed or relevant to the student) and fail to provide a multifaceted view of a student's strengths and needs (Todd, Fiske, & Dopico, 1994). Thus, because language assessments often have low reliability and validity, reliance on the results of these assessments has increased the likelihood of misidentification of ELL students as candidates for placement into special education (Artiles & Trent, 1994). Despite these problems in measurement, language proficiency must be at the center of the valid assessment, for without it the child has not been fully evaluated. Neither dominance nor proficiency in a language can be automatically assumed from observation alone (Hernández, 1994).

The data in this study will be analyzed from a linguistic framework when referring to "oral-language ability," "oral-language competence," or the "ability to speak or understand the language." This construct reflects a grammatical system that consists of the rules and principles governing syntax (word order), morphology (principles of word formation), and phonology (pronunciation), and that interface with principles of discourse, pragmatics, and semantic interpretation, as well as the ability to recombine language forms to make unique utterances. Using a linguistic framework, oral-language assessments must measure the essential elements of knowing the language and should measure competency in all relevant English-language skills and no other skills.

Brown (1973) cites three necessary criteria for knowing a language. First, speakers must be productive in the use of language, that is, able to produce

new utterances and recombine forms to express concepts they have never heard before. Second, language competence requires the ability to represent ideas, events, and objects symbolically. Third, speakers should be able to communicate on an abstract level, not just the immediate context. To do this, children must obtain basic mastery over semantics, morphology, and syntax—the essential components of language (Gleason, 1997). Children acquire the majority of the morphological and syntactic rules of their native language by the time they begin school. Children use and understand increasingly complex grammatical structures until acquisition is considered essentially complete a few years later (Gleason).

Most language assessments use measures of academic achievement to determine language proficiency. Academic achievement denotes the content and skill specific to the school environments. It is assumed that English as a second language acquisition will precede the achievement of parity with English-background children on English measures of academic achievement. Makuta, Butler, and Witt (2000) found that student performance on language assessments are often affected by students' socioeconomic status (SES). This study was conducted to investigate the rate of second language acquisition among ELL students. Given these results, student SES must be accounted for when analyzing language assessments.

No matter what theory of language proficiency is used, if native speakers are labeled not proficient according to a test that purports to measure language proficiency, then the test fails to be descriptive regarding language ability and has become prescriptive. Prescriptivism is the view that one variety of a language has an inherently higher value than another and it should be imposed on all speakers of the language (especially in the areas of grammar, vocabulary, and pronunciation). Descriptive grammar, by contrast, is a true model of a speaker's basic linguistic ability, not the evaluation of a language variation from a prescribed set of rules (Crystal, 1997; Fromkin & Rodman, 1993). Although adhering to a coherent construction of a language is important, to do so by arbitrarily privileging the use of one form of language over another in an assessment deprives the educational community of a realistic evaluation of a student's ability to speak and understand a language. Thus, if the tests do not classify native English speakers as having native English-language ability, then the assessment prescribes what ability native speakers should have, instead of describing a child's linguistic ability. Tests that identify native speakers as having limited English proficiency or being non-proficient cannot be trusted to identify a lack of English-language ability in second-language speakers.

Method

This research is intended to test the validity of the language instruments used to assess ELL students by testing native English speakers on the English version of the language tests. Theoretically, native English speakers should receive scores in the range of native speaking ability. It is important to determine how well children can understand oral instruction and articulate meaning using oral language. Although literacy in a second language is important, the focus of this research is oral-language ability. If native English speakers have wide variability in the scores achieved on these English versions of the assessments, it will be indicative of problems with the way language tests are constructed.

The specific research questions to be posed regarding the native speakers of English are the following:

1. How does each testing instrument identify the English-speaking ability of native English-speaking children?
2. What are the effects of SES and ethnic background on the language proficiency scores for each testing instrument?

To answer these questions, data were obtained from tests given to native English speakers on the oral-language version of each English-language assessment.

Assessment Measures

The tests chosen for the present study were the oral versions of the LAS—O, the WMLS, and the IPT. Nationwide, the LAS and IPT are among the instruments most frequently used by school districts to ascertain whether ELLs qualify to receive (or should exit from) bilingual or English as a Second Language (ESL) classroom instructional settings (Ochoa et al., 1996). For example, when pretest and posttest results are compared, the IPT claims to measure gains in language skills to be used in the evaluation of children for specialized instructional programs (Ballard, Tighe, & Dalton, 1991). Certainly, the results of any or all of these tests are considered by the teacher or special education referral committee when determining the ability of the student and, as such, may trigger a decision to refer a student for placement into special education.

Demographics of the Participating District

The school district that supported this research consists of 32 schools located in diverse socioeconomic pockets of a large urban city in the southwestern United States. Approximately 25,000 students attend school in this district from kindergarten to eighth grade. The district identifies the ethnicity of these children as approximately 55% non-Hispanic White, 33% Hispanic, 6% Black, 3% Native American, and 3% Asian. Thirty-five different

languages are represented in this district, and approximately 4,700 students (roughly 20% of the student population) are classified as limited English proficient, representing an increase of 87% in the last 5 years and 538% in the last 10 years. The most common native language spoken is English, the second most common is Spanish, and the distant third is Serbo-Croatian. Close to 50% of district students are enrolled in the free and reduced lunch program, with eligibility ranging from 9% to 100% in the various schools.

After the district formally agreed to participate, the district released potential names of the students and their teachers. Data from the school district were screened to include students with the following characteristics as designated by the district: (a) ethnic origin of Hispanic or non-Hispanic White; (b) currently enrolled in general education; (c) enrolled in the fourth or fifth grades; and (d) not enrolled in a “gifted” program. Because a brief interview before testing revealed that three children spoke Spanish as their first language and two children were in the district’s gifted program, five children were not invited to participate in the study.

Research Design

The analysis of variance (ANOVA) design consisted of 10 students in each cell. The cells consist of students from low-SES and middle- to high-SES backgrounds and students from non-Hispanic White and Hispanic backgrounds. A total of 40 participants were recruited from a local elementary school district. The subjects were students in the fourth and fifth grades who speak English as their native language. Twenty students were non-Hispanic White, and 20 were of Hispanic descent as determined by the district. Twenty-two students were female, and 18 were male. The ages of the students ranged from 8 to 12 years old (one child was 8 and one child was 12), with a mean age of 9.8. Twenty-eight students were in fourth grade, and 12 students were in fifth grade. Twenty students took part in a free or reduced lunch program, and 20 students paid full price for school lunch. Status in the lunch program was used as a proxy variable to determine SES. Those students who were given a free or reduced lunch were considered “low SES” and those who paid full price for lunches were considered “middle or upper SES.”

The participants were given the assessments one at a time with a week between each assessment. They were first given the IPT, then the LAS, then the WMLS. The children were told that they were taking a test that was designed to give teachers an understanding about how well ELL students knew English. They were being asked to take the tests to help understand how well the tests actually tested English.

All tests were administered by one person, the researcher of this study. For that reason, between-subjects interrater reliability was not a variable to consider regarding analysis. Generally, each assessment was administered

according to the publisher’s guidelines and the *Standards for Educational and Psychological Testing* (American Educational Research Association, 1999). The researcher read the manual describing the assessment procedures and practiced administering the tests to children not in the study prior to administering the tests to the study participants. The researcher has received specific training in administering the LAS—O and the WMLS. The training and experience administering and scoring the LAS—O had been undertaken in the context of her experience as an elementary school ESL teacher, and the training and experience in administering the WMLS was gained as a research assistant in a separate, unrelated research study. No training was received for the administration of the IPT, but the researcher carefully read and followed the directions for test administration provided by the publisher.

Findings

The language assessment scores were analyzed using frequency and *t*-tests and analysis of variance (ANOVA) in an effort to compare the mean assessment scores and the independent variable and correlations.

Frequency Analyses

To demonstrate the distribution of student test scores on each assessment, frequencies of student test scores were analyzed. This simple analysis was conducted to determine in general terms how well the tests measured the English oral-language proficiency of these native English-speaking children.

The IPT results indicate that according to the test scores, 3% were classified as non-English speaking, 12% were classified as limited English speaking, and 85% were classified as fluent English speaking. Table 1 depicts the frequency distribution of the IPT scores.

The LAS—O frequency results indicate that no student was classified as a non-English speaker. Table 2 depicts the frequency distribution of the LAS—O scores.

Table 1
Frequency Distribution of Student Scores on the IDEA Proficiency Test (IPT)

IPT level	Distribution
NES (non-English speaking)	1 (3%)
LES (limited English speaking)	5 (12%)
FES (fluent English speaking)	34 (85%)

Table 2

Frequency Distribution of Student Scores on the Language Assessment Scales—Oral (LAS—O)

LAS—O category	Distribution
Non-English speaking	0
Limited English speaking	0
Fluent English speaking	40 (100%)

Table 3

Frequency Distribution of Student Scores on the Woodcock-Muñoz Language Survey (WMLS)

WMLS level	Distribution
Negligible English	4 (10%)
Very limited English	20 (50%)
Limited English	16 (40%)
Fluent English	0
Advanced English	0

The WMLS results indicate that according to the test scores, no child scored in the “fluent speaking ability” or “advanced English speaking ability” range. In contrast, 10% were classified as negligible English speakers, 50% were classified as very limited English speakers, and 40% were classified as limited English speaking. Table 3 depicts the frequency distribution of the WMLS scores.

The frequency analysis for each instrument shows great variation in the scores that native English-speaking children received on the tests. None of the native English-speaking children were considered fluent or advanced fluent according to the WMLS. According to the LAS, 100% of the students were considered fluent, and according to the IPT, 85% of the children were considered fluent speakers of their native language.

Independent Samples *t*-test

Independent samples *t*-tests were conducted for the IPT and WMLS to individually evaluate differences in scores of non-Hispanic White and

Hispanic students or students from high- and medium-SES and low-SES backgrounds. The LAS scores were excluded from the analyses because 100% of the students were considered fluent speakers of English. The *t*-test results are reported for each language instrument.

The results for the IPT analysis found no statistically significant difference in the test scores between Hispanic and non-Hispanic White students. The next test conducted was to determine the differences in scores of children from medium- to high-SES and low-SES backgrounds. Using the statistic for equal variances assumed, the test was just short of being statistically significant, $t(1.955) = 4.85, p = .058$.

The results for the WMLS indicated that there were no statistically significant differences in the test scores between Hispanic and non-Hispanic White students, nor were there statistically significant differences between student scores from children of medium- or upper-SES and low-SES backgrounds.

Discussion

This section will respond to the research questions, discuss each of the language assessments, and then consider the implications of this research in the context of other “non–non” research. The discussion is layered over the important premise that by third or fourth grade, most children should generally know and be proficient speakers of their native language, and that by this age, children develop a basic mastery of semantics, morphology, and syntax, as well as the ability to recombine language forms to make unique utterances. Thus, when measuring oral-language ability, oral-language assessments should measure these skills and nothing else. Also of importance, none of the children in this study exhibited language impairments or other conditions interfering with the normal development of language, and therefore ought to score in the proficient or fluent range of each assessment.

The WMLS scores indicate that none of the native English-speaking children in this study were fluent in their English ability, the IPT categorized a majority of the students as fluent in English, and the LAS categorized all the children as fluent in English. Given what is known about language competence in children, the WMLS scores were startling and refute the assertion that the WMLS measures oral-language ability, diminishing the test’s construct validity as an assessment of oral-language ability.

The WMLS measure of oral-language proficiency is based on two language subsections, vocabulary and verbal analogies, neither of which should be considered an essential criterion for oral-language proficiency. To get an age equivalent score for 10-year-olds, students had to look at pictures to identify items, such as *candelabra*, *pendulum*, *vise*, *tourniquet*, *panning for gold*. Vocabulary learning continues indefinitely, which is why tests of

vocabulary knowledge cannot be used to properly assess language knowledge. Nobody would ever know all vocabulary; everyone has different yet equally elaborate vocabularies, and a presumption of a value of one vocabulary word over another is a signal that the instrument is not a valid measure of oral-language ability.

Observations made during the testing procedure revealed that many students had particular difficulty identifying a tourniquet, printing press, or men panning for gold because of confusing prompts in the test booklet. Those pictures depict historical views of the world. In particular, the tourniquet picture looks to be an old-fashioned bandage, or a piece of cloth wrapped around a man's forearm. The only clue that the picture was to represent a tourniquet is a stick wrapped on the outside of the cloth. Even if a student had some real-life experience with a tourniquet, the picture was unlikely to represent anything a child may have seen, as modern tourniquets are strings of rubber. No student tested was able to respond correctly to that prompt. Many students were unable to correctly identify "easier" pictures such as a grasshopper. In the southwestern desert, grasshoppers are relatively rare. However, there is a very high population of crickets, which look very much like small grasshoppers. Unfortunately, if a child responded to the grasshopper prompt with *cricket*, the answer was deemed incorrect by the test makers. The test giver is not allowed to direct the student by saying, "What else would you call it?" as is frequently advised on other test items on the WMLS.

A student would need a broad historical knowledge of early machinery to answer the question about the printing press. The printing press depicted in the picture is vintage 19th century. Other pictures, such as that of men panning for gold, seemed to be focused on a specific point in American history, in this case, the gold rush. Additional pictures of the printing press seemed to be derived from early points of history as well. When this test was administered to native Spanish-speaking students in a separate research study (Fey, 2001), students had extreme difficulty correctly identifying those pictures. Fey believed the Spanish-speaking students' difficulty was due to their lack of prerequisite historical knowledge of the gold rush or printing presses used in the late 19th century. Logically, one could assume that native English-speaking students would fare better on those items because they may have had previous school experience that corresponded with these items. In fact, in the present study, both Hispanics and non-Hispanic Whites often failed to respond correctly to those questions. Students would need to possess specific experiences to correctly identify such pictures. For the most part, these prompts are specific to American cultural knowledge and therefore to a particular domain of academic achievement. As stated by MacSwan and Rolstad (2003), "If we define language proficiency in such a way as to include this sort of highly particular cultural knowledge, what should be regarded as a simple cultural difference suddenly becomes a linguistic dividing line which enormously privileges those with more socially valued cultural capital in hand" (p. 329).

The WMLS states that it measures cognitive academic language proficiency (CALP). CALP was developed by Cummins (1980, 2000) in an effort to provide evidence against prematurely exiting ELL students from a bilingual or other language-supported program to a mainstream classroom. His model of second language acquisition is widely accepted in the educational community and frequently cited in education literature. Many teachers find it to be a useful theory when planning instruction for ELL students. CALP, language claimed to be cognitively demanding and context reduced, requires a more sophisticated use of a second language for more complex and abstract thought than conversational skills. Hence, CALP, according to Cummins (1980, 2000), is highly related to academic achievement. Although most in the mainstream educational community adhere to this psychological model, there is growing criticism of this theory by linguists and other experts in education who claim that CALP imposes a prescriptivist or deficit view of “nonacademic” language use (MacSwan, 2000; MacSwan & Rolstad, 2003). This is especially relevant when CALP is used as a theoretical construct of language proficiency in an oral-language assessment. When assessments are aligned with a prescriptivist view of language ability, many proficient or competent speakers of the language will not fare well on the test for reasons unrelated to language competence. If used as a theoretical foundation for a language assessment, CALP may confound second-language proficiency with academic achievement, thereby making it difficult to determine whether a child is having difficulty in a classroom because of an inability to understand the language, or because he or she is unable to perform the academic tasks expected in the classroom environment.

Thus, the CALP theoretical framework is a problematic construct for the assessment of language-minority students for two reasons. First, by definition, it presumes a significant link between language and academic ability, confounding language knowledge with cognitive achievement and making it difficult to tease out academic achievement from normal developmental delays in second language acquisition. Second, the prescriptivist nature of the CALP, especially when CALP is assessed using measures of vocabulary, leads to test bias. If students are asked to identify vocabulary about which they have little prior knowledge or experience, the test score will not adequately reflect their language ability.

The WMLS views academic achievement in reading and writing as indicators of CALP. In fact, according to the test manual, the WMLS test designers chose measures of IQ and academic achievement for their correlational analysis, assessments that the test makers say “measure similar abilities.” Those comparative tests include three varieties of the Wechsler Intelligence Tests (one for preschool children, one for school-age children, and one for adults) and two varieties of achievement tests (the Differential Ability Scales and the Wide Range Achievement Test). The only test related

Table 4

Assessments Correlated With the Woodcock-Muñoz Language Survey to Establish Concurrent Validity

Test	Type
Wechsler Preschool and Primary Scale of Intelligence—Revised	Intelligence
Differential Ability Scales	Cognitive abilities, achievement
Wechsler Intelligence Scale for Children (3rd ed.)	Intelligence
Wechsler Adult Intelligence Scale	Intelligence
Oral and Written Language Scales (OWLS)	Listening, oral, and written expression
Wide Range Achievement Test	Academic skills

to language ability is the Oral and Written Language Scales (OWLS), which measures “listening comprehension, oral expression, written expression, and oral comprehension.” A summary of the tests and test types can be found in Table 4.

Based on the analysis performed by the publisher, the developers of the WMLS suggests that the strong correlation between the verbal portion of the Wechsler Intelligence Tests and the WMLS is evidence that the WMLS oral-language test is cognitive and verbal in nature. The correlation between the WMLS items intended to measure oral language and the OWLS (the only language ability test included) was far weaker than the relationship between the intelligence or academic skills measures. The question is whether these “cognitive and verbal” skills are actually a construct of oral-language proficiency.

Although purporting to measure oral-language ability, the WMLS omits measuring important linguistic functions such as morphology, syntax, and semantics. Instead, it measures knowledge of picture vocabulary, verbal analogies, word identification (reading), and dictation (writing), constructs typically associated with academic achievement measures. Makers of the WMLS find that “measures of CALP are more relevant for an assessment of language proficiency in academic settings than are measures of surface fluency or BICS” (Woodcock & Muñoz-Sandoval, 2001, p. 42). Thus, the WMLS equates oral-language proficiency with CALP. By measuring oral-language ability using these academic achievement measures, the WMLS has

pronounced that if a child does not have CALP, the child does not have oral-language ability in an academic setting. Although it is important to assess a child's reading and writing ability, to do so as a measure of oral-language proficiency does not serve the child being tested or the educational community. Knowledge of oral language is independent of literacy and ought to be separated out, especially when educational placement decisions are to be made based on the results of the assessment.

The alarming results from the frequency analysis that none of the 40 native English-speaking children tested in this study were considered fluent in their native language constitute solid evidence that the WMLS is not an accurate measure of oral-language proficiency. Without a linguistically informed construct (such as syntax, morphology, and semantics) and with tests where questions are weighted heavily on specific vocabulary knowledge, a child such as Genie² could be classified as language proficient, while millions of normally linguistically developed children would be classified as non-proficient.

The creators of the IPT address English oral linguistic ability (both comprehension and production) by measuring phonological, grammatical, syntactical, and semantic structures of English. However, upon review of the test, the items demand academic skills similar to the CALP construction of language proficiency. Specifically, the test requires that the child respond to many items using a complete sentence. For example, when a child is asked, "Do you drive your teacher's car?" it is not enough to answer "no." The child must say, "No, I don't" or "No, I do not." Many children being tested for this study were amused by the question and forgot to answer using the institutional requirement of a complete sentence. While it is true that the test administrator can prompt the child to answer in a complete sentence, this often constrains the child's natural utterance. In another example, children are asked, "What do you do during lunchtime after you eat?" Instead of the natural inclination of the child to say "talk to my friends" or "play soccer" or "play on the monkey bars with my friends," the child is reminded to speak in a complete sentence and to produce a far more artificial utterance, such as, "I like to play on the monkey bars with my friends during lunchtime recess." As MacSwan and Rolstad (2003) ask, "We must ask whether the ability to recognize or produce a complete sentence on demand ought to factor into a native speaker's knowledge of language" (p. 64).

In addition, many students struggled with the story retelling portions of the IPT. At the end of almost every level of the assessment, the child is asked to listen to a short story that is to be read only one time. After the story is read, the student is either asked questions about the story. The first story is the following: "John is going to the airport with his parents. They are taking a trip to visit his grandmother who lives in the city. John and his parents will fly on a plane." The students are then asked four questions asking for certain details

about the story. The questions appeared to be quizzing the children more on their memory of these details than on their English comprehension. For example, one question is: "Where does his grandmother live?" Could a native English-speaking adult remember the answer to this question without referring back to the story? Sometimes the child was asked to retell the story, at which time the person assessing the child was asked to determine if the child correctly explained prescribed elements of the story. For example, on one item, the child had to explain that the main characters of the story are brothers. When taking the test, many children appeared to be confused about the story and the related questions, especially because this story appears at the end of the first level the child undertook. The questions leading up to the story are of a different format, requiring the children to mentally adjust to the new format of questions.

The LAS—O was more mechanized than any of the other assessments, as most of the test concerned listening that was done using prerecorded tapes supplied by the test maker. The first portion of the assessment tests vocabulary and asks the students to name various items found in and around the classroom and playground, then asks students to identify what a person is doing (such as "eating" or "climbing") in a picture. Some of the pictures associated with the action words tended to be confusing to the student, but most students were able to identify the actions without missing more than one or two items. The vocabulary portion of the test is weighted in such a way that missing one or two words does not significantly affect the overall score. The next portion of the test asks the students to listen to a conversation between a girl and a man who serves lunch in the cafeteria. The student is then asked 10 yes–no questions about the conversation. Many children were confused by some of the questions. For example, in the audio story the girl, Liza, selected a brownie for dessert from a choice of brownies, pudding, or ice cream. The question about that portion of the story was, "Were the dessert choices brownies, pudding, or ice cream?" The voice on the tape slightly inflects the word "or" so many children answered "brownie" instead of the correct answer, "yes." Fortunately, even if a student missed one or two questions on this portion of the test, it did not affect his or her overall score because of the way the scores were weighted.

The final LAS—O score is weighted heavily on the story-retelling portion of the test. This portion is scored by comparing the child's utterance to a rubric provided by the publisher. A difference of one point on this portion of the test can change the child's overall language proficiency score from being a limited speaker of the language to a fluent speaker of the language. A surprising and troubling aspect of analyzing the data for this project was the scoring of the LAS story-retelling portion of the assessment. That portion of the assessment was administered by first having the child listen to a tape-recorded story called "Angelina's Uncle." The child was then recorded retelling

the story. Later, the tapes were transcribed by an assistant to the researcher and “second-listened” by the researcher to ensure the accuracy of the transcription. The scoring on the story-retelling section was completed by comparing the child’s utterance to a rubric to establish a test score that ranges from 1 to 5. Initially, the researcher and an assistant graded the utterances according to the rubric. The scores were compared, and frequent discrepancies were observed between the scores. These discrepancies were sometimes by as much as 2 points (for example, one person scored a “2” and the other scored a “4”), which made a significant difference in the overall score of language ability. Comments made by the assistant suggested that the rubric was confusing, and there appeared to be an overlap in the distinctions between levels. It was then that a decision was made to have the story-retelling section scored by an independent company authorized by the publisher to grade the LAS—O to avoid the appearance of bias in the scoring of the assessment and provide accurate and consistent results.

There were significant discrepancies between the LAS—O scores provided by the independent scoring company that scores the test for profit, the researcher, and the research assistant. Of the 40 samples, the researcher and research assistant were in agreement with only 8 out of the 40 samples. On 4 of the samples, there was more than a 2-point discrepancy. Although the scores supplied by the independent scoring company were used in the analysis for the LAS—O, it should be noted that only 5 of the 40 story-retelling samples were scored with total agreement among the researcher, the research assistant, and the scoring company. Twenty-one samples were scored with a 1-point difference among the scores provided by the scoring company, the researcher, and the research assistant. Fourteen samples had a 2-point difference among the scores provided by the scoring company, the researcher’s scores, and the research assistant’s scores. Depending on which scores were used, 100% of the students were considered fluent in English, as scored by the independent scoring company; 93% of the students were considered English fluent, and 7% were considered limited English fluent, as scored by the researcher; or 65% of the students were considered fluent and 35% considered limited English fluent, as scored by the research assistant.

A one-way repeated measures ANOVA was conducted with the factor being the LAS scores for each rater. The results for the ANOVA indicated a significant difference between the scores of entity that scored the sample (Wilks’ $\Lambda = .56$, $F(2, 38) = 14.94$, $p \leq .001$). A follow-up pairwise comparison indicated significant affect between each pair of scores in the $p \leq .001$ range.

This causes some concern about the interrater reliability of the LAS and its practical use in a public school setting. For purposes of consistency, use of a company authorized by the test publisher to objectively score tests should be the best method to evaluate this portion of the test. However,

schools often do not have a budget to cover the costs (\$4.25 per test, plus shipping and handling) to have each test independently evaluated. Consequently, teachers or test administrators (often teachers' aides) are left to this task, and there can be great variation in the way the tests are scored and the children are labeled. The discrepancies between scores bear notation and merit future study on a wider scale.

Comparison with MacSwan's Non-Non study

MacSwan (2001), Ortiz (2001), Fey (2001), and Mahoney (2001) analyzed the scores of approximately 135 native Spanish-speaking children, ages 6 to 8, who were assessed using the Spanish version of the LAS—O (Ortiz), the IPT (Mahoney), and the WMLS (Fey). MacSwan compared these results with a natural language sample taken from the children. The sample consisted of a child telling a story from a wordless storybook illustrated by Mercer Mayer, *Frog, Where Are You?* (1969). The language samples were recorded and transcribed, then meticulously coded for lexical, morphological, and syntactic structures and errors. The format used for the coding was the standard Codes for the Human Analysis of Transcripts (CHAT) format, as modified by Curtiss, Schaeffer, Sano, MacSwan, and Masilon (1996) and adapted to Spanish by Valadez, MacSwan, and Martínez (1997). This system of marking errors is consistent with research on child language and language impairments. Once the transcripts were coded, computer analysis provided evidence of the child's error rate.

Every native speaker of a language demonstrates some degree of error due to slips of the tongue, tiredness, fatigue, or other factors and, according to researchers cited by MacSwan (2001), an error rate as high as 10% may be considered normal. In his report, MacSwan found that of the 134 students sampled, 96% had a syntax error rate of less than 6%, and 3% of the students had an error rate of between 6% and 10%. The students' error rates in morphology were similar in that 97% of the students had an error rate of 10% or less. Table 5 illustrates the rate of error in the natural language samples.

Based on these reported error rates, between 97% and 99% of the students sampled were determined by linguistic analysis to be proficient in their native language. When we compared these children's scores on the WMLS, the IPT, and the LAS—O, we found clear discrepancies between the language-speaking abilities as evidenced by the natural language sample and the language assessment scores. Results of the LAS—O indicated that 33% of the students were classified as non-speakers of their native language. The IPT classified almost 10% of the students as non-speakers of Spanish. The WMLS showed that 9% of the students were classified as limited, or very limited, in their native language. Table 6 illustrates the results of the MacSwan (2001) study.

Table 5

Rate of Error in Morphology and Syntax in the Spanish Natural Language Sample (n = 134)

Proportion of error	Morphology	Syntax
< 6%	94.0%	96.0%
6–10%	3.0%	3.0%
11%	1.4%	0
13%	.7%	0
17%	.7%	0
19%	.7%	0

Table 6

Frequencies of Spanish-Language Proficiency Levels (n = 134)

Assessment	Proficiency designation
Woodcock-Muñoz Language Survey	
Advanced fluency	4.5%
Fluent	38.1%
Limited fluency	42.4%
Very limited fluency	9.0%
IDEA Proficiency Test	
Fluent	9.7%
Limited	77.6%
Non-speaker	9.7%
Language Assessment Scales—Oral	
Fluent	24.6%
Limited fluent	32.8%
Non-speaker	32.8%

MacSwan's (2001) study and the present study yielded similar results in many respects. By analyzing natural language samples of children, MacSwan was able to empirically confirm that children are competent in their native language by the early elementary school years. The children studied in the MacSwan research were Spanish-speaking general education students ages 6 to 8. The children in the present study were native English speakers between the ages of 8 and 12. Surely the native English-speaking students in this study, given their age advantage, would have mastered language skills in English equivalent to those mastered by their younger Spanish-speaking peers in Spanish.

The age differences among these children may be one factor that accounts for differences in the proficiency designation of the children. This may be due to a difference in the construct of the assessment by age group. The difficulty of the items for one age group in comparison to another might differ. In MacSwan's (2001) study, 9.7% of the children were considered fluent Spanish speakers according to the Spanish version of the IPT; in the present study, 85% of the children were considered fluent English speakers according to the English version of the IPT. In MacSwan's study, 24.6% of the children were considered fluent Spanish speakers according to the Spanish version of the LAS—O; in this study, 100% of the children were considered fluent English speakers according to the LAS—O. In MacSwan's study, 42.6% of the children were considered fluent Spanish speakers according to the Spanish version of the WMLS; in this study, none of the children were considered fluent English speakers according to the WMLS. The IPT and LAS—O results may be considered normal based on the age differences of the children; more children in the present study may have been considered fluent because they were older and have developed more savvy test-taking skills. For reasons unknown, the differences between MacSwan's WMLS results and the WMLS results from the present study do not follow this pattern.

The English and Spanish versions of each of these assessments should be considered entirely different measurements, not just a translated version from English to Spanish or vice versa. Therefore, it should not be considered unusual that there is some difference between the frequencies of the English and Spanish proficiency levels of the WMLS, the IPT, or the LAS—O. This point is often lost on educators, who treat these test scores as being identical in both the English and Spanish versions. Both MacSwan's (2001) research and the present research offer strong evidence that children can be erroneously considered non-proficient in both their native language and their second language, thereby making them strong candidates for placement into special education.

Conclusion

If monolingual native speakers of English cannot achieve a score of “fluent” on an assessment of English oral-language proficiency, the assessment does not correctly measure the construct of oral-language ability. Oral-language assessments must measure the essential elements of knowing a language, not just lexical knowledge. This includes the ability to produce new utterances and recombine forms to represent ideas, events, and objects on an abstract level, to produce forms of the language they have never heard before, and to demonstrate mastery over the general functions of language such as syntax, morphology, semantics, and pragmatics. This point is important because we cannot assume that a single psychologically constructed test will accurately describe language fluency. To develop a complete picture of a child’s oral-language ability, a complete evaluation of data from multiple sources must be conducted.

On the whole, these tests may contribute to a deficit view of any child who speaks English as a second language. To be considered fit for general education, ELL children must perform well on a test that befuddles many of their native English-speaking peers. How can children who speak English as a second language be expected to perform well on tests that fail to measure native speaking ability of even native English-speaking children? These tests lack validity because they often assess something different from English-language proficiency. When assessments are constructed on the basis of CALP, ELL students are asked to provide a kind of language proficiency not demanded of native English-speaking students. For example, this research found that none of the native English-speaking children achieved a level of fluency beyond the “limited English” range on the WMLS, yet these students were normally developing children in a mainstream academic program. The students in this research were in no danger of being placed into a special education program because of the results of this assessment. Generally, native speakers of English are presumed to have “academic language proficiency” and not given such an assessment. ELL students are required to take assessments of language proficiency upon entry into school district, and often on the basis of those scores, they are considered candidates for referral into special education when other types of interventions would be more appropriate.

ELL children, on the basis of low scores on these language measures, are therefore targeted as children “at risk” of academic failure and as possible candidates for special education referral. This situation is exacerbated by state laws that forbid the teaching of academic content areas in an ELL’s native language. With few exceptions, teachers in Arizona, California, and Massachusetts are prohibited from teaching students in their native language and are limited to using solely formal language assessments to determine a

child's oral-language ability. Because teachers are not allowed to use an ELL student's native language when discussing content-area instruction, they have a limited ability to make classroom observations of the child using both his or her native language and English—an essential element in determining oral-language ability.

Ortiz and García (1995) called for increased research on two issues plaguing the field of bilingual special education. The first is the absence of research examining the influence of language proficiency on special education placement. The second is the absence of data that help to distinguish cultural or linguistic differences from disabilities. The first step in addressing these pressing needs must be to develop a valid and reliable assessment of oral language that is clearly distinguishable from measures of literacy or other forms of academic achievement. The development of oral language is an important step in developing literacy and achievement within the classroom, but there must be an assessment that measures oral language without determining if the child is also literate. This type of language assessment research is critical to inform educators who place students in language support programs according to their level of English-language proficiency, and those who refer ELL students into special education.

References

- American Educational Research Association. (1999). *Standards for educational and psychological testing*. (1999). Washington, DC: Author.
- Artiles, A. J. (1994). Overrepresentation of minority students in special education: A continuing debate. *The Journal of Special Education*, 27(4), 410–437.
- Artiles, A. J., Rueda, R., Salazar, J. J., & Higareda, I. (2000). Factors associated with English-language learner representation in special education: Evidence from urban school districts in California. In D. Losen & G. Orfield (Eds.), *Minority issues in special education in the public schools*. Cambridge, MA: Harvard Publishing Group.
- Brown, R. W. (1973). *A first language*. Cambridge, MA: Harvard University Press.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81–105.
- Crystal, D. (1997). *The Cambridge encyclopedia of language* (2nd ed.). New York: Cambridge University Press.

- Cummins, J. (1980). The cross-linguistic dimensions of language proficiency: Implications for bilingual education and the optimal age issue. *TESOL Quarterly*, 14(3), 175–187.
- Cummins, J. (2000). *Language, power and pedagogy: Bilingual children in the crossfire*. Clevedon, England: Multilingual Matters.
- Curtiss, S., Schaeffer, J., Sano, T., MacSwan, J., & Masilon, T. (1996). A grammatical coding and analysis system for language data from normal and brain-damaged children. Paper presented at the Joint International Conference of the Association for Literacy and Linguistic Computing and the Association for Computers and the Humanities, University of Bergen, Norway.
- Dalton, E., & Barrett, T. (2001). *IDEA oral language proficiency test, Forms E & F English technical manual*. Brea, CA: Ballard & Tighe.
- DeAvila, A., & Duncan, S. (1990). *Language Assessment Scales oral technical report*. Monterey, CA: CTB Macmillan McGraw-Hill.
- Fey, P. (2001, April). *Of mushrooms, igloos and the Woodcock-Muñoz Language Survey—Spanish (LSS), Subtests I and II*. Paper presented at the annual meeting of the American Educational Research Association, Seattle.
- Figuera, R. (1991). Bilingualism and psychometrics. *Diagnostique*, 17, 70–85.
- Fromkin, V., & Rodman, R. (1993). *An introduction to language* (5th ed.). Orlando, FL: Harcourt, Brace, Jovanovich.
- Geisinger, K. F., & Carlson, J. (1992). *Assessing language-minority students*. (ERIC Document Reproduction Service No. ED 356232)
- Gleason, J. B. (1997). *The development of language*. Needham Heights, MA: Allyn & Bacon.
- Hakuta, K., Butler, Y. G., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* (policy Report 2000–1). Los Angeles, CA: University of California Linguistic Minority Research Institute.
- Hernández, R. D. (1994). Reducing bias in the assessment of culturally and linguistically diverse populations. *The Journal of Educational Issues of Language Minority Students*, 14, 269–300.
- Limbos, M., & Geva, E. (2001). Accuracy of teaching assessments of second-language students at risk for reading disability. *Journal of Learning Disabilities*, 34(2), 136–151.
- MacSwan, J. (2000). The threshold hypothesis, semilingualism, and other contribution to a deficit view of linguistic minorities. *Hispanic Journal of Behavioral Sciences*, 20(1), 3–45.

- MacSwan, J. (2001, April). *The non–non crisis: Knowledge of language and problems of construct validity in native language assessment*. Paper presented at the annual meeting of the American Educational Research Association, Seattle.
- MacSwan, J., Rolstad, K., & Glass, G. (2002). Do some school-age children have no language? Some problems of construct validity in the Pre-LAS Español. *Bilingual Research Journal*, 26(2), 213–238.
- MacSwan, J., & Rolstad, K. (2003). Linguistic diversity, schooling, and social class: Rethinking our conception of language proficiency in language minority education. In C. B. Paulston & R. Tucker (Eds.), *Sociolinguistics: The essential readings*. Oxford, England: Blackwell.
- Mahoney, K. (2001, April). *The Idea Proficiency Test (IPT) Spanish: Seven strikes and you're out!* Paper presented at the annual meeting of the American Educational Research Association, Seattle.
- Mayer, M. (1969). *Frog, where are you?* New York: Dial Books.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.
- Ortiz, A., & García, S. (1995). Serving Hispanic students with learning disabilities: Recommended policies and practices. *Urban Education*, 29(4), 471–481
- Ortiz, J. (2001, April). *The LAS—O Español: Language assessment scale or losing the ability to speak?* Paper presented at the annual meeting of the American Educational Research Association, Seattle.
- Ochoa, S. H., Galarza, A., & González, D. (1996). An investigation of school psychologists' assessment practices of language proficiency with bilingual and limited-English proficient students. *Diagnostique*, 21, 17–36.
- Ochoa, S., & González, D. (1996). The training and use of interpreters in bilingual psycho-educational assessment: An alternative in need of study. *Diagnostique*, 21, 19–22.
- Sommers, R. (1989). Language assessment: Issues in the use and interpretation of tests and measures. *School Psychology Review*, 18(4) 452–462.
- Todd, S., Fiske, K., & Dopico, H. (1994). Vocational assessment for LEP students. *Journal for Vocational and Special Needs Education*, 16(2), 16–28.
- Valadez, C., MacSwan, J., & Martínez, C. (1997, April). *Toward a new view of low achieving bilinguals: Syntactic competence in designated "semilinguals."* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

- Woodcock, R., & Muñoz-Sandoval, A. (2001). *Woodcock-Muñoz language survey normative update comprehensive manual*. Itasca, IL: Riverside.
- Zehler, A. M., Hopstock, P. J., Fleischman, H. L., & Greniuk, C. (1994). An examination of assessment of limited English proficient students. Retrieved June 25, 2005, from <http://www.ncela.gwu.edu/pubs/siac/lepases.htm>

Endnotes

- ¹ The “non–non” crisis refers to students who, by virtue of language assessment results, are considered non-proficient in their native language and non-proficient in their second language. Students who receive such scores on language assessments are very likely to be candidates for referral and placement into special education.
- ² Genie, a linguistic isolate, was confined to a small bedroom where she was tied to an infant potty seat and purposely not spoken to for 12 years starting at the age of 20 months. Genie was rescued at the age of 13, at which time she could barely walk, couldn’t chew or bite, and neither understood nor spoke any language. Despite years of subsequent education, Genie demonstrated impaired grammar and pragmatic performance but adequate vocabulary knowledge (Curtiss, 1989).