LEADERS

Test Review:
Clinical Evaluation of Language Fundamentals – Fourth Edition (CELF-4) Spanish

Version: 4th Edition
Copyright date: 2006
Grade or Age Range: 5 through 21 years
Author: Eleanor Semel, Ed.D. ; Elisabeth H. Wiig, Ph.D.; and Wayne A. Secord, Ph.D.
Publisher: PsychCorp

Table of Contents

1. PURPOSE

The CELF-4 Spanish is designed to assess the presence of a language disorder or delay in Spanish speaking students aged 5-21. Subtests were designed to aid in determining a student's diagnosis, strengths and weakness, eligibility for services, and how their performance on tasks affects the student's access to the general education curriculum. The CELF-4 Spanish contains a four-level assessment process in which the presence of a language disorder can be determined by calculating a Core Language score using only four subtests. Additional subtests allow the clinician to gain more information regarding the nature and extent of the disorder. Content areas include: morphology and syntax, semantics, pragmatics, and phonological awareness.

2. DESCRIPTION

Subtest	Age Range	Purpose	Format
Conceptos y Siguiendo Direcciones [1,2] (Concepts and Following Directions[1,2])	5-12	To measure the student's ability to: (a) Interpret oral directions of increasing length and complexity; (b) Recall names, characteristics, and order of objects from orally presented material; (c) Discrimination of pictured objects from several choices.	Student identifies pictured objects following oral directions from test administrator.
Estructura de palabras [1] (Word Structure[1])	5-8	To measure the student's use of morphological rules.	Student completes an orally presented sentence in reference to a visual stimulus.
Formulación de oraciones[1, 2, 3] (Formulated Sentences[1,2, 3])	5-21	To measure a student's ability to generate grammatically and semantically correct sentences given a visual stimulus.	Following an orally presented target word, the student generates a sentence in reference to a visual stimulus.
Recordando oraciones [1,2, 3] (Recalling Sentences [1,2, 3])	5-21	To measure a student's ability to recall and repeat sentences of increasing length and complexity.	The student repeats sentences orally presented by the administrator.
Clases de palabras 1 (Word Classes 1)	5-7	To measure an individual's ability to comprehend and explain relationships between 3-4 images.	Given 3-4 words, the student selects two words that go together and explains their relationship.
Clases de palabras 2 [2, 3] (Word Classes 2 [2, 3])	8-21	To measure an individual's ability to comprehend and explain relationships between orally presented target words.	Given 3-4 words, the student selects two words that go together and explains their relationship.

Definiciones de palabras [3] (Word Definitions [3])	10-21	To measure a student's ability to infer word meanings based on class relationships and shared meanings.	Student defines an orally presented word presented orally in a sentence.
Estructura de oraciones (Sentence Structure)	5-8	To measure a child's receptive language ability to interpret sentences of increasing length and semantic and grammatical complexity.	Following an orally presented sentence, the student points to the corresponding stimulus image.
Vocabulario expresivo (Expressive Vocabulary)	5-9	To measure a student's ability to use referential naming.	Student identifies an object, person, or action presented by the administrator.
Entendiendo párrafos (Understanding Spoken Paragraphs)	9-21	To measure an individual's ability to comprehend a verbally presented story by answering factual questions. The student's ability to infer is also measured as some questions require this skill.	The student answers comprehension questions related to an orally presented story.
Conocimiento fonológico (Phonological Awareness)	5-12	To measure a student's acquisition of sound structure and ability to manipulate sound through: (a) Syllable and phoneme segmentation; (b) Syllable blending and deletion (c) Phoneme manipulation and identification.	Comprised of 11 tasks of varying directives involving phonological and metalinguistic awareness.
Enumeración rápida y automática (Rapid Automatic Naming)	5-21	To measure the student's ability to produced automatic speech.	The student is timed during naming of color, shapes, and color-shape combinations.
Asociación de palabras (Word Associations)	5-21	To measure the student's ability to recall objects from a semantic category within a fixed time limit.	The student lists objects belonging to a semantic category within one minute.
Repetición de numeros 1 (Number Repetition 1)	5-16	To measure the student's working memory.	The student repeats a series of digits in the exact order presented by the administrator. Following, the student repeats digits in reverse order of an orally presented string of numbers.

Repetición de numeros 2 (Number Repetition 2)	17-21	To measure the student's working memory.	The student repeats a series of digits in the exact order presented by the administrator. Following, the student repeats digits in reverse order of an orally presented string of numbers.
Sequencias familiares 1 (Familiar Sequences 1)	5-16	To measure the student's ability to retrieve common information.	The student recites common information (e.g. days of the week, counting backwards, etc.) while being timed.
Sequencias familiares 1 (Familiar Sequences 2)	17-21	To measure the student's ability to retrieve common information.	The student recites common information (e.g. days of the week, counting backwards, etc.) while being timed.

[1] Core Language Score (Ages 5-8)
[2] Core Language Score (Ages 9-12)
[3] Core Language Score (Ages 13-21)

The CELF-4 Spanish Manual del Examinador (Examiner's Manual) describes examiner qualifications for the test. The test may be administered by Spanish-speaking SLPs, school psychologists, special educators, and diagnosticians. The manual cautions that the examiner must speak Spanish fluently with near-native proficiency in order to accurately conduct the test and record the students' responses. If a qualified examiner is not available, the test may be administered "by a paraprofessional with near-native proficiency in Spanish who has been trained in test administration" (Semel, Wiig, & Secord, 2006). Chapter 2 of the manual provides additional information regarding using interpreters to administer the CELF-4 Spanish. Specifically, the manual cautions, "a great deal of background training is necessary to prepare an interpreter to administer the test. Training should include information about normal speech and language development, appropriate testing practices, cultural factors among subgroups, regulations governing testing, educational backgrounds, and behavior management" (p. 12). Providing such extensive training would require a large investment of time and resources that may not be available in many settings.. Further, even if an individual was adequately trained, they will not have the clinical judgment necessary to determine if a child presents with a language disorder. Children may be misdiagnosed as a result of being administered the test by an under-qualified individual.

3. STANDARIZATION SAMPLE

The standardization sample for the CELF-4 Spanish used data collected from May 2004 through May 2005. The sample was representative of the Hispanic population in the United States for individuals 5 through 21 years of age and was stratified by demographic factors including age, gender, and parental education level. The standardization, reliability, and validity studies involved more than 1,100 children, adolescents, and young adults. Normative data is reported in 6 month intervals from 5;0-6;11, in 1 year intervals from 7;0-16;11, and in one 5 year interval from 17;0-21;11. Sample sizes for each age group ranged from 50 to 80 students. Inclusion into the standardization sample required completion of the test in a standard manner (e.g. no sign language was permitted).

Although bilingual children were included in the sample, Spanish was reported to be the first language of all participants in the standardization group. Approximately 30% of the sample lived in homes in which a language other than Spanish was also spoken (primarily English, but also French, Italian, and Portuguese). Further, the following percentages of the sample reportedly never spoke Spanish in the following contexts: 2.25% with siblings, 4.38% with friends, 9.63% in the classroom, and 6.25% at recess/leisure (Manual Técnico, p. 31). This reduces the validity of the test as children included in the standardization sample do not match the intended test population.

It should be noted that 39.7% of the sample reportedly received school services and/or were identified as having a "specific condition impacting educational performance" (p. 27). Specifically, 2% were identified as learning disabled, 1% was identified with ADHD, and less than 1% was identified with developmental delays, hearing impairments, or other conditions. About 11% of the sample was receiving speech and language services. The manual did not state whether or not the students who were previously identified as having a disability were accurately identified by the CELF-4 Spanish. According to Peña, Spaulding, & Plante (2006), inclusion of individuals with disabilities in the normative sample can negatively impact the test's discriminant accuracy, or its ability to differentiate between typically developing and disordered children. Specifically, when individuals with disabilities are included in the normative sample, the mean scores are lowered. As a result, children will only be identified as having a disability with an even *lower* score. Thus, children with mild disabilities will not be identified, compromising the sensitivity of the test.

4. VALIDITY

Content - Content Validity is how representative the test items are of the content that is being assessed (Paul, 2007). According to the Manual Técnico, the "content and subtest construction [of the CELF-4 Spanish] was designed to ensure that the subtests and subtest items adequately sample the language domains with particular attention to content that reflects more complex linguistic processes evident in adolescents and young adults" (p. 72). The manual notes that the content for "younger" students was "reviewed for comprehensiveness and appropriateness for those age ranges. The subtest content is indicative of the linguistic milestones achieved at various stages of development" (p. 72). The manual also notes that a group of experts reviewed items on the CELF-4 Spanish for potential biases based on their "understanding of language usage among the Hispanic population". The manual also states that test items were reviewed to ensure that they addressed the targeted skill for the subtest, that confounding elements of the task were minimized, and that the content of the item was familiar and appropriate for Spanish-speaking children. Responses to new tasks were analyzed to determine if additional responses could be considered correct to account for additional cultural and linguistic variability. Test developers used information from student's responses, research data, clinician input resulting from the standardization record forms, and input from scorers to determine test items that were particularly challenging or confusing.

The content validity is insufficient for several reasons. First, although the manual mentions procedures that were used to ensure content validity, details were not provided regarding specific protocols that were in place. For example, although they mention that the test items "adequately sample the language domains …" and that "subtest content is indicative of the linguistic milestones achieved at various stages of development" (p. 74), information was not provided regarding *how* the test items were determined to be appropriate. Further, the manual does not mention the level of expertise of the "expert panel" that reviewed test content for potential biases. Expert knowledge of a variety of dialects requires an enormous and sophisticated knowledge base. In some cases, the intricacies of dialectal variations are so small that even highly educated linguists find it difficult to determine differences between cultures. As specific information regarding the background and training of the panel was not provided, one cannot be confident that the items in this test are completely free of bias.

Construct – Construct validity assesses if the test measures what it purports to measure (Paul, 2007). It was measured using special group studies comprised of typically developing and language disordered individuals. The diagnosis of these students was compared with their status as determined by the CELF-4 Spanish to determine the test's diagnostic accuracy.

Reference Standard

In considering the diagnostic accuracy of an index measure such as the CELF-4 Spanish, it is important to compare the child's diagnostic status (affected or unaffected) with their status as determined by another measure. This additional measure, which is used to determine the child's 'true' diagnostic status, is often referred to as the "gold standard." However, as Dollaghan & Horner (2011) note, it is rare to have a perfect diagnostic indicator, because diagnostic categories are constantly being refined. Thus, a *reference standard* is used. This is a measure that is widely considered to have a high degree of accuracy in classifying individuals as being affected or unaffected by a particular disorder, even accounting for the imperfections inherent in diagnostic measures (Dollaghan & Horner, 2011).

The reference standard used to identify children as having a language disorder (part of the sensitivity group) was a score below 1.5 SD on a standardized test of language skills. The clinical study group included 116 children, aged 5-21, who were tested by speech-language pathologists in the United States Puerto Rico, and Mexico. It should be noted that when divided into the 15 normative age ranges, this would result in 7-8 children per age group. It should be noted that this sample is too small as compared to the standards in the field, which recommends sample sizes of 100 or more (Guadagnoli & Velicer, 1988). The manual does not include information regarding the tests that were used to diagnose the children with a language disorder. Therefore, we cannot be sure of the test's diagnostic accuracy. It is also important to note that arbitrary cut off scores on standardized language tests do not accurately discriminate between typically developing children and children with a language disorder (Spaulding, Plante, & Farinella, 2006). Thus, the true diagnostic status of the sensitivity group is unknown. Further, no information was provided regarding the background and training of the speech pathologists who diagnosed the children. Therefore we cannot be sure they had the knowledge and skills necessary to diagnose bilingual and/or culturally and linguistically diverse children as language disordered. The reference standard is insufficient because students were included based on previous test performance on potentially invalid measures using an arbitrary cut off score.

The reference standard used to identify the *specificity* group was a control group of 116 students who were randomly selected from the standardization sample. It should be noted that when divided into the 15 normative age ranges, this would result in 7-8 children per age group. It should be noted that this sample is too small as compared to the standards in the field, which recommends sample sizes of 100 or more (Guadagnoli & Velicer, 1988). They were matched to the sensitivity group based on age, parental education, and sex. According to Dollaghan (2007) performance on the reference standard cannot be assumed. As the same reference standard (a score above 1.5 SD below the mean on a standardized test) for the sensitivity group was not applied to the specificity group, one cannot be certain of the diagnostic status of the specificity group. Further, as students in this group were randomly selected from the standardization sample, they are not guaranteed to be typically developing. As noted above, 39.7% of the standardization sample reportedly received school services and/or were identified as having a "specific condition impacting educational performance" (p. 27). Additionally, 11% of the sample was receiving speech and language services. Thus, the reference standard for the specificity group is insufficient as the diagnostic status of the students was not confirmed

Sensitivity and Specificity

Sensitivity measures the proportion of students who have a language disorder that will be accurately identified as such on the assessment (Dollaghan, 2007). For example, sensitivity means that a child previously diagnosed with a language disorder will score within the limits to be identified as having a language disorder on this assessment. The CELF-4 Spanish reports sensitivity measures to be .96, .86, and .52 for -1, -1.5, and -2 standard deviations (SD) below the mean. According to Plante & Vance (1994), validity measures above .9 are good, measures between .8 and .89 are fair, and measures below .8 are unacceptable. Sensitivity measures reported in the manual of the CELF-4 Spanish range from "unacceptable" to "good" according to standards in the field. However, since the reference standard was not specifically identified, the true status of the children in the sensitivity group is not known. Therefore, the sensitivity measures reported cannot be considered valid measures. Even if reported sensitivity values were valid, it is important to consider the implications of these measures. A sensitivity of .86 means that 14/100 children who have a language disorder will not be identified as such by the CELF-4 Spanish, and therefore will not receive the extra academic and language support that they need. As a result of the lack of a valid reference standard, sensitivity measures are considered insufficient and invalid.

Specificity measures the proportion of students who are typically developing who will be accurately identified as such on the assessment (Dollaghan, 2007). For example, specificity means that a child with no history of a language disorder will score within normal limits on the assessment. The CELF-4 Spanish reports specificity measures to be 0.87 at -1 SD below the mean, 0.95 at -1.5 SD below the mean, and 1.00 at -2 SD below the mean. These measures range from "fair" to "good" (Plante & Vance, 1994), but as with the sensitivity group, since the reference standard was not valid (i.e. not identified so validity cannot be confirmed) we cannot be sure of the true status of the students. Therefore these measures cannot be considered valid. It is also important to consider the implications. A specificity of .87 means that 13/100 typically developing children will be identified as having a language disorder and may be unnecessarily referred for special education services. Further, as the reference standard (score below 1.5 SD on a standardized language assessment) was not applied to the specificity group, it is important to consider that students in this group cannot be guaranteed to be free of a disorder. As mentioned above, 39.7% of the standardization sample was receiving special related services and 11% of the sample was receiving speech and language services. Further, this decreases the validity of the test due to spectrum bias which occurs when the sample population does not represent the full spectrum of the clinical population (Dollaghan & Horner, 2011). Therefore, the specificity measures reported are insufficient and invalid.

Likelihood Ratio

According to Dollaghan (2007), likelihood ratios are used to examine how accurate an assessment is at distinguishing individuals who have a disorder from those who do not. A positive likelihood ratio (LR+) represents the likelihood that an individual, who is given a positive (disordered) score on an assessment, actually has a disorder. The higher the LR+ (e.g. >10), the greater confidence the test user can have that the person who obtained the score has the target disorder. Similarly, a negative likelihood ratio (LR-) represents the likelihood that an individual who is given a negative (non-disordered) score actually does not have a disorder. The lower the LR- (e.g. < .10), the greater confidence the test user can have that the person who obtained a score within normal range is, in fact, unaffected. Likelihood ratios were not calculated due to insufficient and invalid sensitivity and specificity measures resulting from lack of a valid reference standard for either group.

Overall, construct validity, including the reference standard, sensitivity and specificity, and likelihood ratios, was determined to be insufficient due to lack of validity of the reference standard as well as unacceptable sensitivity and specificity measures.

Concurrent - Concurrent validity is the extent to which a test agrees with other valid tests of the same measure (Paul, 2007). According to McCauley & Swisher (1984) concurrent validity can be assessed using indirect estimates involving comparisons amongst other tests designed to measure *similar behaviors*. If both test batteries result in similar scores, the tests "are assumed to be measuring the same thing" (McCauley & Swisher, 1984, p. 35). Concurrent validity was measured by comparing the CELF-4 Spanish to the CELF-3 Spanish using a group of 91 typically developing children aged 6-21 years. The CELF-4 Spanish was administered before the CELF-3 Spanish for all students; time between the tests ranged from 7 to 60 days. Corrected correlation coefficients were .76 for Core Language Score, .61 for the Receptive Language Score, and .67 for the Expressive Language score. According to the standard as recommended by Plante & Vance (1994), these concurrent validity measures are considered "unacceptable."

This comparison is invalid for several reasons. First, the definition of concurrent validity requires a "valid test" is used as the comparison measure but the CELF-3 Spanish is not valid. Second, the correlation measures between the CELF-4 Spanish and the CELF-3 Spanish are considered "unacceptable" as compared to the standard in the field (Plante & Vance, 1994). Further, the age range of the students included in the study was 6-21, which does not cover the entire age range of the CELF-4 Spanish. Finally, this concurrent validity measure was determined using an old version of the same test. By using this comparison, concurrent validity may appear higher as the same test is being used, just modified. Therefore, it is not representative of how the CELF-4 Spanish compares to other tests.

5. RELIABILITY

According to Paul (2007, p. 41), an instrument is reliable if “its measurements are consistent and accurate or near the ‘true’ value.” Reliability may be assessed using different methods, which are discussed below. It is important to note, however, a high degree of reliability alone does not ensure diagnostic accuracy. For example, consider a standard scale in the produce section of a grocery store. Say a consumer put on 3 oranges and they weighed 1 pound. If she weighed the same 3 oranges multiple times, and each time they weighed one pound, the scale would have *test-retest reliability*. If other consumers in the store put the same 3 oranges on the scale and they still weighed 1 pound, the scale would have *inter-examiner reliability*. Now say an official were to put a 1 pound calibrated weight on the scale and it weighed 2 pounds. The scale is not measuring what it purports to measure—it is not valid. Therefore, even if the reliability appears to be sufficient as compared to the standards in the field, if it is not valid it is still not appropriate to use in assessment and diagnosis of language disorder.

Test-Retest Reliability – Test-retest reliability is a measure used to represent how stable a test score is over time (McCauley & Swisher, 1984). This means that despite the test being administered several times, the results are similar for the same individual. Test-retest reliability was calculated by retesting 132 students from the standardization sample 7-30 days (mean = 14 days) after the initial administration. The same test examiner was used each time. A correlation coefficient was calculated for individuals across the age range of the test. Salvia, Ysseldyke, and Bolt (2010) recommend the *minimum* standard for reliability be .9 when using the test to make educational placement decisions, including SLP services. Correlation coefficients were corrected to account for the variability of the standardization sample (Allen & Yen, 1979; Magnusson, 1967, as cited in Wiig, Semel, & Secord, 2006). According to the Manual Técnico, across ages, and subtests, corrected reliability coefficients ranged from .52 to .93. Thus, test-retest reliability is insufficient.

Inter-examiner Reliability– Inter-examiner reliability is used to measure the influence of different test scorers or different test administrators on test results (McCauley & Swisher, 1984). It should be noted that the inter-examiner reliability for index measures is often calculated using specially trained examiners. When used in the field, however, the average clinician will likely not have specific training in test administration for that specific test and thus the inter-examiner reliability may be lower in reality. Inter-examiner reliability was calculated using a group of 14 trained raters. Each subtest that required subjective scoring (*Formulación de oraciones, Asociación de palabras, Definiciones de palabras, Vocabulario expresivo, Estructura de palabras, Clases de palabras 1/2*) was rated independently by two raters and then compared. A third rater resolved any discrepancies. According to the Manual Técnico, correlation between scorers ranged from .81 to .99. According to the standard as recommended by Salvia, Ysseldyke, and Bolt (2010), 3 out of 7 subtests did not meet the minimum standard for reliability. Therefore, inter-examiner reliability is insufficient.

Inter-item Consistency – Inter-item consistency assesses whether parts of an assessment are in fact measuring something similar to what the whole assessment claims to measure (Paul, 2007). Inter-item consistency was calculated using a Coefficient Alpha and the split-half method. For the standardization sample, across ages and subtests, coefficient alphas ranged from .65 to .97. Across subtests, at least one age group did not meet the minimum standard in the field for reliability (Salvia, Ysseldyke, and Bolt, 2010). In the split half method, the authors divided the targets into two groups and calculated the correlation between the test halves for each subtest. Across age ranges and subtests, average coefficients ranged from .62 to .98. As with the coefficient alpha, across subtests, at least one age group did not meet the minimum standard in the field for reliability (Salvia, Ysseldyke, and Bolt, 2010). Inter-item consistency was also calculated for 116 students who were diagnosed with a language disorder. Across subtests, the coefficient alpha ranged from .83 to .96; three out of 10 subtests did not meet the minimum standard (Salvia, Ysseldyke, and Bolt, 2010). For the split half method, coefficients ranged from .83 to .97; 2 out of 10 subtests did not meet the minimum standard (Salvia, Ysseldyke, and Bolt, 2010). Therefore, inter-item consistency is insufficient.

Overall, the reliability, including the test-retest, and inter-examiner reliability, is considered insufficient. Across subtests, at least one age range did not meet the minimum standard for reliability as defined by Salvia, Ysseldyke, and Bolt (2010). In addition, the use of trained scorers does not reflect the real world use of the CELF-4 Spanish and contributes to its lack of reliability.

6. STANDARD ERROR OF MEASUREMENT

According to Betz, Eickhoff, and Sullivan (2013, p.135), the Standard Error of Measurement (SEM) and the related Confidence Intervals (CI), “indicate the degree of confidence that the child’s true score on a test is represented by the actual score the child received.” They yield a range of scores around the child’s score, which suggests the range in which their “true” score falls. Children’s performance on standardized assessments may vary based on their mood, health, and motivation. For example, a child may be tested one day and receive a standard score of 90. Say he was tested a second time and he was promised a reward for performing well; he may receive a score of 96. If he were to be tested a third time, he may not be feeling well on that day, and thus receive a score of 84. As children are not able to be assessed multiple times to acquire their “true” score, the SEM and CIs are calculated to account for variability that is inherent in individuals. Current assessment guidelines in New York City require that scores be presented within confidence intervals whose size is determined by the reliability of the test. This is done to better describe the student’s abilities and to acknowledge the limitations of standardized test scores (NYCDOE CSE SOPM 2008).

The clinician chooses a confidence level (usually 90% or 95%) at which to calculate the confidence interval. A higher confidence level will yield a larger range of possible test scores, including a child's true range of possible scores. Although a larger range is yielded with a higher confidence interval, the clinician can be more *confident* that the child's 'true' score falls within that range. A lower level of confidence will produce a smaller range of scores but the clinician will be less confident that the child's true score falls within that range. The wide range of scores necessary to achieve a high level of confidence, often covering two or more standard deviations, demonstrates how little information is gained by administration of a standardized test. For example, for a child between 5;0-5;5 on the CELF-4 Spanish, the SEM at the 90% confidence level is +/- 6 for Core Language Score (CLS). If the child were to achieve a 75 as his CLS, considering the CI, users can be 90% confident that the child's true language abilities would be represented by a score between 69 and 81. Thus, all the clinician can determine from administration of the CELF-4 Spanish is that this child's true language ability (according to the CELF-4 Spanish) ranges from moderately-severely impaired to low normal. Without considering the CI, this child would be labeled LD inappropriately and given special education services unnecessarily. This has serious long-term consequences on the child's development and achievement. The wide range of the CI makes the scores from the CELF-4 Spanish insufficient as children may be misdiagnosed.

7. BIAS:

According to Crowley (2010), IDEA 2004 regulations stress that assessment instruments must not only be "valid and reliable" but also free of "discriminat[ion]on a racial or cultural basis." The CELF-4 Spanish contains many inherent biases against culturally and linguistically diverse children. Some of these biases are described in Chapter 2 of the examiner's manual (p. 13) and should be carefully considered when administering the CELF-4 Spanish to culturally and linguistically diverse children.

Linguistic Bias

English as a Second Language

Paradis (2005) found that children learning English as a Second Language (ESL) may show similar characteristics to children with Specific Language Impairments (SLI) when assessed by language tests that are not valid, reliable, and free of bias. Thus, typically developing students learning English as a Second Language may be diagnosed as having a language disorder when, in reality, they are showing signs of typical second language acquisition. Many students who will be administered the CELF-4 Spanish may be students who are learning English as a second language in school. Consider, for example, a child from a Spanish speaking family who enters kindergarten. Although they only spoke Spanish until they started school at 5 years old, they may refuse to speak it once they start learning English in school. Thus, they may be referred for an evaluation in English, a language they have only been learning for about one year. Although Spanish was their first language, after a year of little to no practice using it, they may be experiencing subtractive bilingualism. This occurs when "acquisition of the majority language comes at the cost of loss of the native language" (Paradis, Genesee, & Crago, 2011, p. 49). As a child gains skills in their second language and ceases using their first language, their proficiency in the first language declines. Since language tests are cognitively demanding and require significant amounts of metalinguistic and academic language skills and vocabulary, a typically developing child experiencing subtractive bilingualism may show depressed skills in both languages. According to ASHA, clinicians working with diverse and bilingual backgrounds must be familiar with how elements of language differences and second language acquisition differ from a true disorder (ASHA, 2004). Only a clinician with significant training and experience evaluating bilingual children and using other assessment tools (i.e. not a norm-referenced test) would be able to pick up on why a bilingual child would have delayed skills in both languages. If bilingual students are tested using only the CELF-4 Spanish, they may be falsely identified as having a language disorder when, in reality, they are experiencing subtractive bilingualism.

On the CELF-4 Spanish, children learning English as a second language may be falsely identified as having a language disorder on subtests that use academic language and concepts. Many children who learn English as a second language once they enter school still only speak Spanish in their homes. Thus students may have stronger skills in English for academic concepts that they have learned in school, but may not necessarily have equivalent skills in Spanish since such concepts are not discussed at home. For example, on the *Conceptos y Siguiendo Direcciones* subtest, students may be learning temporal and spatial concepts in school and be familiar with that vocabulary in English, but may not necessarily be familiar with the terms in Spanish. Therefore, when administered the CELF-4 Spanish, they may be falsely identified as having a language disorder.

Dialectal Variations

A child's performance on the CELF-4 Spanish may also be affected by the dialect of Spanish that is spoken in their homes and communities. The manual does not provide information regarding the dialect of Spanish that is used. It is important to note that there are many different dialects of Spanish from different regions that vary significantly. In the normative sample alone, 6 different countries of origin are reported: Mexico, Central and South America, Puerto Rico, Dominican Republic, Cuba, and "other". It can be safely assumed that this test will be administered to children who speak even more dialects of Spanish. It is important to consider the issues of the test being administered in a child's non-native dialect of either Spanish or English. For example, imagine being asked to repeat the following sentence, written in Early Modern English: "Whether 'tis nobler in the mind to suffer the slings and arrows of outrageous fortune or to take arms

> against a sea of troubles And by opposing end them" (Shakespeare, 2007). Although the content of the sentence consists of words in English, because of the unfamiliar structure and semantic meaning, it would be difficult for a speaker of SAE to repeat this sentence as compared to a similar sentence in SAE. The same would hold true for being asked to repeat a sentence in a dialect of Spanish that was different from the child's.
>
> Speakers of various dialects face a similar challenge when asked to complete tasks such as the *Recordando Oraciones* subtest of the CELF-4 Spanish. The goal of this subtest is to assess a child's syntactical development. Such tests are inappropriate for speakers of other dialects, as their syntactical structure may not correlate to that of the stimulus item. Speakers of various dialects may also be at a disadvantage during the *Vocabulario Expresivo* subtest. Specific words vary drastically by region and dialect. A child may provide a response that is consistent with their own dialect but if it is not a suggested response, they may be penalized. Thus, speakers of different dialects may be falsely identified with a language disorder.

Socioeconomic Status Bias

Research has shown that SES positively correlates with vocabulary knowledge; children from low SES families have been shown to have smaller vocabularies than their higher SES peers. Hart & Risley (1995) found that a child's vocabulary correlates with his/her family's socio-economic status; parents with low SES (working class, welfare) used fewer words per hour when speaking to their children than parents with professional skills and higher SES. Thus, children from families with a higher SES will likely have larger vocabularies and thus will likely show a higher performance on standardized child language tests. Horton-Ikard & Weismer (2007) found that children from low SES homes performed worse than higher SES peers on norm-referenced vocabulary tests (Peabody Picture Vocabulary Test-III and Expressive Vocabulary Test), and on a measure of lexical diversity (Number of Different Words) during a spontaneous language sample. However, SES was not a factor in the child's performance on a fast mapping task for novel word learning. These, along with other studies that have come out in the last decade, demonstrate that using norm-referenced vocabulary tests to identify disability is an important factor in the over-referral of minority and low SES students for special education services.

A child from a lower SES background may be falsely identified as having a language disorder on standardized language tests due to a smaller vocabulary than his higher SES peers. The CELF-4 Spanish contains items that are biased against children from low SES backgrounds because they require knowledge of lower frequency vocabulary items. For example, on the *Vocabulario Expresivo* subtest, a child from a lower SES may not have exposure to some of the lower frequency vocabulary words such as esqueleto [skeleton], or estampilla [stamp]. Further, the *Conceptos y siguiendo direcciones* subtest requires prior knowledge of words in the target to provide an appropriate response (e.g. última [last], están separadas por [are separated by]). As a result of vocabulary items on the CELF-4 Spanish, children from low SES backgrounds may have lower scores when compared to higher SES peers, making this an inappropriate test for children from low SES backgrounds.

Prior Knowledge/Experience

A child's performance on the CELF-4 Spanish may also be affected by their prior knowledge and experiences. For example, a child who has never attended school may not be familiar with some of the academic vocabulary in the *Conceptos y siguiendo direcciones* subtest; a child who does not have a pet may not be familiar with the word *veterinaria* [veterinarian] in the *Vocabulario Expresivo* subtest. It is also important to consider that the format of the test may affect a child's performance if they do not have prior experiences with the specific type of testing. According to Peña, & Quinn (1997), children from culturally and linguistically diverse backgrounds do not perform as well on assessments that contain tasks such as labeling and known information questions as they are not exposed to these tasks in their culture. The CELF-4 Spanish contains various testing formats, many of which are dependent upon prior knowledge and experience. The *Vocabulario Expresivo* subtest is a task based entirely on labeling. A child who has not been exposed to this type of testing may label a *periódico [newspaper]* as "para leer" as they have been exposed to function-type description tasks rather than labeling the object itself. Further, the *Entendiendo Párrafos* subtest requires the child to respond to a "known information question." In this subtest, the student is required to listen to a short passage read aloud by the clinician and then responds to questions related to the story. A child who is not accustomed to an adult asking questions to which they already know the answer may fail to respond appropriately.

Further, a child's performance on the test may have been affected by their prior exposure to books. According to Peña and Quinn (1997), some infants are not exposed to books, print, take-apart toys, or puzzles. The CELF-4 Spanish requires children to attend to the test book for the length of the assessment, which can often take hours. This may be challenging for a child who has not had prior exposure with structured tasks. He or she must also realize that pictures and symbols have meaning and attend to them (print awareness); this is not an innate skill but a learned one. In addition, lack of access to books and print materials results in a lack of familiarity with letters and sounds and delayed pre-literacy skills including letter knowledge and phonological awareness; this leads to reduced metalinguistic ability. For example, the *Formulación de Oraciones* subtest also requires significant metalinguistic ability. It requires the student to manipulate words by putting them in different positions to create various meanings. A child without the chance to play with and realize the value of language through books and word games may experience significant difficulty with this task due to lack of opportunity to develop his or her metalinguistic skills.

Cultural Bias

According to Peña & Quinn (1997), tasks on language assessments often do not take into account variations in socialization practices. For example, the child's response to the type of questions that are asked (e.g. known questions, labeling), the manner in which they are asked, and how the child is required to interact with the examiner during testing, may be affected by the child's cultural experiences and practices. Please see specific information above regarding cultural biases in testing format.

It is also important to consider that during test administration, children are expected to interact with strangers. In middle class mainstream American culture, young children are expected to converse with unfamiliar adults as well as ask questions. In other cultures, however, it may be customary for a child to not speak until spoken to. When he does speak, the child often will speak as little as possible or only to do what he is told. If a child does not respond to the clinician's questions because of cultural traditions, they may be falsely identified as having a language disorder.

Attention and Memory

Significant attention is required during administration of standardized tests. If the child is not motivated by the test's content, or they exhibit a lack of attention or disinterest, they will not perform at their true capacity on this assessment. Further, fatigue may affect performance on later items in the test's administration. Even a child without an attention deficit may not be used to sitting in a chair looking at a picture book for an hour. A child that has never been in preschool and has spent most of his days in an unstructured environment and playing with peers and siblings may find it very challenging to sit in front of a book for extended periods of time. It should be noted that according to the Examiner's manual, short breaks are permitted during the test if the clinician determines them to be necessary. Items may be repeated *once* if the student requests repetition, or if the clinician suspects they were not attending. However, items from *Conceptos y Siguiendo Direcciones, Recordando Oraciones,* Repetición de Numeros 1/2, and *Sequencias Familiares* 1/2 may not be repeated. It is important for the clinician to consider the child's behavior and attention across the assessment and allow for breaks as necessary to ensure optimal performance.

Short term memory deficits could also falsely indicate a speech and/or language disorder. Many of the test items require the child to hold several items in short term memory at once, then compare/analyze them and come up with a right answer (e.g. *Conceptos y Siguiendo Dirrectiones, Recordando Oraciones).* A child with limited short-term memory may perform poorly on standardized assessments due to the demands of the tasks. However, he may not need speech and language therapy but rather techniques and strategies to compensate for short-term or auditory memory deficits.

Motor/Sensory Impairments

In order for a child to participate in administration of this assessment, they must have a degree of fine motor and sensory (e.g. visual, auditory) abilities. If a child has deficits in any of these domains, their performance will be compromised. For example, for a child with vision deficits, if they are not using proper accommodations, they may not be able to fully see the test stimuli, and thus their performance may not reflect their true abilities. A child with motor deficits, such as a child with typical language development but living with cerebral palsy (CP), may find it much more frustrating and tiring to be pointing to/attending to pictures for an extended period of time than a typically developing non-disabled child. The child with CP may not perform at his highest capacity due to his motor impairments and would produce a lower score than he or she is actually capable of achieving. It is crucial that the examiner consider a child's motor/sensory limitations when administering the CELF-P2 Spanish to ensure the child is not falsely identified with a language disorder as a result of these impairments.

8. SPECIAL ALERTS/COMMENTS

The CELF-4 Spanish was designed to assess the presence of a language disorder or delay in Spanish speaking students aged 5;0-21;11. Subtests were designed to aid in determining a student's diagnosis, strengths and weakness, eligibility for services, and how their performance on tasks affects the student's access to the standard educational curriculum. Despite the CELF-4 Spanish's attempt to design a comprehensive language battery, results obtained from administration are not valid due to lack of information as to how tasks and items were deemed appropriate, and an insufficient reference standard. The insufficient reference standard in turn affects the diagnostic accuracy of the CELF-4 Spanish, including the sensitivity, specificity, and likelihood ratios, rendering these measures invalid. Therefore, even if the CELF-4 Spanish were a valid, reliable and unbiased assessment, it lacks sufficient discriminant accuracy in order to determine the presence or absence of a language disorder.

Items from the CELF-4 Spanish rely heavily on vocabulary dependent and labeling tasks. As a result, this test will likely identify socioeconomic status and second language acquisition issues, not a disorder or disability, in children learning English as a second language and those from homes of lower socioeconomic status. According to the Examiner's Manual, test administrators should be aware of a number of factors that may affect the performance of a student from diverse cultural and linguistic backgrounds. Some clinicians may choose to use items from the CELF-4 Spanish as probes to determine receptive and expressive language skills. In this case, modifications to the standard test administration may be used. Suggested modifications are available on page 20-22 of the Examiner's Manual. If such modifications are used however, normative test scores cannot be calculated. Clinical judgment must always be used to determine if performance was typical as compared to the student's peers in the student's speech community, NOT those from the normative sample as it is not *representative. Therefore, scores should not be calculated nor should they be used for classification or referral to special education services.*

According to the Examiner's Manual, "for an overall evaluation of a student's language ability, the results of the CELF-4 Spanish should be supplemented with a complete family and academic history, parent interview, results of other and informal measures, an analysis of a spontaneous language sample, the results of other linguistic and metalinguistic abilities tests, classroom behavioral observations, observations with peers, and evaluations of pragmatic and interpersonal communication abilities" (p. 18). One may question why the CELF-4 Spanish should be administered if the manual itself states that it should be supplemented with various other measures that require clinical judgment and essentially constitute a complete, more appropriate and valid evaluation.

Diagnosing children as language disordered or delayed and placing them in a special education program when they may not require services has many long lasting and detrimental consequences. These consequences may include a limited and less rigorous curriculum (Harry & Klingner, 2006), and lowered expectations which can lead to diminished academic and post-secondary opportunities (National Research Council, 2002; Harry & Klingner, 2006) and higher dropout rates (Hehir,2005).

Due to cultural and linguistic biases (for example, exposure to books, repetition of unfamiliar syntactic structures, cultural labeling practices, communication with strangers, responses to known questions, etc.) and assumptions about past knowledge and experiences, this test should only be used to probe for information and not to identify a disorder or disability. Therefore, scores should not be calculated and used as the determinant of classification or referral to special education services.

REFERENCES

American Speech-Language-Hearing Association. (2004). Knowledge and skills needed by speech-language pathologists and audiologists to provide culturally and linguistically appropriate services [Knowledge and Skills]. Available from www.asha.org/policy.

Betz, S. K., Eickhoff, J. R., & Sullivan, S. F. (2013). Factors influencing the selection of test for the diagnosis of specific language impairment. *Language, Speech, and Hearing Services in Schools, 44,* 133-146.

Dollaghan, C. (2007). *The handbook for evidence-based practice in communication disorders*. Baltimore, MD: Paul H. Brooks Publishing Co.

Dollaghan, C., & Horner, E. A. (2011). Bilingual language assessment: a meta-analysis of diagnostic accuracy. *Journal of Speech, Language, and Hearing Research, 54,* 1077-1088.

Guadagnoli, E. and Velicer, W. F. (1988). Relation to sample size to the stability of component patterns. Psychological Bulletin, 103, 2, 265-275. doi: 10.1037/0033-2909.103.2.265

Hart, B & Risley, T.R. (1995) *Meaningful Differences in the Everyday Experience of Young American Children.* Baltimore: Paul Brookes.

Harry, B. & Klingner, J., (2006). *Why are so many minority students in special education?: Understanding race and disability in schools.* New York: Teachers College Press, Columbia University.

Hehir, T. (2005). New directions in special education: Eliminating ableism in policy and practice. Cambridge, MA: Harvard Educational Publishing Group

McCauley, R. J. & Swisher, L. (1984). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders, 49*(1), 34-42.

New York City Department of Education (2009). Standard operating procedures manual: The referral, evaluation, and placement of school-age students with disabilities. Retrieved from http://schools.nyc.gov/nr/rdonlyres/5f3a5562-563c-4870-871f bb9156eee60b/0/03062009sopm.pdf.

National Research Council. (2002). *Minority students in special and gifted education.* Committee on Minority Representation in Special Education. M. Suzanne Donovan and Christopher T. Cross (Eds.), Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Paul, R. (2007). *Language disorders from infancy through adolescence (3rd ed.).* St. Louis, MO: Mosby Elsevier.

Paradis, J. (2005). Grammatical morphology in children learning English as a second language: Implications of similarities with Specific Language Impairment. *Language, Speech and Hearing Services in the Schools*, 36, 172-187.

Paradis, J., Genesee, F., & Crago, M. B. (2011). Dual language development & disorders: A handbook on bilingualism & second language learning (2nd ed.). Baltimore, MD: Paul H. Brookes.

Peña, E., & Quinn, R. (1997). Task familiarity: Effects on the test performance of Puerto Rican and African American children. *Language, Speech, and Hearing Services in Schools*, 28, 323–332.

Peña, E.D., Spaulding, T.J., & Plante, E. (2006). The Composition of Normative Groups and Diagnostic Decision Making: Shooting Ourselves in the Foot. American Journal of Speech-Language Pathology, 15, 247-254.

Plante, E. & Vance, R. (1994). Selection of preschool language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools, 25,* 15-24.

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2010). *Assessment in special and inclusive education (11th edition).* Belmont, CA: Wadsworth Cengage Learning.

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals (4th ed.) [CELF-4 Spanish]*. San Antonio, TX: PsychCorp.

Shakespeare, W. (2007). *Hamlet.* David Scott Kastan and Jeff Dolven (eds.). New York, NY:

Barnes & Noble.

Spaulding, T. J., Plante, E., & Farinella, K. A. (2006). Eligibility criteria for language impairment: is the low end of normal always appropriate? Language, Speech, and Hearing Services in Schools, 37, 61-72.