Test Review:
Clinical Evaluation of Language Fundamentals Preschool – Second Edition (CELF-P2) Spanish

---

Version: 2nd Edition
Copyright date: 2009
Grade or Age Range: 3 years through 6 years, 11 months
Author: Elisabeth H. Wiig, Ph.D.;Wayne A. Secord, Ph.D., and Eleanor Semel, Ed.D.
Publisher: Pearson

Table of Contents

1. PURPOSE

The CELF-P2 Spanish is designed to assess the presence of a language disorder or delay in Spanish speaking students aged 3;0-6;11. Subtests were designed to aid in determining a student's diagnosis, strengths and weaknesses, and eligibility for services. The CELF-P2 Spanish contains a four-level assessment process and was designed as a parallel, not translated, version of the CELF-P2. The presence of a language disorder can be determined by calculating a Core Language score using only three subtests, and additional subtests aid in gaining more information regarding the nature and extent of the disorder. Content areas include: morphology, syntax, semantics, and pragmatics.

2. DESCRIPTION

Subtest	Age Range	Purpose	Format
Conceptos Básicos [1, 2] (Basic Concepts) [1, 2]	3-6	To assess the child's understanding of concepts such as direction/location, sequence, size, number, quantity, and same/different.	The student points to a picture that represents an orally presented sentence.
Estructura de palabras [1, 2, 3] (Word Structure[1, 2, 3])	3-6	To determine the student's use of morphological rules.	The student completes an orally presented sentence in reference to a visual stimulus.
Recordando oraciones [1,2, 3] (Recalling Sentences [1,2, 3])	3-6	To measure a student's ability to recall and imitate sentences of variable length and complexity.	The student repeats sentences orally presented by the administrator.
Vocabulario expresivo (Expressive Vocabulary)	3-6	To measure a student's ability to use referential naming.	The student identifies an object, person, or action presented by the administrator.
Estructura de oraciones (Sentence Structure)	3-6	To measure a child's expressive language ability to formulate grammatically correct sentences.	Following an orally presented sentence, the student points to the corresponding stimulus image.

Conceptos y Siguiendo Direcciones [3] (Concepts and Following Directions[3])	4-6	To determine 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.	Identification of pictured objects following oral directions from test administrator.
Clases de palabras (Word Classes)	5-6	To measure an individual's ability to comprehend and explain relationships between images or orally presented target words.	Given 3-4 words, the student selects two words that go together and explains their relationship.
Conocimiento fonológico (Phonological Awareness)	4-6	To measure a student's acquisition of sound structure and ability to manipulate sound through: (a) Syllable segmentation and blending (b) Phoneme identification.	Comprised of 5 tasks of varying directives.

[1] Core Language Score (3 year olds)
[2] Core Language Score (4 year olds)
[3] Core Language Score (5-6 year olds)

The CELF-P2 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 the examiner does not have near-native proficiency, "the test can be administered in collaboration with a trained and qualified interpreter" (Wiig, Secord, & Semel, 2009). The manual cautions that the SLP should be the person "responsible for the process and outcome of the assessment" (p. 19). This includes summarizing and interpreting results of the assessment, and planning for intervention if necessary. Specific information regarding using the test with interpreters is provided in Chapter 2 of the Examiner's Manual.

3. **STANDARIZATION SAMPLE**

The standardization sample for the CELF-P2 Spanish used data collected from November 2007 – May 2008 and was comprised of a sample of 464 individuals from across the United States and Puerto Rico aged 3;0 through 6;11 years. Inclusion into the standardization sample required completion of the test in a standard manner (e.g. no sign language was permitted). The students were also required to speak Spanish to communicate and have no current diagnosis of a behavioral or emotional disorder. The standardization sample was stratified by demographic factors including age, gender, race, parental education level, and geographic location as compared to the 2009 national census for the Hispanic population of the USA for children aged 3;0-6;11. It should be noted that although the test is intended for Spanish speaking children, 122 out of 464 children were reported to be bilingual (p. 153) and no information regarding what was considered bilingual was provided in the manual. Further, the following percentages of the sample reportedly never spoke Spanish in the following contexts: 48% with friends, 14.2% in the classroom, and .2% with caregivers. So, although a child may be reportedly "bilingual" they may only speak Spanish with their family or in their community, and not in other contexts, such as with their friends. This reduces the validity of the test as children included in the standardization sample do not match the intended test population.

Approximately 42% of the sample reported receiving special services, including 26.3% for second language services (e.g. ESL, LEP), and 22.6% for bilingual education; 6.5% of the sample reportedly was receiving Speech and Language services and less than 1% of the sample was receiving Occupational Therapy and Physical Therapy. The manual did not state whether or not the students who were previously identified as having a disability were accurately identified by the CELF-P2 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 (TD) 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). Content validity was analyzed using a literature review and feedback from SLPs across the country, an expert panel to review item content, and pilot and standardization data. Pilot studies were conducted from January 2007 – March 2007 on seven measures: *Conceptos Básicos, Estructura de Palabras, Recordando Oraciones, Vocabulario Expresivo, Estructura de Oraciones, Clases de Palabras,* and *Conocimiento Fonológico*. It should be noted that this does not include *Conceptos y Siguiendo Direcciones,* which is

included in the CELF-P2 Spanish. The sample included 89 children aged 3;0-6;11 from across the USA and Puerto Rico. In addition, a clinical study was conducted with 47 children aged 3;0-6;11 who had been previously identified as having a language disorder and were receiving language services at the time of testing. All the children in both groups lived in Spanish speaking homes and spoke Spanish fluently enough to take the test in a "standard manner" (p. 144). A scoring panel analyzed responses obtained during the tryout test and clinical studies. If responses differed from the target response but were determined to be accurate by the panel, they were added to the scoring criteria to increase sensitivity to linguistic variation.

An expert panel comprised of seven SLPs from the USA and Puerto Rico reviewed the test items and materials of the CELF-P2 Spanish to identify biases, clinical utility, and cultural and content relevance. Following the panel member's feedback, items that were considered biased were edited, rewritten, or eliminated.

The content validity of the CELF-P2 Spanish is insufficient for several reasons. First, regarding the pilot study, although the test intends to identify a language disorder in Spanish speaking children, 16 out of 89 children in the pilot study and 5 out of 47 children in the clinical study were "bilingual." Further, the manual failed to mention how the children in the typical group were determined to be typically developing, and thus we cannot be certain of their diagnostic status. Regarding the expert panel, specific information was not provided regarding the background and training of the panel members. 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. Therefore, one cannot be confident that the items in this test are completely free of bias.

Construct – Construct validity assesses whether the test measures what it purports to measure (Paul, 2007). It was measured using special group studies comprised of typically developing children and children with language disorders. The diagnosis of these students was compared with their status as determined by the CELF-P2 Spanish to determine the test's diagnostic accuracy.

Reference Standard

In considering the diagnostic accuracy of an index measure such as the CELF-P2 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 sensitivity group included 90 children, ages 3-6, who were identified and tested by speech language pathologists in the United States, Puerto Rico, and Mexico. However, there are several causes for concern regarding the reference standard used to identify the sensitivity group. First, the manual is not clear if the students in the sample included those through 6;0 or 6;11. If students from 6;0-6;11 were not included, then the reference standard is not representative of the test population as it did not include the entire age range. Further, the tests that were used to identify the children were not identified in the manual. Therefore, we cannot be sure if they are able to accurately discriminate between TD children and children with a language disorder. Additionally, arbitrary cut off scores on standardized language tests, such as 1.5 SD from the mean, often do not accurately discriminate between TD children and children with a language disorder (Spaulding, Plante, & Farinella, 2006). Due to the reasons mentioned above, the reference standard used for the sensitivity group is considered insufficient.

The reference standard used to identify the *specificity* group was a control group of 90 TD students who were matched to the sensitivity group based on age, gender, place of origin, and level of parental education. The manual did not include information regarding what reference standard was used or how the children were identified as TD. The reference standard for the specificity group is considered insufficient for several reasons. First, similar to the sensitivity reference standard, if the sample did not include children from 6;0-6;11, it is not representative of the test population. Further, according to Dollaghan (2007) performance on the reference standard cannot be assumed. The same reference standard (a score above 1.5 SD below the mean on a standardized test) was not applied to both the specificity and sensitivity groups. 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).

Sensitivity and Specificity

Sensitivity is the proportion of students who actually have a language disorder and are accurately identified as such on the assessment (Dollaghan, 2007). For example, sensitivity means that an eight-year-old boy previously diagnosed with a language disorder will score within limits to be identified as having a language disorder on this assessment. According to Plante & Vance (1994), validity measures above .9 are good, measures between .8 and .89 are fair, and measures below .8 are unacceptable. The

CELF-P2 Spanish reports the sensitivity to be .86, which is considered to be "fair" by the standards in the field. However, it should be noted that the sensitivity measure does not provide a complete picture of the test's accuracy. As mentioned above, the content validity and reference standards were determined to be insufficient. Since the reference standard is not a valid measure for distinguishing between disordered and typical individuals, the user cannot be confident that the sensitivity measure reported is accurate. These issues decrease the overall accuracy of the CELF-P2 Spanish. In addition, even if a specificity of .86 reflected the test's true accuracy, 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-P2 Spanish, and therefore will not receive the extra academic and language support that they need.

Specificity is 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 an eight-year-old boy with no history of a language disorder will score within normal limits on the assessment. The CELF-P2 Spanish reports specificity measures to be 0.89, or "fair" as compared to the standard in the field. However, the user cannot know if this is a true reflection of the CELF-P2 Spanish's specificity since the reference standard (score below 1.5 SD on a standardized language assessment) was not applied to the specificity group. As a result, their true diagnostic status is unknown and the children in the specificity sample cannot be guaranteed to be free of a disorder. Due to the uncertainty of the true diagnostic status of the children in the specificity group (language disordered vs. typically developing) the specificity measure does not provide a complete picture of the test's accuracy. As mentioned above, the content validity and reference standards were determined to be insufficient, thus decreasing the overall accuracy of the CELF-P2 Spanish. Finally, if we assume that .89 is a true measure of the CELF-P2 Spanish's specificity, it is also important to consider the implication of this measure. A specificity of .89 means that 11/100 typically developing children will be identified as having a language disorder and may be unnecessarily referred for special education services.

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. These measures are preferred to sensitivity and specificity in determining diagnostic accuracy as the sensitivity and specificity are susceptible to changes in the base rate of the standardization sample, or the percentage of students in the sample who have the disorder. The lower the base rate is in a sample, the fewer people there are who are affected. Therefore, the specificity will be higher because there is a higher probability that the individual is unaffected (Dollaghan, 2007). Likelihood ratios are less affected by

changes to the base rate. 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 the lack of a valid reference standard used when determining sensitivity and specificity values.

Overall, construct validity, was determined to be insufficient for several reasons. First, the reference standard consisted of several assessments not mentioned in the manual and thus we cannot be certain of their diagnostic accuracy or the true diagnostic status of the individuals used in the special group studies. Also, the sensitivity and specificity groups were created based on arbitrary cut off scores that may not necessarily discriminate typically developing children from children with disabilities (Spaulding, Plante, & Farinella, 2006). The sensitivity and specificity were also determined to be insufficient due to the issues mentioned above regarding the reference standard and how children were determined to be typically developing versus language disordered.

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-P2 Spanish to the CELF-P2, CELF-4 Spanish, and the PLS-4 Spanish. This is strange, however, because these tests are intended for vastly different populations. See the targeted populations for the index and comparison tests in the table below.

Test	Language of Target Population	Age of Target Population
CELF-P2 Spanish	Spanish	3;0-6;11
CELF-P2	English	3;0-6;11
CELF-4 Spanish	Spanish	5;0-21;11
PLS-4 Spanish	Spanish	Birth – 7;11

As shown above, the target population of the index measure does not match that of any of the comparison tests, thus rendering them inappropriate comparison measures. Five children were administered both the CELF-P2 (English) and the CELF-P2 Spanish. 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).

According to the Examiner's Manual, "given that the subtests in the CELF-P2 Spanish parallel those in the CELF-P2, it is anticipated that a bilingual child would perform similarly on the 2 tests" (Wiig, Semel, & Secord, 2009, p. 186). This makes *vast* assumptions about language acquisition and proficiency. Bilingual children will likely not exhibit equal skills in Spanish and English and thus it cannot be assumed that their performance on standardized tests of language in the two languages would be equal. The children were reported to understand some concepts in Spanish and some concepts in English and to converse fluently in both languages. Specific information regarding the correlation between the tests was not provided; the manual only reports, "the data indicates a positive correlation between the 2 tests… [for example] when a child's composite score on the CELF-P2 Spanish indicated poor performance, his or her composite score on the CELF-P2 was also poor (CELF P2 Spanish RLI = 75, CELF P2 RLI = 65). It is important to consider these scores in the context of a standard error of measurement (see section 6: *Standard Error of Measurement* for more information). Given the SEM, the test user can only determine that the child's score falls within a range of possible scores. Thus, by reporting that the scores exhibited similar trends, considering the SEM, we gain very little information regarding the child's performance in either language.

Information regarding the sample group and correlations between the comparison measures are reported in the table below.

Comparison Measure	Description of the Sample	CLI	RLI	ELI
CELF-4 Spanish	30 TD children (ages 5-6 years)	.87	.79	.65
PLS-4 Spanish	36 TD children	.6	.48	.56

Concurrent validity "requires that the comparison test be a measure that is itself valid for a particular purpose" (APA, 1985, as cited in Plante & Vance, 1994). Correlations between the subtests for CLI, RLI, and ELI ranged from .48 to .87. In order to examine the validity of a comparison measure, and thus the concurrent validity, its discriminant accuracy, or its sensitivity and specificity, must also be considered. The comparison tests used to demonstrate concurrent validity of the CELF-P2 Spanish have similar validity issues as those described for the CELF-P2 Spanish above. Therefore, they are considered insufficient measures of concurrent validity. The sensitivities and specificities for the comparison tests are reported in the table below. According

to Plante & Vance (1994), validity measures above .9 are good, measures between .8 and .89 are fair, and measures below .8 are unacceptable. Even if the comparison tests were themselves valid, as shown in the table above, none of the comparison measures achieve consistently "good" measures of sensitivity and specificity.

	Sensitivity	Specificity
CELF-P2	.6-.85	.82-.95
CELF-4 Spanish	.87-1	.82-.96
PLS-4 Spanish	.83-.91	.56-.68

Ranges of values represent sensitivity and specificity data from across subtests and cutoff scores.

Overall, concurrent validity for the CELF-P2 Spanish is considered insufficient due to inappropriate index measures (i.e. tests with different target populations), lack of inclusion of the entire age range of the index measure, lack of information regarding how the children were determined to be TD, and insufficient psychometric data (sensitivity and specificity) of the comparison 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 customer put on three oranges and they weighed one pound. If she weighed the same three oranges multiple times, and each time they weighed one pound, the scale would have *test-retest reliability*. If other customers in the store put the same three oranges on the scale and they still weighed 1 pound, the scale would have *inter-examiner reliability.* It is a reliable measure. Now say an official were to put a one pound calibrated weight on the scale and it weighed two 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. Standardized tests often report high measures of reliability while choosing not to report or emphasize the lack of validity in order to present the test as an accurate measure of language. However, as you can see, reliability does not equal accuracy.

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 administering the test at two separate times (2-29 days; mean = 10 days) to 66 individuals from the standardization sample. Children from across the age range of the test were used for this study. 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, 1997; Magnusson, 1967, as cited in Wiig, Semel, & Secord, 2009). According to the Examiners Manual, across ages, corrected reliability coefficients for CLS, RLI, and ELI ranged from .81 to .95. Considering the coefficients of the individual subtests across ages, coefficients ranged from .68 to .96. Only 3 out of 26 subtest coefficients met the standard to be considered acceptable. Thus, test-retest reliability is considered 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 5 trained raters. During standardization testing, each subtest that requires subjective scoring (Estructura de palabras, Recordando oraciones, Vocabulario expresivo and Clases de palabras-expresivo) was rated independently by two raters and then compared. A third rater resolved any discrepancies. According to the Examiners Manual, agreement between scorers ranged from .93 to .98 across the four selected subtests. Although this indicates a high level of reliability, "a high degree of reliability alone does not ensure validity" (McCauley & Swisher, 1984, p. 35) and, as described above, the CELF-P2 Spanish did not achieve any type of validity considered acceptable within the field.

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 the split half method. 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 (3-6;11) and subtests, average coefficients ranged from .79- to .97; 9 out of 12 subtests did not meet the standard of .9 for reliability (Salvia, Ysseldyke, & Bolt, 2010). Inter-item consistency was also calculated for 90 students who were diagnosed with a language disorder. Across subtests, the split-half coefficient ranged from .80-.98; 6 out of 12 subtests did not meet the minimum standard (Salvia, Ysseldyke, & Bolt, 2010). As the range of coefficients for TD and LD children were similar, the CELF-P2 Spanish has comparable reliability for measuring language skills of clinical groups and TD children. However, at least half of the subtests did not meet the minimum standard and thus the inter-item consistency is insufficient.

Overall, the reliability, including the test-retest, and inter-examiner reliability, is considered insufficient. Specifically, across ages for the range of the test, only 3 out of 26 subtests received sufficient test-retest reliability. Further, for inter-item consistency, 9/12 subtests did not meet the standard for TD children and 6/12 subtests did not been the standard for the clinical group.

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 actual 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 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 CIs 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, p. 52). The clinician chooses a confidence level (usually 90% or 95%) at which to calculate the CI. A higher confidence level will yield a larger range of possible test scores, so as to be more likely to include the child's true range of possible scores. Although a larger range is yielded with a higher CI, 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 CI 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 4;6-4;11 on the CELF-P2 Spanish, the SEM at the 90% confidence level is +/- 7 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 68 and 82. Thus, all the clinician can determine from administration of the CELF-P2 Spanish is that this child's true language ability (according to the CELF-P2 Spanish) ranges from moderately-severely impaired to low average. 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.

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-P2 Spanish contains inherent biases against culturally and linguistically diverse children.

Linguistic Bias

Bilingual Speakers

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-P2 Spanish may be students who are learning English as a second language in school (and many students from the standardization sample were receiving second language services of some type). 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-P2 Spanish, they may be falsely identified as having a language disorder when, in reality, they are experiencing subtractive bilingualism.

On the CELF-P2 Spanish, children learning English as a second language may be falsely identified as having a language disorder on subtests of the CELF-P2 Spanish 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. These 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-P2 Spanish, they may be falsely identified as having a language disorder.

Dialectal Variations

A child's performance on the CELF-P2 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: Central and South America, Cuba, Dominican Republic, Mexico, Puerto Rico, 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-P2 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. It should also be noted that in the Examiner's Manual, additional responses for select subtests are suggested (*Estructura de Palabras, Vocabulario Expresivo, Clases de Palabras*). According to the manual, responses on the Folleto de Registro reflect responses that were provided a minimum of 10% of the time. However, additional responses may be considered correct, especially from speakers of other dialects. Although providing additional responses increases scoring sensitivity, it is impossible to provide a complete list of responses that would be considered correct. It is crucial that the examiner consider the child's dialectal background and current speech community in scoring subtests of the CELF-P2 Spanish to ensure that a child is not falsely identified with a language disorder as a result of dialectal variations.

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 (SES); 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 higher performance on standardized child language tests, which often place significant emphasis on vocabulary-based tasks. 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-P2 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 hueso [bone], or huellas [footprints]. Further, the *Conceptos Básicos* and *Clases de Palabras* subtests require prior knowledge of words in the target to provide an appropriate response (e.g. lleno [full], vacío [empty]). As a result of vocabulary items on the CELF-P2 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-P2 Spanish may also be affected by their prior knowledge and experiences. For example, a child from a low SES family may not have prior experience with the word *cámara* [*camera*] in the *Vocabulario Expresivo* subtest; a child who has never been a part of organized school or activities may not be familiar with *una fila* [*a line*] in the *Conceptos Básicos* 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-P2 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 guitar as "para tocar musica" as they have been exposed to function-type description tasks rather than labeling the object itself.

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-P2 Spanish requires children to attend to the test book for the length of the assessment, which 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. Therefore a child without prior educational experience may have difficulty with tasks requiring this skill, such as the *Conocimiento fonológico* subtest.

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 information 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. 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 is 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. Performance on subtests administered later in the session may be affected by the child's fatigue. 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. According to the manual, "administration time for the 3 subtests required to determine the core language score [*Conceptos Basicos, Estructura de Palabras, Recordando Oraciones*] is 15-20 minutes, depending on the age and responsiveness of the child" (p. 1). However, the clinician should be aware that administration time may take longer for a preschool child, especially if more subtests are administered. 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 and *Recordando Oraciones* 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 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 on these subtests due to the demands of the tasks. 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 fully 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 disabled child who is not physically disabled. 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-P2 Spanish is designed to assess the presence of a language disorder or delay in Spanish speaking students aged 3;0-6;11. Subtests were designed to aid in determining a student's diagnosis, strengths and weaknesses, and eligibility for services. The CELF-P2 Spanish contains a four-level assessment process and was designed as a parallel, not translated, version of the CELF-P2. Despite the CELF-P2 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-P2 Spanish, including the sensitivity, specificity, and likelihood ratios, rendering these measures invalid. Therefore, even if the CELF-P2 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.

As this test contains tasks that rely on vocabulary and prior knowledge and experience, for bilingual children and children from lower income homes, test scores will reflect differences due to SES and second language acquisition, not a true disorder or disability. 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-P2 Spanish as probes to determine receptive and expressive language skills. In this case, modifications to the standard test administration may be used. If such modifications are used, normative test scores cannot be calculated, and clinical judgment must always be used to determine if performance was typical as compared to the student's true peers. 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 child's language ability, the results of the CELF-P2 Spanish should be supplemented with a complete family and academic history, caregiver interview, analysis of a spontaneous language sample, classroom behavioral observations, observations of peer interactions, evaluations of pragmatic and interpersonal communication abilities, and the result of other linguistic and metalinguistic abilities tests" (p. 28). One may question why the CELF-P2 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 assessment.

Diagnosing children as language disordered or delayed and placing them in a special education program when they may not require the 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 (e.g. exposure to books, dialectal variations, cultural labeling practices, communication with strangers, responses to known questions), 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 sole 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.

Crowley, C. (2010) A Critical Analysis of the CELF-4: The Responsible Clinician's Guide to the CELF-4. Dissertation.

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.

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.

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

Wiig, E. H., Secord, W. A., & Semel, E. (2009). *Clinical Evaluation of Language Fundamentals- Preschool, Spanish edition (2nd ed.). [CELF-P2 Spanish].* San Antonio, TX: Pearson.