LEADERS

Test Review:
Clinical Evaluation of Language Fundamentals – Fifth Edition (CELF-5)

Version: 5th Edition
Copyright date: 2013
Grade or Age Range: 5-21
Author: Elizabeth Wiig, Eleanor Semel and Wayne Secord
Publisher: Pearson

Table of Contents

1. **PURPOSE**
The Clinical Evaluation of Language Fundamentals (CELF-5) was designed to assess a student’s language and communication skills in a variety of contexts, determine the presence of a language disorder, describe the nature of the language disorder and plan for intervention or treatment. The CELF-5 is a comprehensive and flexible assessment procedure. The test identifies a student’s language strengths and weaknesses and can be used to determine eligibility for services, plan “curriculum relevant treatment,” recommend classroom language adaptations or accommodations and provide performance-based assessment that corresponds to educational objectives.

2. **DESCRIPTION**

The CELF-5 consists of a number of tests. Each test can be administered as an independent test and is designed to assess specific language skills. More detailed information regarding each test is listed in Table 1.

Table 1. CELF-5 Tests [in appendix]

TEST	**Age Range**	**Purpose**	**Format**
Observational Rating Scale (ORS)	5-21	Systematic observation of a student’s listening, speaking, reading and writing skills in the classroom and at home. Identifies situations where reduced language performance occurs.	Multiple raters (e.g. teachers, parents/ caregivers etc.) complete a form rating student’s classroom and home interaction and communication skills according to how frequently the behavior occurs. Examiner summarizes the raters’ responses.
Sentence Comprehension	5-8	Measures comprehension of grammatical rules at the sentence level.	Following an orally presented stimulus, the student points to the corresponding stimulus image.
Linguistic Concepts	5-8	Measures understanding of linguistic concepts, including comprehension of logical operations or connectives.	Following oral directions that contain embedded concepts, the student points to a corresponding image.
Word Structure	5-8	Measures the acquisition of English morphological rules.	The student completes an orally presented sentence in reference to visual

			stimuli.
Word Classes	5-21	Measures the ability to understand relationships between associated words.	Given 3-4 orally presented words or visually presented pictures, student selects the two words that are most related.
Following Directions	5-21	Measures the ability to interpret, recall and execute oral directions of increasing length and complexity, remember the names, characteristics and order of objects.	Following oral directions, the student points to correct shapes in order in the stimulus book.
Formulated Sentences	5-21	Measures the ability to formulate semantically and grammatically correct sentences of increasing length and complexity.	Student formulates a sentence about a picture using 1-2 target words presented orally by the examiner.
Recalling Sentences	5-21	Measures the ability to recall and reproduce sentences.	Student imitates orally presented sentences of increasing length and complexity.
Understanding Spoken Paragraphs	5-21	Measures the ability to interpret factual and inferential information.	Following oral presentation of a paragraph, student answers questions targeting the paragraph’s main idea, details, sequencing and inferential information.
Word Definitions	9-21	Measures the ability to define word meanings by describing features of the words.	Following oral presentation of a sentence, student defines the target word used in the sentence.
Sentence Assembly	9-21	Measures the ability to assemble words and word combinations into grammatically correct sentences.	Following presentation of visual or oral word combinations, the student produces syntactically and semantically correct sentences.
Semantic Relationships	9-21	Measures the ability to interpret sentences that	Following presentation of an oral stimulus, the

		include semantic relationships.	student selects 2 correct choices from 4 visually presented options that answer a target question.
Pragmatics Profile	5-21	Provides information regarding development of verbal and non-verbal social communication.	A 4-point Likert scale questionnaire, completed by examiner or parent/caregiver.
Reading Comprehension	8-21	Measures the ability to interpret information presented in written paragraphs.	The student reads a written paragraph and then answers questions presented orally targeting the paragraph's main idea, details, sequencing and inferential information.
Structured Writing	8-21	Measures the ability to interpret written sentences to complete a story.	Student writes a short story by completing a sentence and writing one or more additional sentence(s).
Pragmatics Activities Checklist	5-21	Provides information related to student's verbal and non-verbal social interactions	The examiner completes a checklist about their interaction with the student as observed during formal testing and selected activities.

3. STANDARIZATION SAMPLE

The standardization sample was based on the March 2010 US Census Update and was stratified by age, sex, race/ethnicity, geographic region, and parent education level. Inclusion into the sample required completion of the test in the standard oral manner (e.g., didn't need sign language). Of the 3,000 participants, 20% were bilingual, 27% spoke a dialect other than Standard American English (SAE), 4% were gifted or talented, 11% had diagnoses including but not limited to attention deficit hyperactivity disorder (ADHD), learning disability (LD), intellectual disability (ID), pervasive developmental disorder (PDD), Down Syndrome, cerebral palsy, developmental delay, or emotional disturbance, 12% were diagnosed with speech and/or language disorders, and 3% were receiving occupational or physical therapy. The manual did not state how the students classified as having a disability were identified. According to Peña, Spaulding and Plante (2006), inclusion of children with disabilities in the normative sample can negatively impact the test's discriminant accuracy, or ability to differentiate between typically developing and disordered children. Specifically, inclusion of individuals with disabilities in the normative sample lowers the mean score, which limits the tests ability to diagnose children with mild disabilities.

4. VALIDITY

Content - Content Validity is how representative the test items are of the content that is being assessed (Paul, 2007). Content validity was determined in a variety of ways, including: literature review; users' feedback; expert review; pilot studies and response process. Content construction was designed to ensure adequate sampling of various language domains (Technical Manual, p. 52). Three pilot studies were conducted to determine test modifications, evaluate effectiveness of revisions from the CELF-4, improve test floors and ceilings and improve visual stimuli. The pilot study sample consisted of 195 students in three age groups (4-6 years, 8 years and 9-16 years) and included 102 females and 93 males. Pilot studies determined adaption of subtests into tests, elimination of subtests and addition of new tests to meet the goals of the CELF-5 revision. National tryout studies were conducted by 154 Speech-Language Pathologists to determine appropriateness of content revisions and determine scoring rules. CELF-5 pilot and tryout items were reviewed by a panel of speech pathologists from across the country with "expertise in assessment of diverse populations" to minimize cultural and linguistic biases in test content (Technical Manual, 22).

Several factors contribute to lack of content validity for the CELF-5. First, there is a lack of information regarding how individuals who participated in the pilot and try out studies were identified as typically developing or language impaired. The pilot sample also used sample sizes smaller than what is considered acceptable in the field. In addition, information regarding the panel's level of expertise was not provided. ASHA (2004) has described the knowledge and skills needed to provide culturally and linguistically appropriate services, but whether the panel has that level of expertise is not described. As a result, the expert review panel may have been limited in its ability to accurately assess the test content for bias.

Construct – Construct validity assesses the extent to which a test can be used for as a specific purpose, such as to identify children with a language disorder (Vance & Plante, 2004). The authors of the CELF-5 measured construct validity using a study of students diagnosed with and without language disorders.

Reference Standard

In considering the diagnostic accuracy of an index measure such as the CELF-5 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 at 1.5 SDs or below on a standardized language test. The study included 67 children, recruited from Speech-Language Pathologists in multiple centers across the United States ranging in age between 5;0 – 15;11. It is important to note that this does not include the entire age range of the CELF-5, and thus is not representative of the test population. According to the APA (2004) these samples are too small to be considered representative and do not meet the minimum standard of 100 per age group. Dollaghan (2007) argues that the bigger the sample size, the more power it yields to detect differences between groups. With small sample sizes, particularly with young children there is a high chance of false negatives and misdiagnoses. The standardized tests that were used to identify children as language disordered included the CELF-4 (49%), CELF-P2 (7.5%), Test of Language Development (TOLD) primary or intermediate (8%), PLS-3 (17.9%) and Oral and Written Language Scales (OWLS) (13%) and Comprehensive Assessment of Spoken Language (CASL) (4.5%). Over half of the students were identified using the CELF-4 and PLS-3, both of which have been identified as instruments with unacceptable diagnostic accuracy (Plante & Vance, 1994). In addition, according to Spaulding, Plant and Farinella (2006), the use of arbitrary cut scores on standardized tests does not accurately distinguish children with a language disorder from children who are typically developing. Therefore, the true diagnostic status of these children is unknown and their inclusion in the reference standard is based on unacceptable measures. Therefore, the diagnostic accuracy of the CELF-5 is subject to potential spectrum bias, which occurs when "diagnostic accuracy is calculated from a sample of participants who do not represent the full spectrum of characteristics" (Dollaghan & Horner, 2011). The reference standard is insufficient because it does not include the entire age range of the CELF-5 and students included had unknown diagnostic status.

The reference standard used to identify the specificity group was no previous referral for speech and language services, matched to the sensitivity group, selected from the normative sample. The reference standard does not include students from the entire age range of the index measure and is not representative of the population. Students were classified as typically developing if they had not previously been diagnosed with a language disorder and were not currently receiving speech and language services. This does not meet the standards set forth by Dollaghan (2007) who states that a reference standard must be applied to the sensitivity and specificity groups, in order to determine the test's discriminant accuracy. According to Dollaghan (2007), "the reference standard and the index measure both need to be described clearly enough that an experienced clinician can understand their differences and similarities and can envision applying them" (p. 85). Therefore, the reference standard used for the specificity group is not a valid measure.

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 when given the CELF-5, Johnny, an eight-year-old boy previously diagnosed with a language disorder, will score within the limits to be identified as having a language disorder on this assessment. 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 Peter, an eight-year-old boy with no history of a language disorder, when he is given the CELF-5 will score within normal limits on the assessment.

No test is 100% accurate in its discriminant accuracy—that is the test's ability to accurately distinguish between children with and without language disorders. Vance and Plante (2004) set forth the standard used to determine whether a test is "accurate enough." That standard is as follows: a test that accurately identifies children with language disorders and those without language disorders is considered "good" if it is 90% to 100% accurate; "fair" if it is accurate 80 to 89 percent of the time. Less than 80% accuracy in identifying disorder, or specificity which is absence of disorder, is considered "unacceptable" because such a high rate of misdiagnosis can lead to serious social consequences.

The CELF-5 reports sensitivity and specificity measures at 4 cut scores: 1 SD; 1.3 SD; 1.5 SD and 2 SD below the mean. At 1, 1.3 and 1.5 SD below the mean sensitivity and specificity range from fair to good according to Plante and Vance (1994). According to the Technical Manual, the optimal cut score is 1.3 SD below the mean as this best balances sensitivity and specificity values and results in sensitivity and specificity of .97, which is good according to the standards in the field. A sensitivity of .97 means that only 3% of children with a language disorder will not be diagnosed as such and specificity of .97 means 3% of children who do not have a language disability will be identified as such and referred for special education services.

It must be noted that the sensitivity group included only 67 children ranging from 5;0 to 15;11. This is a very small group to rely upon. Also, the only requirement to be included in the sensitivity group is that each of the 67 children had to score below 1.5 Standard Deviations below the mean on any standardized language test. This means that the 67 children in the sensitivity group could all have had severe disabilities. They might have multiple disabilities in addition to severe language disorders including severe intellectual disabilities or Autism Spectrum Disorder making it easy for a language disorder test to identify this group as having language disorders with extremely high accuracy. The few numbers of students with disorders in the sensitivity group and the lack of information on the severity and kinds of disabilities of those 67, makes it hard to rely upon, and trust, the high sensitivity numbers offered in the CELF-5.

It is important to emphasize that at two standard deviations (2SD) below the mean, the CELF-5 is only 57% accurate in identifying children with language disorders as having language disorders. For those districts—and even state regulations--that continue to require performance below two standard deviations below the mean, the CELF-5 will correctly identify children with language disorders with only about as much accuracy as a flip of a coin.

Base rate must also be considered. Base rate refers to the number of affected individuals within a sample, and is important to consider when assessing sensitivity and specificity values. For example, if there are only a few affected individuals, the specificity will be higher because there is a higher probability that the individual is unaffected (Dollaghan, 2007). For example, if one applies a cut score of 1.3 standard deviations below the mean and a base rate of 70%, there is a 7% chance that a child with a language disorder will be identified as typically developing (false negative). With a base rate of 80% and a cut score of -1.3 SD below the mean, the CELF-5 has an 11% false negative rate.

An additional serious concern with the CELF-5's construct validity analysis has to do with the test used to identify the sensitivity and specificity groups—which is called the "reference standard." According to Dollaghan (2007), sensitivity and specificity groups should be identified using the same reference standard, which did not happen in the CELF-5 discriminant accuracy analysis. In addition, the reference standard used to identify the sensitivity group is insufficient as discussed in the sensitivity section above.

Based on the information provided in the test manual, construct validity is insufficient. Evaluators, school districts, and families cannot take any comfort in the extremely high sensitivity and specificity provided by the CELF-5 at 1.3 standard deviations. There were only 67 children with language disorders in the sensitivity group and those children could easily have severe language disorders, intellectual disabilities and/or autism spectrum disorder. A sensitivity group made up of children with such severe disabilities would easily score as disordered on virtually any test to identify language disorders. But, whether the test is valid must be assessed with children in the low average to moderately disordered range. We simply cannot rely on the sensitivity and specificity information provided in the CELF-5 to support its construct validity.

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.

These measures are preferred to sensitivity and specificity in determining diagnostic accuracy because they are not susceptible to changes in base rate. Base rate refers to the prevalence of the clinical condition in a given population. Sensitivity and specificity are particularly sensitive to changes in base rate while likelihood ratios are less affected by changes to base rate (Dollaghan, 2007). However, since LR+ and LR- are calculated using sensitivity and specificity measures, which have already been shown to be unacceptable due to an insufficient reference standard, likelihood ratios for the CELF-5 are invalid. It is also important to take base rate and sample size into account when assessing sensitivity and specificity.

Overall, construct validity, including the reference standard, sensitivity and specificity, and likelihood ratios of the CELF-5 were determined to be unacceptable due to spectrum bias in the standardization sample and invalid reference standards.

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-5 to the CELF-4 as well as the Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4) and its expressive counterpart, Expressive Vocabulary Test, Second Edition (EVT-2).

The study conducted to compare the CELF-5 with the CELF-4 consisted of 1000 typically developing students between the ages of 5-16. Correlations between overall scores and index scores were high and ranged from .78 - .92. Overall CELF-5 scores were higher and the authors attribute this to changes in scoring procedures due to increased awareness of dialectal differences as well as “due to differences in test difficulty and differences in the normative sample” (Technical Manual, 58). It is important to note that the entire age range for which the CELF-4 and 5 are intended was not compared. The CELF-4 and CELF-5 were both standardized for use with students aged 5-21;11. However, only the age range of 5-16 was compared. Further, 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). The CELF-4 lacks sufficient discriminant accuracy according to the standards in the field. Therefore, it cannot be used to determine concurrent validity of the CELF-5.

The study conducted to compare the CELF-5 with the PPVT-4 and EVT-2 consisted of 117 typically developing students aged 5;0-16;11. Again, the entire age range for which the CELF-5 was designed was not compared. The PPVT-4 and EVT-2 are well known vocabulary tests as their test manuals state. They assess the child’s ability to match vocabulary items to drawings from the stimulus book and whether or not the child has acquired the items included in the stimulus book. According to Hart and Risley (1995), children from lower SES are exposed to far fewer words per hour than middle and upper middle class children. These authors elaborate that, vocabulary tests tend to identify socioeconomic status rather than language ability. Specifically, Horton-Ikard & Weismer (2007) found that children from low SES homes performed worse than higher SES peers on the PPVT-4 and EVT-2.These tests are not themselves valid tests of language acquisition or ability, so regardless of the correlation coefficients, concurrent validity of the CELF-5 language test cannot be determined by comparison to vocabulary tests. Therefore, due to insufficient comparison tests, concurrent validity for the CELF-5 is insufficient.

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 administering the CELF-5 to the same group of students on two separate occasions and comparing their performance. The test was administered to 137 students divided across three age bands (5;0 – 6;11, 8;0 – 9;11 and 12;0 – 16;11). The students were tested as part of the standardization sample and then repeated the test between 7-46 days later (mean = 19 days). Children aged 5;0-16;11 were used to determine test retest stability, which does not include the entire age range for the CELF-5. Correlation coefficients for composite scores and index scores ranged between .83-.90. According to Salvia, Ysseldyke, & Bolt (2010, as cited in Betz, Eickhoff, & Sullivan, 2013), many of these reliability coefficients are insufficient. They recommend a minimum standard of .90 for test reliability when using the test to make educational placement decisions, such as speech and language services. Also, the small sample size of children in each age band and limited age range, limits the reliability measure. Thus, the test-retest reliability for the CELF-5 is considered insufficient due to a small sample sizes and correlation coefficients that were less than the accepted minimum standard.

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. Scoring protocols were developed for four of the CELF-5 tests. Inter-examiner reliability was calculated for these four tests by randomly selecting two different scorers to score each protocol independently. The scores were compared and a third independent scorer resolved any differences. Inter-scorer reliability coefficients ranged between .91 and .99 indicating acceptable inter-examiner reliability (Salvia, Ysseldyke, & Bolt, 2010, as cited in Betz, Eickhoff, & Sullivan, 2013). However, inter-examiner reliability was only calculated for the four CELF-5 tests that have specific scoring protocols. It is unclear why the authors did not include inter-scorer reliability coefficients for all tests or for composite and index scores. Therefore, despite good inter-examiner reliability for four CELF-5 tests, overall inter-examiner reliability was determined to be 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 the split half method whereby scores from the first half of the test are compared to scores on the second half of the test. Inter-item consistency was evaluated using children from the normative and clinical samples and included students identified as having a language disorder, students with autism spectrum disorder, and students with reading and learning disabilities. No data is provided regarding how students from the normative sample were selected, or how many students participated in the study examining inter-item consistency. Reliability measures for the test scores for students from the normative sample ranged from .75-.98. Reliability for the composite and index scores for the normative sample ranged between .95-.96. Many reliability coefficients for the individual tests fell below the accepted standard according to Salvia, Ysseldyke, & Bolt (2010, as cited in Betz, Eickhoff, & Sullivan, 2013). While reliability coefficients for the composite and index scores are acceptable, insufficient data regarding how these values were obtained means inter-item consistency cannot be considered sufficient.

Inter-tem consistency for clinical groups was determined using a sample of 166 students aged 5;0-21;11 years previously diagnosed with LD. No mention is made regarding how these students were diagnosed or how they were recruited. Because their diagnostic status is unknown, the information provided by the reliability study cannot be generalized to other students diagnosed with an LD. A sample of 66 students with a learning disability in reading and or writing (LDR) and 69 students with ASD were also used to assess inter-item consistency. Information regarding age ranges, how diagnostic status was determined and how these students were recruited was not provided. Correlation coefficients for index and composite scores were not calculated due to insufficient sample sizes (Technical Manual, pg. 40). For the LD group reliability, coefficients ranged between .81-.97. For the LDR group reliability coefficients ranged between .86-.99, and for the ASD group coefficients ranged between .91-.99. The majority of correlation coefficients are acceptable according to the standards in the field. However, due to unclear diagnostic statuses and small sample sizes, inter-item reliability for the clinical groups is not sufficient.

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.” SEM provides an estimate of the amount of error in a student’s observed test scores. It is inversely related to the reliability of a test, which means that the smaller the SEM, the greater the reliability and confidence in the precision of the observed test score. Confidence intervals are the range of standard scores within which one can have confidence that the child’s true score lies. The CELF-5 provides confidence intervals at 68%, 90%, and 95%. For example if a child receives a scaled score of 6 on the Following Directions test, to be 95% confident that the test results captured the child’s true score, the confidence range would be scaled scores of 4 to 8 or from 2 Standard Deviations below the mean to within one standard deviation below the mean. Another way of describing this is: Based on the administration of the Following

Directions test of the CELF-5 we are 95% sure that the child's true score falls within 4 and 8 scaled scores or the child performed somewhere between moderately disordered to within normal limits. SEM allows evaluators to underscore the limitations of the single score approach to identifying disabilities.

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, in order to be more "confident" that the child's true score is included. 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.

6. BIAS:

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. 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).

According to Paradis (2005), grammatical morphology has been noted as an area of difficulty for children with LEP. In the *Word Structure Test*, students are required to apply morphological rules as well as select and use appropriate pronouns. For a child with LEP, morphological rules including appropriate pronoun selection and irregular plurals could be difficult. While scoring rules were developed to account for dialectical variations, this does not account for the extra challenges placed on a student with LEP who may have particular difficulty with less common grammatical forms for which they lack exposure. For example, in the sentence "Here is one mouse. Here are two _____" the student is required to fill in the grammatically correct word "mice". The incorrect answer, "mouses" may reflect a student's lack of exposure to English irregular plurals rather than a disorder. However, according to the Examiner's Manual, this answer would be incorrect.

Dialectal Variations

A child's performance on the CELF-5 may also be affected by the dialect of English that is spoken in their homes and communities. Consider dialect issues resulting from the test being administered in Standard American English (SAE). 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 of an archaic dialect, it would be difficult for a speaker of SAE to repeat this sentence. Consider how taking a test in your non-native language or dialect will be more taxing- it will take longer and more energy for a dialect speaker to process test items in a dialect that they are not familiar/conformable with.

Speakers of dialects other than SAE (e.g. African American English [AAE], Patois) face a similar challenge when asked to complete tests such as *Word Structure*, *Formulated Sentences* and *Recalling Sentences*. The *Formulating Sentences* test requires students to formulate complete semantically, syntactically and pragmatically appropriate sentences that incorporate a target word and pertain to a specific stimulus picture. For example, on item 1 a correct answer could be "She is washing her hands" or "She is waiting for her sister to finish washing her hands." However a student who speaks a dialect such as AAE may have difficulty with certain SAE grammatical rules and may create the sentence "She be reading." On the *Recalling Sentences* test, the student is required to repeat sentences verbatim. Omitted words, transposing words or extra words count as errors. No accommodations are made for the added memory load and challenge introduced when the student is repeating a sentence that is not typical in his or her native dialect(s). Scoring rules account for dialectal variations and the manual recommends that examiners be familiar with dialectal variations or the language used in the student's home and community to determine appropriate variations. However, if the examiner is not completely familiar with all possible dialectal variations, or the dialectal variations are not listed in the appendix, answers could be marked incorrectly due to dialectal bias.

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. 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 overall referral of minority and low SES students for special education services. Yet the test designers chose to use purely vocabulary based assessments in order to demonstrate concurrent validity, exhibiting the test designers' ignorance of the inherent bias of vocabulary tests.

A number of the CELF-5 tests such as *Formulating Sentences*, *Word Classes* and *Word Definitions* place a heavy emphasis on vocabulary, which may be more difficult for a student from a low SES. For example, item 4 on the *Word Definitions* test requires a student to define the word “cactus”. This word is not a commonly used word in most regions of the USA. Therefore, it may pose a challenge for children from low SES who have decreased opportunities for vacations, or whose parents have less experience with uncommon vocabulary items. As a result these students demonstrate reduced lexical diversity. These tests require prior knowledge of the stimulus words to provide appropriate answers. As a result, test focusing on vocabulary may result in reduced scores for children from low SES relative to their higher SES peers.

Prior Knowledge/Experience

A child's performance on the CELF-5 may also be affected by their prior knowledge and experiences. For example, a child from a large city may not know a “giraffe” (*Word Definitions*) or “snowman” (*Word Classes).* Both these tests contain low frequency vocabulary words that students from certain areas may have less experience with.

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 *Understanding Spoken Paragraphs* and *Reading Comprehension* tests require students to respond to questions to which the adult already knows the answer (“know information”) and converse with unfamiliar adults. Both these tests may be difficult for a student without appropriate prior experience or who is unaccustomed to an adult asking known questions.

Further, a child's performance on the test may have been affected by 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-5 requires students to attend to the test book for the length of the test. This may be difficult for a student who has not had prior experience with books or structured tasks. For a student to succeed on the CELF-5 they must possess skills such as print awareness or the ability to recognize that pictures and symbols convey meaning. These skills are crucial pre-literacy skills that develop through early exposure to books and print materials and do not develop in the absence of such exposure. Delayed pre-literacy skills lead to reduced metalinguistic ability, a critical skill for the *Formulating Sentences* test. This test requires students to create sentences using a provided target word. For example, item 2 requires a student to formulate a sentence using the word “airplane”. A student with limited experience or exposure to toy or real airplanes would have difficulty formulating an appropriate sentence. A student who has had limited exposure to language through books, word games and print material may be challenged on this test. Additionally, the examiner is required to select three pragmatic activities in order to complete the *Pragmatic Activities Checklist*. Many of these activities include playing a game, putting together a puzzle, or using arts and crafts materials to wrap a gift. Some students may have difficulty with these activities due to lack of prior exposure or experience despite typically developing pragmatic 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.

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. Many of the pragmatic activities on the *Pragmatics Activities Checklist* (PAC) relate to culturally specific nonverbal skills including gaze, gesture, expression and body language. The PAC includes games as well as a requirement to interact with an unknown adult are highly culturally dependent and not indicative of a disorder when not compared to typically developing peers from their speech community. These items could be difficult for a child from a cultural background that differs from the examiner or from mainstream American culture.

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 is used to less structured environments may find it very challenging to sit in front of a book for extended periods of time. The CELF-5 can take a minimum of an hour for children aged 5;0-5;5 to over two hours for children aged 17;0-21;11 in order to obtain all core and index scores. This requires the students to focus and remain interested and motivated for lengthy periods of time. This can be very difficult for students, particularly younger students.

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. 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.

A small portion (<11%) of the sample population included students from with attention deficits. However, the sample is too small to be representative of students with attention disorders and results of this assessment are invalid for children with attention 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 not using proper accommodations, 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.

Further, as the sample population did not include children from this population, results of this assessment are invalid for children with motor and sensory impairments.

7. SPECIAL ALERTS/COMMENTS

The Clinical Evaluation of Language Fundamentals (CELF-5) was designed to assess a student's language and communication skills in a variety of contexts, determine the presence of a language disorder, describe the nature of the language disorder and plan for intervention or treatment. The CELF-5 is a comprehensive and flexible assessment that correlates with current educational practices by implementing common core standards that link assessment with *instruction and intervention. The test consists of 16 tests that can be used to determine a student's language strengths and weaknesses, determine eligibility for services, plan curriculum relevant treatment, and be taken into account in the initial steps of clinical decision making regarding recommendations for "classroom language adaptations or accommodations" (Examiner's Manual, pg 1). Despite the CELF-5's attempt to design a comprehensive, valid and reliable language assessment, results obtained cannot be used to determine the presence of a language disorder. The CELF-5 was determined to lack validity due to an unrepresentative standardization sample and an insufficient reference standard. The specificity and sensitivity measures of the CELF-5 are misleading. While these measures are good according to the standards in the field, both groups were not administered the same reference standard. The population for the sensitivity measure was determined through invalid measures (e.g. previous versions of the CELF or the PLS-3 which lack acceptable diagnostic accuracy) and no reference standard was applied to the specificity group. Therefore, it lacks sufficient discriminant accuracy. In addition, there is a lack of information as to how tasks and items were deemed appropriate and free from bias. Upon examination, it was demonstrates that the CELF-5 does contain significant linguistic, cultural and socioeconomic biases, despite having created a panel specifically for the purpose of ensuring the test was without bias. As this test is largely a vocabulary test, it will likely identify socioeconomic status and second language acquisition issues, rather than a language disorder or disability.*

Due to cultural and linguistic biases (e.g. exposure to books, knowledge of SAE, syntactic and grammatical structures, low frequency vocabulary words, 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 or used in educational placement decisions. The Examiner's Manual suggests variations in scoring and administration be made for students from culturally diverse backgrounds, with dialectal variations or motor, sensory or cognitive impairments. However, should modifications be made, the authors caution against the use of scaled scores, standard scores, percentile ranks or age equivalents (Examiner's Manual, pg. 21). Scores should not be calculated. According to IDEA (2004), assessment materials are required to be valid, reliable and free of bias. Therefore, even if scores are not calculated, SLPs and others using the CELF-5 to assess a student's language should keep in mind the biases inherent in the test and testing situation which may negatively affect performance. 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).

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.

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.

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.

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.

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